# Studying the Relationship

# of Smoking to Lung Cancer Using

# the Multistage Model of Carcinogenesis:

<u>A Review</u>

P.N. Lee 1995 •

# <u>Index</u>

			<u>Page no</u>
Glo	ssary	of abbreviations	4
1.	Intro	oduction	5
	1.1	Value of models	5
	1.2	Power law relationship of mortality rates with age and the multistage model	5
	1.3	Difficulties in interpreting published mortality rates	6
2.	Derivation and assumptions		
	2.1	Assumptions	8
	2.2	Exposure constant throughout life	9
	2.3	Exposure varying during life	10
	2.4	Two relevant periods - continuous smokers	11
	2.5	Three relevant periods - giving up smoking	14
	2.6	More than three relevant periods	17
3.	Pred conf	ictions of the multistage model and ormity with observations	19
	3.1	Data sources	19
	3.2	Relationships with age, duration and age of starting to smoke	21
	3.3	Relationships with dose	28
	3.4	Relationships with stopping exposure	36
	3.5	Variation with age in relative risk associated with exposure	48
	3.6	Effects of joint exposure	51
	3.7	Effect of changing the type of cigarette smoked	52
	3.8	Relationship of dose to age of onset of exposure	53
	3.9	Other issues	54

÷

<u>بر</u>

and a local difference

athree spectrum

o a Marthur

A CONTRACTOR OF A CONTRACTOR OF

and the second second second

a succession of the second sec

AND STREET OF

			<u>Page no</u>
4.	Limi	tations of the multistage model	55
	4.1	Stages undefined	55
	4.2	Reversibility of effects may occur	56
	4.3	Transition probabilities may vary from individual to individual for a given exposure	57
	4.4	The model may be inaccurate if the transition probabilities are not small	60
	4.5	Other problems	62
5.	Appl	ications of the multistage model	63
	5.1	Using data on prevalence of smoking at different ages	63
	5.2	Applications to cohort data	65
	5.3	Whittemore (1988)	66
	5.4	Brown and Chu (1987)	68
	5.5	Other authors	71
6.	Modified versions of the multistage model		72
	6.1	Doll and Peto (1978)	72
	6.2	Townsend (1978)	73
7.	Disc mult	cussion and conclusions regarding the tistage model	75
8.	References		
Tał	les 1	1-4 and Figure	91-95

į

# <u>Glossary of abbreviations</u>

a <sub>i</sub>	transition probabilities for stage i during first period considered
В	constant relating incidence to a power of time
b <sub>i</sub>	transition probabilities for stage i during second period considered
c	proportion of susceptible
с	power of dose relationship
°i	transition probabilities for stage i during third period considered
D	duration of exposure
d	dose of carcinogen
F	length of period after stopping exposure
G <sub>T</sub>	cumulative density function at time T
<sup>_</sup> g <sub>1</sub>	Whittemore's packs function
8 <sub>2</sub>	Whittemore's multistage function
I <sub>T</sub>	incidence rate at time T
k	number of stages of the multistage process
N	number of cells at risk
P <sub>i</sub>	transition probability for stage i
R	ratio of incidences of smoker and nonsmoker
S	age of starting to smoke
s <sub>i</sub>	time at which ith period of exposure ends
Т	time
<u>T</u>	median time of tumour induction
u	transition probability for affected stage during first period
	considered
v	transition probability for affected stage during second period
	considered
W	waiting time between last transition and appearance of cancer
α	background transition probabilities for stage i
β	increase in transition probability for stage i per unit dose of
	carcinogen

÷

**.**•

#### 1. <u>INTRODUCTION</u>

#### 1.1 <u>Value of models</u>

A number of mathematical models have been used to attempt to quantify the relationship between lung cancer and various aspects of the smoking habit, such as age of starting to smoke, amount smoked, duration of smoking, and, in ex-smokers, time since stopping. Use of an appropriate model may allow prediction of future lung cancer rates and judgement as to the extent to which trends over time or differences between countries in incidence of lung cancer are explicable in terms of smoking habits or depend on other lung cancer risk factors. Ideally, a good model should not only describe well how incidence depends on smoking, but should have some biological meaning, giving insight into the mechanisms by which cancer develops. Even a good model can only approximate the truth and cannot be expected to take into account precisely the interplay of susceptibility, exposure and disease.

# 1.2 <u>Power law relationship of mortality rates with age and the</u> <u>multistage model</u>

Early interest in mathematical models for cancer started shortly after the second World War with the observation (e.g. Fisher and Holloman, 1951; Nordling, 1953) that, for many types of cancer, mortality rates rose with age according to an approximate power law, with the exponent often about 6. There are a number of difficulties in interpreting published mortality rates, described in section 1.3 below. Despite these difficulties, and despite it being apparent that the simple power law relationship did not fit for all types of

cancer (as later confirmed in a detailed analysis of 338 data sets by Cook, Doll and Fellingham (1969)), a number of models have been postulated in an attempt to try to explain this relationship. The most important of these has been the multistage model of Armitage and Doll (1954), which predicts a power law when exposure is constant and continuous, and a more complex relationship when it is The multistage model is discussed in detail in this document, not. which not only gives its derivation, but also describes how well it explains a variety of aspects of the smoking/lung cancer relationship. Other models will be considered later.

#### 1.3 <u>Difficulties in interpreting published mortality rates</u>

The major difficulties in interpreting published mortality rates can be summarized as follows:

- (a) For some cancers, though not for lung cancer, which usually is rapidly fatal, mortality rates may not bear a close correspondence to incidence rates;
- (b) Recorded mortality rates, based on death certificates, usually carried out in the absence of a post-mortem, will be inaccurate due to errors in diagnosis. For lung cancer, the techniques for diagnosing lung cancer have enormously improved between 1900 and 1950 due to the introduction of X-rays, bronchoscopy, intrathoracic surgery, sputum cytology, sulfa drugs and antibiotics (Doll and Peto, 1981), though even now the rate of false-positive and false-negative diagnosis remains quite high (e.g. Szende <u>et al</u>, 1994), particularly at ages 80 or over (Doll, 1971).

- (c) Mortality rates, and indeed incidence rates from cancer registries, do not distinguish between the different histological types of lung cancer, such as squamous cell cancer and adenocarcinoma, which may show different relationships with age, smoking habits and other factors.
- (d) Experimental studies are often conducted on genetically similar animals and exposure to the agent of interest is carefully controlled. Human populations, however, vary widely both in susceptibility and exposure. The observed patterns of incidence may be very different for different subsets of the population.
- (e) Studying variation in rates by age for one particular year inevitably means one is comparing different birth cohorts at each age, with differing patterns of smoking habits and exposure to other risk factors. The study of variation in rates by age for one particular birth cohort, on the other hand, means comparison over a long time period during which <u>inter</u> <u>alia</u> diagnostic standards may have changed.
- (f) Because of competing risk of death from other diseases, people surviving to older ages may be unrepresentative, in respect of susceptibility and exposure, of the whole population from which (Indeed, even in the absence of deaths from they are derived surviving population may be other causes, the unrepresentative, especially for genetic diseases, such as familial polyposis coli and Huntingdon's chorea, where risk rises with age and then falls off, to zero, as the susceptible pool is eliminated.)

- (g) There may be inadequate available comparable data on variation by age, sex and year in smoking habits. Data on cigarette consumption per head drawn from sales statistics are usually not age or sex specific; averages may be more appropriate to age groups 20 or 30 years younger than the ages at which lung cancer normally occurs.
- (h) Published mortality rates typically do not take account of the effect of variations in exposure to other risk factors for lung cancer, such as occupational exposure, air pollution and diet.

# 2. DERIVATION AND ASSUMPTIONS

## 2.1 Assumptions

The multistage model assumes a single cell can generate a malignant tumour only after undergoing a certain number, k, of heritable changes, a cell having undergone s stages being said to be "at stage s." For each person at risk, the tissue in question is assumed to consist initially of N normal cells (at stage zero), each with the same likelihood of independently progressing through the multistage process. The model also assumes the progression of changes must occur in a specific order and that the background rate of occurrence of each change is a constant which is independent of age, carcinogens acting by increasing the rate of occurrence of one or more stages above the background rate. If  $p_i$  is the "transition" probability (per unit time) of change for a cell having experienced exactly i-1 changes, the probability that the kth change occurs in the short time interval (t, t+dt) is approximated by

$$\frac{\mathbf{p}_{1}\mathbf{p}_{2} \dots \mathbf{p}_{k} \mathbf{t}^{k-1} \, \mathrm{dt}}{(k-1)!} \tag{1}$$

as t+0. This result will be valid for large values of t (of the order of a human lifetime) provided that  $p_1t$ ,  $p_2t$ , ...  $p_kt$  are all sufficiently small. The incidence rate per person is obtained by multiplying (1) by N. For a rigorous proof, see Armitage (1953); for a less rigorous proof, see Armitage and Doll (1954).

## 2.2 Exposure constant throughout life

Providing that the transition probabilities remain constant throughout life, the incidence rate,  $I_T$ , of cancer at time T will be given by the simple formula

$$I_{T} = BT^{k-1}$$
 (2)

where B is a constant equal to  $Np_1p_2 \dots p_k / (k-1)!$ 

This is the simple power law relationship observed by Fisher and Holloman (1951) and by Nordling (1953). The incidence rate is that for a Weibull distribution, where the cumulative density function,  $G_{\rm T}$ , is given by

$$G_{T} = 1 - \exp(-BT^{k})$$
 (3/1)

As noted by Pike (1966), this distribution may actually arise under quite broad assumptions concerning the distribution of time to onset of cancer in individual cells (i.e. the model implies the formula; but the formula does not imply the model). The Weibull distribution. is in fact also known as the "third asymptotic distribution of smallest values" discovered by Frechet (1927) and by Fisher and Tippett (1928) (see Gumbel (1958) for a discussion of the derivation of the three distributions and of their properties). This distribution is often expressed with an extra parameter W as

$$G_{T} = 1 - \exp(-B(T-W)^{k})$$
 (3/2)

In the context of the multistage model, W is often interpreted as the "waiting time" between the last transition occurring and clinical appearance of, or death from, lung cancer. To simplify the presentation that follows we ignore W, though note that some researchers, when fitting the multistage model, ignore exposure up to a short period (eg. 2 years) before recorded diagnosis or death to try to take account of this waiting time.

## 2.3 Exposure varying during life

In the simplest use of the multistage model, the transition probabilities are assumed to remain constant throughout life. A strength of the model is that incidence can readily be calculated for varying probabilities, e.g. resulting from varying exposure. Again assuming transition probabilities are small, and, for convenience, taking k=5, the incidence rate at time T is given by the formula

$$I_{T} = p_{5} \int_{0}^{T} p_{4} \int_{0}^{t_{4}} p_{3} \int_{0}^{t_{3}} p_{2} \int_{0}^{t_{2}} p_{1} dt_{1} dt_{2} dt_{3} dt_{4} dt_{5}$$
(4)

-10-

where the p<sub>i</sub> are the time-dependent transition probabilities for each stage.

Although it is in theory possible to take into account any form of functional dependence of the transition probabilities on age, the most common uses of the multistage model have been where transition probabilities are either unaffected by exposure, and take "background" values  $\alpha_i$  which are invariant of age, or are affected by exposure, taking the constant value  $\alpha_1 + \beta_i d = \gamma_i$  when exposure occurs, d being dose of carcinogen applied. In the simpler applications, dose is constant during exposure. In some contexts,  $\beta_i d$  may be large with respect of  $\alpha_i$ , so that the transition probability is approximately directly proportional to dose.

# 2.4 <u>Two relevant periods - continuous smokers</u>

One particularly useful form of the incidence rate formula applies where there are two periods of time, during the first of which [0,S] the transition probabilities are  $a_i$  and during the second of which [S,T] the transition probabilities are  $b_i$ . In the context of smoking, S can be viewed as the age of starting to smoke, smoking continuing subsequently.  $a_i$  are background probabilities in the absence of smoking,  $b_i$  the probabilities during smoking. Up to time S, the incidence rate is as for formula (2). Subsequently, the formula is given by

2 stage process

 $I_{T} = N [a_{1}b_{2}S + b_{1}b_{2} (T-S)]$ (5/2)

$$I_{T} = N \left[ \frac{a_{1}a_{2}b_{3}S^{2}}{2} + a_{1}b_{2}b_{3}S(T-S) + \frac{b_{1}b_{2}b_{3}(T-S)^{2}}{2} \right]$$
(5/3)

<u>4 stage process</u>

$$I_{T} = N \left[ \frac{a_{1}a_{2}a_{3}b_{4}S^{3}}{6} + \frac{a_{1}a_{2}b_{3}b_{4}S^{2}(T-S)}{2} + \dots \right]$$
  
$$\dots + \frac{a_{1}b_{2}b_{3}b_{4}S(T-S)^{2}}{2} + \frac{b_{1}b_{2}b_{3}b_{4}(T-S)^{3}}{6} \right]$$
(5/4)

$$\frac{5 \text{ stage process}}{I_{T}} = N \left[ \frac{a_{1}a_{2}a_{3}a_{4}b_{5}S^{4}}{24} + \frac{a_{1}a_{2}a_{3}b_{4}b_{5}S^{3}(T-S)}{6} + \dots + \frac{a_{1}a_{2}b_{3}b_{4}b_{5}S^{2}(T-S)^{2}}{4} + \frac{a_{1}b_{2}b_{3}b_{4}b_{5}S(T-S)^{3}}{6} + \frac{b_{1}b_{2}b_{3}b_{4}b_{5}(T-S)^{4}}{24} \right] (5/5)$$

More generally, for a k stage process, the formula can be derived noting that the terms within the square bracket arise from a binomial expansion of  $[S + (T-S)]^{k-1} / (k-1)!$  with each term being multiplied by appropriate values of  $a_i$  or  $b_i$ , the first term relating to cancers where the first k-1 transitions occur before S, the second term to cancers where the first k-2 transitions occur before S, and so on (the last transition must occur after S, at time T, by definition).

Note that these formulae can be considerably simplified when only one, or a limited number of stages, are affected by exposure. As an example consider the four stage process where only the first stage is affected. If a are the background transition probabilities for unaffected stages, u is the transition probability for the affected stage during the period [0,S] and v the transition probability for the affected stage during the period [S,T], we have

$$I_{T} = \frac{Na_{2}a_{3}a_{4}}{6} [uS^{3} + 3uS^{2}(T-S) + 3uS(T-S)^{2} + v(T-S)^{3}]$$
  

$$\approx uT^{3} + (v-u)(T-S)^{3}$$

More generally, for a k stage process with the <u>first stage</u> affected

$$I_{T} \simeq uT^{k-1} + (v-u)(T-S)^{k-1}$$
 (6/1)

With the penultimate stage affected, we have

$$I_{T} \simeq (u-v)S^{k-1} + vT^{k-1}$$
 (6/2)

With the first and penultimate stages affected, we have

$$I_{T} \approx u_{1}u_{2}s^{k-1} + v_{1}v_{2} (T-s)^{k-1} + u_{1}v_{2} (T^{k-1}-s^{k-1}-(T-s)^{k-1})$$
(6/3)

(Here  $u_1$  and  $v_1$  refer to the first stage transition probabilities, and  $u_2$  and  $v_2$  refer to the penultimate stage transition probabilities.)

As discussed elsewhere, e.g. by Day and Brown (1980), Brown and Chu (1983b) and Brown and Chu (1987), these formulae allow some fairly simple conclusions. Let us consider firstly <u>excess</u> incidence at age T in relation to exposure starting at time S. Where only the <u>first stage</u> is affected, since the incidence at age T in the absence of carcinogenic exposure would be  $uT^{k-1}$ , since the duration of exposure, D, equals (T-S) and since v-u is linearly proportional to dose d, we have (from formula 6/1)

$$I_{\rm T} \simeq dD^{\rm k-1} \tag{7/1}$$

i.e. the excess risk at a given age is proportional to dose, depends (by a power-law relationship) on duration of exposure, but is independent of age of starting to smoke. Where the <u>penultimate</u> <u>stage</u> is affected we have (from formula 6/2)

$$I_{T} \approx d[(D+S)^{k-1} - S^{k-1}]$$
 (7/2)

i.e. the excess risk is proportional to the dose d and is an increasing function of both duration given age of start, and of age of start given duration. Where the <u>first and penultimate stages</u> are affected, the excess risk can be expressed by the formula

$$I_{T} \simeq d_{1}D^{k-1} + d_{2}[(D+S)^{k-1} - S^{k-1}] + d_{1}d_{2}D^{k-1}$$
 (7/3)

Here  $d_1$  and  $d_2$  are the effective excess doses, relative to background, for the first and penultimate stages (i.e. if the dose increases the background risk by a factor q, the effective dose is q-1). Note that setting  $d_2 = 0$  gives formula (7/1) and setting  $d_1 =$ 0 gives formula (7/2).

# 2.5 <u>Three relevant periods - giving up smoking</u>

The same authors note that inferences can similarly be made by examining the excess risk patterns for those individuals who have stopped their exposure. When the exposure starts at age S, continues for a duration D, then stops, and follow-up continues for a period of length F, the excess risk at age S+D+F = T is given by

$$I_{T} \simeq d[(D+F)^{k-1} - F^{k-1}]$$
 (8/1)

when only the first stage is affected by the carcinogen, by

$$I_{T} \simeq d[(D+S)^{k-1} - S^{k-1}]$$
 (8/2)

where only the penultimate stage is affected, and by

$$I_T \simeq d_1 [(D+F)^{k-1} - F^{k-1}] + d_2 [(D+S)^{k-1} - S^{k-1}] + d_1 d_2 D^{k-1}$$
 (8/3)

where both the <u>first and penultimate stages</u> are affected. Note that Whittemore (1988) gives a version of this formula (her formula 12 using different notation) which is incorrect, including a term  $d_1 D^{k-1}$  rather than the correct term  $d_1 [(D+F)^{k-1} - F^{k-1}]$ . These terms are the same where exposure is not discontinued (F = 0) but not otherwise.

These inferences for stopping smoking can be derived from formulae (analogous to formulae 5) in which there are three periods of time, during the first of which  $[0,S_1]$  the transition probabilities are  $a_i$ , during the second of which  $[S_1,S_2]$  the transition probabilities are  $b_i$ , and during the third of which  $[S_2,T]$  the transition probabilities are  $c_i$ . Below we give the formulae for a <u>4 stage process</u>.

$$I_{T} = N \left[ \frac{a_{1}a_{2}a_{3}c_{4}S_{1}^{3}}{6} + \frac{a_{1}a_{2}b_{3}c_{4}S_{1}^{2}(S_{2}-S_{1})}{2} + ... \right]$$

$$+ \frac{a_{1}a_{2}c_{3}c_{4}S_{1}^{2}(T-S_{2})}{2} + \frac{a_{1}b_{2}b_{3}c_{4}S_{1}(S_{2}-S_{1})^{2}}{2} + ...$$

$$+ \frac{a_{1}b_{2}c_{3}c_{4}S_{1}(S_{2}-S_{1})(T-S_{2})}{1} + \frac{a_{1}c_{2}c_{3}c_{4}S_{1}(T-S_{2})^{2}}{2} + ...$$

$$+ \frac{b_{1}b_{2}b_{3}c_{4}(S_{2}-S_{1})^{3}}{6} + \frac{b_{1}b_{2}c_{3}c_{4}(S_{2}-S_{1})^{2}(T-S_{2})}{2} + ...$$

$$+ \frac{b_{1}c_{2}c_{3}c_{4}(S_{2}-S_{1})(T-S_{2})^{2}}{2} + \frac{c_{1}c_{2}c_{3}c_{4}(T-S_{2})^{3}}{6}$$

$$(9)$$

More generally, for a k stage process, the formula can be derived noting that the terms within the square brackets arise from a multinomial expansion of  $[S_1 + (S_2 - S_1) + (T - S_2)]^{k-1}/(k-1)!$  with each term being multiplied by appropriate values of  $a_i$ ,  $b_i$  or  $c_i$ , to describe the various sequences by which cancer can arise. For example the 5th term above describes the cases where the first transition occurs in  $[0,S_1]$ , with contribution  $a_1S_1$  to the formula (probability x length of period), the second transition occurs in  $[S_1, S_2]$ , with contribution  $b_2(S_2-S_1)$ , and the third occurs in  $[S_2, S_1]$ T], with contribution  $c_3(T-S_2)$ , the fourth occurring at T, with contribution  $c_{L}$ . Where multiple (z) transitions occur in one in the first term the first three changes occur in period, e.g. the denominator includes a term z! to take account of the [0,S<sub>1</sub>], fact that only one of the possible sequences of transition is allowed (the transitions must be in order).

Formulae 8 can readily be shown to be special cases of formula 9.

-16-

# 2.6 More than three relevant periods

It may also be useful to write down the formula for the situation where there are two periods of identical exposure, a person having periods of length U, V, W, X, Y respectively unexposed, exposed, unexposed, exposed and unexposed, i.e. the person starts smoking and gives up twice. Where both the <u>first and</u> <u>penultimate stages</u> are affected, the excess risk is given by

$$I_{T} \approx d_{1}[(V+W+X+Y)^{k-1} - (W+X+Y)^{k-1} + (X+Y)^{k-1} - Y^{k-1}] + d_{2}[(U+V+W+X)^{k-1} - (U+V+W)^{k-1} + (U+V)^{k-1} - V^{k-1}] + d_{1}d_{2}[(V+W+X)^{k-1} + V^{k-1} + W^{k-1} + X^{k-1} - (V+W)^{k-1} - (W+X)^{k-1}] (10)$$

The simpler formulae when only the <u>first</u> or only the <u>penultimate stages</u> are affected are given by setting  $d_2 = 0$  or  $d_1 = 0$ , respectively, in the above formula.

This formula can be extended to larger numbers of exposure periods by realizing that:

- (a) the term in d<sub>1</sub> (the first stage effect) is the sum of (k-1)th powers of the length of all periods starting at the beginning of an exposure period and ending at t, minus the sum of (k-1)th powers of the length of all periods starting at the end of an exposure period and ending at t;
- (b) the term in d<sub>2</sub> (the penultimate stage effect) is the sum of
   (k-1)th powers of the length of all periods starting at time 0

and ending at the end of an exposure period, minus the sum of (k-1)th powers of the length of all periods starting at time 0 and ending at the beginning of an exposure period;

(c) the term in  $d_1d_2$  (the joint effect) is the sum of (k-1)th powers of the length of all periods starting at the beginning of an exposure period and ending at the end of an exposure period, minus the sum of (k-1)th powers of the length of all periods which either start at the beginning of one exposure period and end at the beginning of another or start at the end of one exposure period and end at the end of another.

Where both the first and penultimate stages are affected, it is possible to write down the formula for the excess at time T for a subject whose life can be divided into T equal annual periods (i = 1, ... T) during which the standardized "dose" from cigarettes is  $Z_i$ . It is given by

$$I_{T} \simeq d_{1} \sum_{i=1}^{T} F_{i} Z_{i} + d_{2} \sum_{i=1}^{T} G_{i} Z_{i} + d_{1} d_{2} \sum_{i=1}^{T} \sum_{j=i}^{T} H_{ij} Z_{i} Z_{j}$$
(11)

where  $F_i = (T - i + 1)^{K-1} - (T - i)^{K-1}$ 

 $G_i = F_{T-i+1}$ 

and  $H_{ij} = 1$  if i = j or

= 
$$(j - i + 1)^{K-1} - 2(j - i)^{K-1} + (j - i - 1)^{K-1}$$
 if  $i < j$ 

In this formulation, a smoker smoking at a unit dose  $(Z_i = 1)$ in a time period increases the first and penultimate stage transition probabilities by factors of, respectively,  $1 + d_1$  and  $1 + d_2$ , with smokers smoking at dose Z increasing the probabilities by factors of  $1 + d_1Z$  and  $1 + d_2Z$ . Dose may, for example, be proportional to the product of number smoked per day and tar level of brand smoked to allow for the effects of variation in these aspects of smoking.

## 3. PREDICTIONS OF THE MULTISTAGE MODEL AND CONFORMITY WITH OBSERVATIONS

The multistage model makes a number of predictions as to how the cancer incidence rate will depend on various aspects of the data. These are considered in some detail, comparing the predictions as appropriate with epidemiological and animal data. Before looking at these various aspects in turn, we first summarize some of the key data sources we will use as reference for comparison.

#### 3.1 Data sources

<u>British Doctors Study</u>. In 1951 Doll and Hill sent a questionnaire on smoking habits to all men and women on the British Medical Register. The 34,000 men and 6,000 women who replied have been followed up for mortality ever since. Results of 20 year follow-up for men are given in Doll and Peto (1976) and of 22 year follow-up for women are given in Doll <u>et al</u> (1980). Doll and Peto (1978) give a detailed tabulation of lung cancers and man-years at risk by age and amount smoked for men who had never smoked and for men who started smoking at ages 16-25 and continued to smoke.

<u>US Veterans' Study</u>. In 1954 Dorn mailed questionnaires to US veterans, mainly of World War I, who held Government life insurance policies. Almost all policy holders were white males. Almost 250,000 responses were received. Kahn (1966) gives extensive tables or results relating to follow up after 8½ years. Rogot (1974) gives less detailed results for 16 years follow-up.

American Cancer Society (ACS) Cancer Prevention Studies I and II (CPS I and II). The ACS have sponsored two huge prospective studies of smoking and mortality in the United States. In the first study about 1 million persons were followed from 1959 until 1972, in the second study about 1.2 million persons were followed from 1982 until 1988. There have been a very large number of papers published about In particular Hammond (1966) gave very detailed results for CPS I. four years follow-up, and various reports of the US Surgeon-General (particularly 1979, 1982 and 1989) have presented summary results. The 1989 report has also presented some results for CPS II, though It should be noted that extensive tables have yet to be published. the sampling in both studies was by ACS volunteers and those interviewed are not representative of the US population. In particular they are far more likely than average to be white, have higher education and income and lower exposure to occupational carcinogens and lower mortality than average.

<u>Studies of skin painting of mice</u>. During the 1960's and early 1970's a large number of studies were carried out in which the backs of mice were painted regularly with tobacco smoke condensate or with known carcinogens as a model for human carcinogenesis. Studies were carried out by the Tobacco Research Council at Harrogate, by the Medical Research Council at Pollard's Wood and by other laboratories. Relevant papers include Lee (1974), Lee and O'Neill (1971), Lee, Rothwell and Whitehead (1977) and Peto <u>et al</u> (1975).

## 3.2 <u>Relationships with age, duration and age of starting to smoke</u>

As shown by formula 2, the multistage model predicts that if the transition probabilities remain constant throughout life the incidence rate of cancer will bear a simple power law relationship to <u>age</u>. Where the first stage is very strongly affected then, regardless of which other stages are affected, the incidence rate will have a simple power law relationship to <u>duration of exposure</u>. For example, take formula 6/3 and let  $u_1$  tend to zero. However, where the first stage is not affected, one may get a more complex relationship (see formula 7/3).

As noted above, the multistage model was actually derived to explain the fact that, for many cancers, incidence (or mortality) rates tend to rise approximately according to a power of age (Fisher and Holloman, 1951; Nordling, 1953), although the relationship shows upward or downward curvature from this general pattern in many cases (Cook, Doll and Fellingham, 1969), even if one excludes from

-21-

analysis incidence rates observed at high age, where diagnosis is unreliable.

A particularly important study was that on mouse skin reported by Peto <u>et</u> <u>al</u> (1975). In this study a total of 950 mice with a normal lifespan of two to three years were exposed to regular application of benzpyrene (a proven carcinogen) starting at 10, 25, 40 or 55 weeks of age. In each group the incidence rate of malignant epithelial skin tumours among the survivors increased similarly according to a power of duration of exposure. Given duration of exposure, incidence was shown to be completely independent of age. These results suggested that observed approximate power-law increases in most human adult cancer incidence rates with age could exist merely because age equals duration of exposure to background and carcinogenic stimuli. The results could be explained without postulating any intrinsic effects of ageing (such as failing immunological surveillance or age related hormonal changes), and are consistent with our multistage hypotheses in which benzpyrene strongly affected the first stage (and perhaps also other stages) of a multistage process, with background transition probabilities invariant of age.

Another interesting observation consistent with the notion that age <u>per se</u> need not be relevant to risk of cancer occurrence is that reported by Lijinsky (1993). Collecting evidence from studies in 20 species of mammals, reptiles, birds, amphibians and fish exposed to approximately 1000 mg/kg body weight lifetime dose of nitrosodiethylamine, he noted that, despite the great variation in lifespan (from 3 years in mice to over 50 years in snakes), tumours developed within a similar period, of about a year. He felt that "the evidence suggests that the time dependence of tumour development is more likely related to the cumulative dose of carcinogen than to lifespan and the rate of aging".

The results of a study by Stenbäck <u>et al</u> (1981), in which mouse skin tumours were induced by a single initiating dose of DMBA followed three weeks later by application of the tumour promoter TPA, do not fit in so well with the simple multistage theory. They reported a highly significantly lower yield of tumours when initiation took place at 68 weeks of age than when it took place at 8 or at 48 weeks of age. The authors suggested that this difference was chiefly due not to changes in the number of cells initiated by DMBA but rather to a decrease in the promotional efficacy of TPA in ageing mice.

Peto <u>et al</u> (1985) consider these and additional animal experiments, concluding that the observations "argue strongly that there is no systematic tendency for old animals to be more susceptible to the processes of carcinogenesis than younger animals are", a conclusion reflected in the provocative title of their paper, "There is no such thing as ageing, and cancer is not related to it".

-23-

Turning now to humans, Seidman (1985) and Peto <u>et al</u> (1982), have analysed data relating incidence of mesothelioma in asbestos workers to age, age at start of exposure and duration of exposure. Just as in the Peto <u>et al</u> (1975) benzpyrene mouse study, they found that, given duration of exposure, age at start of exposure was irrelevant. Peto <u>et al</u> (1982) concluded that their results support the multistage model of carcinogenesis "under which the increase in most cancer incidence rates with age is due to a constant incidence of genetic or epigenetic accidents, rather than to progressive generalized changes in regulatory or immune function".

Given duration of exposure, age at start of exposure is associated with risk of some cancers. One case in point is lung cancer due to arsenic exposure. Brown and Chu (1983a,b) compared risk of lung cancer in groups of copper smelter workers exposed to arsenic and found that risk increased steadily as age at start of exposure increased from <20, through 20-29 and 30-35, up to 40-49 However this does not of itself mean that their results are years. inconsistent with the multistage hypothesis, rather that one needs to assume that arsenic affects a late stage of the process in order In fact, Brown and Chu fitted the actual to explain the results. functional form of the excess cancer risk predicted by the multistage theory to their detailed data on risk of lung cancer by age at initial employment and duration of level of exposure, employment and found an excellent fit to formula 7/2, in which the penultimate stage of a four stage process is affected. This formula fitted the data considerably better than formula 7/1, in which the

first stage is affected and the authors concluded that "the results indicate that arsenic exerts a definite late stage effect though an additional effect at the initial stage cannot be ruled out".

Doll (1971), using data from his British Doctors Study, plotted, on a double logarithmic scale, lung cancer incidence rates in man

(a) for nonsmokers, against age,

(b) for smokers, against age, and

(c) for smokers, against duration of smoking.

Since the amount smoked varied with age, the incidence rates in smokers were standardized for smoking habits. Equations (a) and (b) both showed a good linear relationship (consistent with formula 2) but the slopes of the lines varied markedly, with k estimated as 5 for nonsmokers and about 8.5 for cigarette smokers. However, when plot (c) was considered, the position was changed. In this case the relationship remained linear, but the value of k for smokers became much lower and very similar to that for nonsmokers. The graphical results presented by Doll were consistent with lung cancer resulting from a 5 stage process, with risk related to duration of exposure. In nonsmokers exposure is from birth to a weak carcinogen; in smokers exposure is from start of smoking to a stronger carcinogen. Note that, in theory (see formula 7/1), excess, not absolute, risk in smokers should be proportional to a power of duration of However, since risk in smokers is so much higher than in exposure. nonsmokers (relative risk of about 14 in the British Doctors Study), excess and absolute risk are very similar.

While many studies other than that on the British Doctors allow one to investigate how risk rises with age in smokers and nonsmokers, relatively few studies provide useful data on how risk varies by age of starting to smoke given duration of exposure. A problem of course is that most smokers tend to start smoking within a relatively short period of time and it is difficult to accumulate sufficient data on people starting very early or very late to allow reliable comparison. Perhaps the best data, reproduced in Table 1, comes from the Veterans' Study (Kahn, 1966). If one looks at the data for all cigarette smokers a striking fact emerges, namely that increasing age by 10 years has a virtually identical effect to decreasing age of starting to smoke by 10 years. Thus comparing two groups of smokers, both with a duration of about 43 years, one aged 55-64 and starting to smoke at age 15-19, the other aged 65-74 and starting to smoke at age 25+, we see their lung cancer rates (168 and 162 per 10<sup>5</sup> per year) are virtually identical. Similarly comparing two groups of smokers, both with a duration of about 48 years, one aged 55-64 and starting to smoke at age <15, the other aged 65-74 and starting to smoke at age 20-24, we again see lung cancer rates (251 and 241 per  $10^5$  per year) that are very similar. At first sight these results are consistent with the Peto et al (1975) mouse skin results showing irrelevance of age given duration of smoking. However, if one looks at the results in Table 1 broken down further by amount smoked, the pattern is not so clear cut. Where adequate numbers of deaths are available (in the 10-20 and 21-39 cigs/day group) there is a consistent tendency for risk to be somewhat higher in the older smokers in the above comparisons. The

-26-

simple comparison for all cigarette smokers appears to be somewhat biassed because it fails to take into account the fact that people who start to smoke younger smoke rather more cigarettes a day than those who start to smoke older. However the inference that age is important given duration is not totally secure, bearing in mind the uncertainty present in what the mean durations in the various groups are, given the relatively wide and in some cases open-ended intervals. Thus, for example, if the average age of starting in the <15 group is say 13.5 and that in the 20-24 group is say 21.5, one may not be comparing groups with identical durations (when one compares 55-64 year olds and 65-74 year olds) but groups which differ in duration by two years.

Another study that has provided relevant data is that by Lubin <u>et al</u> (1984). As described in more detail below (section 5.4), Brown and Chu (1987) found that a multistage model in which the first and penultimate stages were affected by smoking predicted reasonably well the variation observed in risk of lung cancer by age of starting to smoke, given age.

Hegmann <u>et al</u> (1993) have also presented data consistent with a major effect of age of starting to smoke. Based on a case-control study in Utah involving 282 lung cancer cases and 3282 population controls they found that, after adjusting for age and amount smoked, men who started to smoke before age 20 had a substantially higher risk of lung cancer (RR compared to nonsmokers = 12.7, 95% CI 6.39-25.2) than men who started later (6.03, 2.82-12.9). For women the heavy increase in risk continued until age 25 (9.97, 4.68-21.2) compared with women who began smoking at age 26 or older (2.58, 0.53-12.4). No analyses were presented comparing risk in smokers of the same duration but of differing ages.

Perhaps the safest conclusions to draw are those given in the IARC (1986) monograph on tobacco smoking. They note that "the effects of the duration of smoking are so strong, and so closely correlated with age, that it is virtually impossible to determine exactly whether ageing <u>per</u> <u>se</u> has any independent effect on excess lung cancer rates among people of different ages who have all smoked similarly for a similar number of years. If age has any independent effect, however, this would be small compared with the accumulative effect of duration of smoking (Peto <u>et al</u>, 1975, 1985; see also Likhachev <u>et al</u>, 1985)".

The data in Table 1 can be used not only to demonstrate that risk depends much more strongly on duration of smoking than on age given duration, but also to demonstrate an approximate power law relationship between duration and risk. Table 2 shows the result of fitting a fourth power relationship of duration to lung cancer risk. It can be seen that the fit is very adequate.

# 3.3 <u>Relationships with dose</u>

Given continuous exposure to a dose of a carcinogen, then under the multistage assumptions it has already been shown that the risk of lung cancer at a given age is proportional to the product of the

-28-

individual transition probabilities. For a stage affected by the carcinogen one might assume that the transition probability,  $p_i$ , is linearly related to dose d by the formula

$$p_{i} = \alpha_{i} + \beta_{i} d \tag{12}$$

Here  $\alpha_i$  is the background value of the transition probability, and  $\beta_i$  is the coefficient of the regression of the transition probability on dose. Where the carcinogen strongly affects risk, so that  $\beta_i d \gg \alpha_i$  one would then get the approximate relationship

$$p_i = \beta_i d \tag{13}$$

i.e. a direct linear relationship of transition probability with dose. Where the particular stage is unaffected by the carcinogen, one would have  $\beta_i = 0$  so that

$$p_{i} = \alpha_{i} \quad (constant) \tag{14}$$

Based on this formulation one would expect the following relationship between incidence rate and the number of stages affected:

- (i) <u>One stage strongly affected</u>. Risk proportional to dose, linear through the origin.
- (ii) <u>One stage weakly affected</u>. Risk proportional to dose, linear not through the origin.
- (iii) <u>Two stages strongly affected</u>. Risk directly proportional to dose squared.
- (iv) <u>Two stages affected</u>, one or both weakly. Quadratic relationship of risk to dose.

- (v) <u>C stages strongly affected</u>. Risk directly proportional to dose to the power c.
- (vi) <u>C stages affected</u>, <u>some weakly</u>. Cth power polynomial relationship of risk to dose.

A striking example of data fitting the multistage hypothesis both in respect of dose and time comes from the mouse skin painting studies of Lee and O'Neill (1971). In two separate experiments benzopyrene was painted regularly on the backs of mice at different dose levels (6, 12, 24 and 48  $\mu$ g per week in the Harrogate study; 1, 3, 9 and 27  $\mu$ g per week in the Zurich study). In both studies the incidence, both of tumours and of infiltrating carcinomas, was very well fitted by the expression

$$I_{T} = d^{2}(T-W)^{k}$$
(15)

where T is time from first application, d is the applied dose, and W and k are constants independent of dose. The direct quadratic relationship of incidence with dose was consistent with benzopyrene strongly affecting two stages of mouse skin carcinogenesis.

There are a number of reasons (some applicable to humans only, some to animals also) why one might not always expect to see such a simple relationship of incidence to dose. These include:

 Numbers of cigarettes smoked per day may not be a direct index of exposure to target tissues of relevant smoke constituents, e.g. smokers of differing numbers of cigarettes a day inhale differently;

- (ii) Numbers of cigarettes smoked per day may be inaccurately reported; low numbers may be understatements, high numbers exaggerations. There are no data relating lung cancer risk to objective markers of smoke uptake. (Even if there were, current markers, such as cotinine, only quantify recent exposure to one constituent of smoke.)
- (iii) Numbers of cigarettes smoked per day may depend on susceptibility to disease. Sufferers of symptoms may cut down; those with strong constitutions may stay smoking high numbers.
- (iv) Smokers of different numbers of cigarettes may differ in respect of various other characteristics - age, age of starting to smoke, diet, occupation, etc, etc.
- (v) At high doses cells may be killed off before they get the chance to be transformed into cancerous cells. It is generally believed (Major and Mole, 1978) that cell killing by radiation is an explanation for the fact that the risk of induced leukaemia flattens off and then falls above a given dose, and Davies <u>et al</u> (1974) suggest it may explain why in mouse skin painting studies with various cigarette smoke condensates the log incidence/log dose relationship becomes less steep at high doses.
- (vi) It may not be correct that the transition probability for a given stage is actually directly proportional to dose.

Despite these reasons, dose-response relationships consistent with the multistage formulation are found to fit many data sets quite well. Druckrey (1967) has summarized the results of extensive

-31-

animal studies over more than 25 years involving a total of about 10,000 rats treated with a variety of carcinogenic substances. He noted that for all the carcinogens he studied, the relationship between dose d and median time of tumour induction  $\underline{T}$  could be summarized by the general formula:

$$d\underline{\mathbf{T}}^{\mathbf{n}} = \text{constant}$$
 (16)

(N.B. His studies generally involved such high doses of carcinogenic substances that deaths from other causes did not obscure this simple relationship.) As shown in formula 3/1 the distribution of time to tumour in the absence of death from other causes is given by

$$G = 1 - \exp(-BT^k)$$

and the state of the second

Substituting  $B = d^{C}$  (where a carcinogen strongly affects c stages) we have

$$G = 1 - \exp(-d^{C}T^{K})$$

$$(17)$$

At the median G = 0.5, so we have

$$\exp(-d^{C}\underline{T}^{K}) = 0.5$$
 (18/1)

or 
$$d^{c} \underline{T}^{k} = \log_{e} 2$$
 (18/2)

or 
$$d\underline{T}^{k/c} = (\log_e 2)^{1/c} = constant$$
 (18/3)

which is exactly of the form that Druckrey (who did not invoke multistage assumptions at all) found to hold in practice.

Though Druckrey's simple formula may only hold for studies such as his with strong carcinogens where essentially all the animals get tumours, and deaths from other causes rarely occur (so that the observed median time is close to the true median time in the absence of deaths from other causes), his results are completely consistent with what is predicted by the multistage model. It is interesting to note that Druckrey always found his n to be greater than 1, i.e. the carcinogen never affected all the stages of the multistage process. Peto (1977) has also pointed out the dose power is invariably less than the time power. As Armitage and Doll (1954) note, this observation is inconsistent with the Fisher and Holloman (1951) model (<u>vide infra</u>) which predicts that the two powers should be the same.

A number of the major prospective studies on smoking and health have presented data relating incidence rate of lung cancer with amount smoked (see e.g. USSG 1982). All the studies show that risk increases with amount smoked. Generally the dose-response seems to be approximately linear. In view of evidence described elsewhere in section 2 that risk of lung cancer in ex-smokers rapidly becomes less than that in continuing smokers (which suggests a late stage is affected), and evidence that risk of lung cancer in continuing smokers of a given age depends strongly on age of starting to smoke (which suggests an <u>early</u> stage is affected) this linear dose-response seems somewhat surprising. If two stages are affected then surely the dose-response relationship should have a quadratic component?

Doll and Peto (1978) attempted to answer this point, put forward by Armitage (1971) when discussing a paper by Doll (1971). Based on 20-year follow-up data from the British Doctors study,

-33-

they studied the relationship of annual lung cancer incidence rate to age and number of cigarettes smoked among cigarette smokers of age 40-79 who started to smoke at age 16-25 and who smoked 40 or less per day. They reported an adequate fit to the formula

Lung cancer incidence =  $0.273 \times 10^{-12} (\text{cigs/day} + 6)^2 (\text{age} - 22.5)^{4.5}$ 

They noted that the form of the dependence on dose is "subject not only to random error but also to serious systematic biases", biases which they discussed in the paper. They emphasized that "there was certainly some statistically significant (p<0.01) upward curvature of the dose-response relationship in the range 0-40 cigarettes/day, which is what might be expected if more than one of the stages (in the multistage genesis of bronchial carcinoma) was strongly affected To some extent their conclusions are dependent on the by smoking". extent to which they were justified in omitting results for smokers of more than 40 cigarettes a day from their analysis, since risk in this group was clearly substantially less than predicted from their Some of their reasons for omitting this group from formula. analysis (in which only five lung cancers occurred) have already been discussed.

For a carcinogen continuously applied throughout life, the incidence rate at a given time, t, should, in theory, be proportional to the following function of dose and time

$$I \alpha t_{i=1}^{k} (\alpha_{i} + \beta_{id})$$
(19)

-34-

It should be noted that, as described by e.g. Crump and Howe (1984) it is possible to fit a generalization of this function as follows

$$I \alpha t^{k}(q_{0} + q_{1}d + q_{2}d^{2} + \dots q_{k}d^{k})$$
 (20)

where all the coefficients  $q_i$  are  $\geq 0$ . This model, along with related statistical methods, is routinely used by the EPA and other regulatory agencies to assess low dose cancer risks. It is often referred to as the "multistage model". However formula 20 is actually more general than formula 19, since it contains polynomials not contained in it.

In formulae 19 and 20, the relationships of incidence rate to dose and of incidence rate to time are separable functions which multiply together. Strictly this only applies to continuous exposure throughout life. Where exposure starts at a given point in time, the separability no longer applies, as illustrated by formulae 5 and 6.

Lee (1979) considered a version of the multistage model in which it was assumed that lung cancer was a seven stage multistage process, with smoking only affecting the first and sixth stages. Lee presented a table, reproduced as <u>Table 3</u>, in which relative risk at age 70-74 was related to number of cigarettes smoked under two hypotheses: A - equal effects on stages 1 and 6, and B greater effect on stage 6 than stage 1. Under the column "linear fit" is shown how a straight line going through the dose points 0 and 6 would fit the data. <u>Figure 1</u> (adapted from Lee (1979))
shows that hypothesis B produced a dose-response relationship that is quite close to a linear relationship. In this figure one dose unit from Table 2 has arbitrarily (though not unreasonably in view of the knowledge of the magnitude of relative risk for 20 a day smokers) been taken to be five cigarettes a day. Although inspection of Table 2 shows that hypothesis B fits a linear relationship better than does hypothesis A, it is far from clear that hypothesis A is necessarily ruled out. As Doll and Peto (1978) point out (vide supra) there does appear to be some upward curvature of the dose relationship, and as we have already noted, there are a number of reasons why the observed dose-response may be shallower than the true dose-response. Lee (1979) concluded that it would be difficult to infer reliably from existing data whether late stage effects are stronger than early stage effects. In any event, it is clear that apparent approximate linearity of the dose-response relationship does not exclude the possibility of two stages being affected by the carcinogen, especially when the effects on the transition probabilities, relative to background, may not be very large.

# 3.4 <u>Relationships with stopping exposure</u>

Formulae 8/1, 8/2 and 8/3 relate incidence rate to age T for individuals starting to smoke at age S and then smoking for a duration of D. Using these formulae a number of authors have shown that the rise in incidence with time following stopping depends dramatically on which stages are assumed to be affected. If the <u>first stage</u> only is affected, then for a considerable time after stopping the risk rises nearly as fast as if exposure had been

-36-

continued. This is illustrated in the table below, using formula 8/1 with k = 5, S = 20, d = 10 and D = 20.

	<u>Excess lung_cancer risk (10<sup>4</sup>)</u>		
<u>Age</u>	Continued smoking	Stopped at age 40	
40	160	160	
40 50	810	800	
60	2560	2400	
70	6250	5440	
80	12960	10400	

The relative lack of effect of giving up smoking here results from the fact that most cancers arising come from cells which have undergone their first transition early in life. Giving up after this first transition has occurred has no effect at all on risk of cancer arising from a cell.

If the <u>penultimate stage</u> only is affected, then the effect of stopping is much more dramatic, excess risk not rising at all after stopping, though absolute risk does rise. This is illustrated in the table below, using formula 8/2 - again with k = 5, S = 20, d = 10 and D = 20.

Lung cancer risk (10 <sup>4</sup> )			
<u>Age</u>	Nonsmoker	<u>Stopped at age 40</u>	<u>Excess</u>
40	256	2656	2400
50	625	3025	2400
60	1296	3696	2400
70	2401	4801	2400
80	4096	6496	2400

Compared with the situation where the first stage is affected, where absolute risk after stopping rises from 416 at age 40 to 14496 at age 80 (i.e. by a factor of 34.8), absolute risk only rises by a factor of 2.4 in the situation where only the last stage is affected.

Lee (1979) has investigated how lung cancer risk varies by time since stopping for a multistage model with seven stages where only the first and sixth stages were affected. Taking S = 20 and D = 20and using various assumed values of the two stage effects all of which predicted the same multiplication in risk (25) at age 60-64 for continuous smoking, he showed that provided that the sixth stage was affected at least as much as the first stage there was relatively little increase in risk with giving up smoking for at least 10 years after stopping smoking. Some of his results are reproduced below:

				Risk r	elativ	7e to
Нуро-		<u>Stage e</u>	<u>ffects</u>	<u>risk a</u>	<u>it age</u>	<u>50-54</u>
<u>thesis</u>	<u>Description</u>	<u>1</u>	<u>6</u>	<u>50-54</u>	<u>60-64</u>	<u>70-74</u>
1	Only stage 1 affected	275	1	100	544	2039
2	Stage 1 strongly affected, stage 6 more weakly	25	8.05	100	142	272
3	Both stages affected similarly	12.47	12.47	100	123	191
4	Stage 1 affected less than stage 6, but still quite strongly	5	18.52	100	113	147
5	Stage 1 affected weakly, stage 6 strongly	2	23.01	100	109	132
6	Stage 6 only affected	1	25.03	100	108	126

-38-

There are certain problems in interpreting epidemiological data on ex-smokers since those who give up may be unrepresentative in various ways of those who continue to smoke. <u>Inter alia</u>, those who give up may:

- (a) be less committed smokers, smoking less, inhaling less, smoking lower tar brands and starting to smoke later;
- (b) be more health conscious, a decision to give up smoking being linked to reduced levels of other risk factors; or
- (c) be more unhealthy, illness precipitating the decision to give up.

Nevertheless study of trends in rates after giving up smoking gives useful insight into the validity of the multistage model and clues as to the stages likely to be affected.

Data from the British Doctors Study in relation to ex-smoking have been presented in various papers. Doll (1971) gives a detailed table giving man-years at risk and numbers of deaths by amount last smoked, age stopped and period since stopping, Doll and Peto (1976) give estimates of mortality relative to that in continuing smokers and in lifelong nonsmokers, while Doll (1978) gives graphs showing how absolute incidence in ex-smokers, by years stopped, compares with that in continuing smokers and in lifelong nonsmokers. Doll (1978) summarizes the data as follows:

"The effect of stopping smoking is evident with 5 years. On stopping the rate ceases to increase as it would have if smoking had continued, but whether it actually falls is uncertain because the numbers are small ... The trend, however, suggests a fall followed by an increase, which keeps the rate ahead of that in lifelong nonsmokers".

Compared with continuing smokers, ex-smokers were found to have 35% of the lung cancer rate 5-9 years after stopping and 11% of the lung cancer rate 15+ years after stopping. For those periods after stopping risks relative to lifelong nonsmokers were respectively 5.9 and 2.0 times higher.

The multistage model cannot, of course, predict a declining risk after stopping unless the final stage of the process is affected. However, as Doll notes, a true decline may not have occurred, the slight drop being explained by sampling variation or unrepresentativeness of ex-smokers. Doll's results seem not inconsistent with the multistage model, but clearly require that a late stage be affected to fit. The drop off, relative to continuing smoking, is far too large and rapid to be explained if only an early stage were affected. It will be interesting to see whether, when the 40 year results are published, the apparent approximate freezing of incidence rate on stopping continues for a longer period after stopping. As shown in the calculations above, the multistage model does not actually predict that the rate will stay constant on stopping, only that it will approximately do so for a period.

Kahn (1966) presented detailed tabulations, for smokers of cigarettes only, giving observed numbers of lung cancer deaths and annual death rates per 100,000 per year broken down by age (55-64, 65-74), age of starting to smoke (<15, 15-19, 20-24, 25+),

-40-

maximum number of cigarettes smoked per day (1-9, 10-20, 21-39. and years since cigarette smoking stopped (continuing, 1-4, 40+), 5-9. 10-14 and 15+) based on 8<sup>1/2</sup> years follow-up of the US Veterans Those who had stopped smoking because of "doctor's orders" Study. were excluded from analysis. Given age, it was generally evident that those who had given up smoking for more than 5 years had lower risks than those who continued to smoke, with risk declining with time given up. Smokers of age 65-74 who had given up for 10-14 years had higher risks (258) than those of age 55-64 who continued to smoke (158), suggesting that the absolute risk did not freeze on A limitation of this study is the fact that smoking stopping. habits were only determined at one point in time.

Freedman and Navidi (1987, 1990) describe results of analyses based on a longer follow-up of the US Veterans Study, from 1954/57 to 1969. Again smokers giving up because of doctor's orders are 169 lung cancer deaths in ex-smokers of omitted from analysis. cigarettes only are considered compared to 113 reported by Kahn (1966). Freedman and Navidi compare risk by years of giving up smoking, standardized for amount smoked and age at giving up, i.e. they are testing whether absolute risk freezes on giving up smoking. For years of giving up of 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34 and 35+ the standardized risks (numbers of lung cancers) were respectively 87 (26), 98 (45), 88 (52), 74 (25), 48 (11), 16 (6), 520 (4) and 0 (0). The risks for long-term giving up are based on small numbers of deaths and are difficult to interpret, but the pattern suggests some decline over a 20 year period. Compared to

nonsmokers, the risk declined with increasing time of giving up, with no excess evident by 25 years. Without detailed study of the data, which are not presented so as to allow this, it is unclear why Freedman and Navidi's analysis appears to differ in conclusions from that of Kahn.

Hammond (1966) presents only limited data on ex-smoking from the first American Cancer Society Cancer Prevention Study. For men of age 50-69 who smoked (or had smoked) 20+ cigarettes a day age-standardized death rates for lung cancer were 15 in never smokers, 205 in current smokers and, respectively, 437, 180, 108 and 16 in smokers who had given up for <1, 1-4, 5-9 and 10+ years. Following an initially higher rate for very short term ex-smokers, presumably related to why they gave up, the risk declined until no increase was evident for smokers who had given up for 10+ years. The pattern was similar for ex-smokers of 1-19 cigarettes a day, though less stable, being based on only 10 deaths in ex-smokers as against 93 for ex-smokers of 20+ cigarettes a day.

Freedman and Navidi (1987, 1990) also describe results of analyses based on the first ACS study. Based on five years follow-up and a total of 294 deaths in ex-smokers, they again compared risks by years of giving up smoking standardized for amount smoked and age at giving up. For years since quitting of <1, 1-4, 5-9 and 10+ years the standardized risks (numbers of lung cancers) were 158 (69), 114 (111), 83 (108) and 53 (6). Relative to nonsmokers the risks were estimated as 12.8, 7.8, 3.5 and 0.4. The

-42-

decline in absolute risk, with risk going below that of nonsmokers after 10 years of quitting, are notable features of the data. Freedman and Navidi note that declining excess risk is not compatible with the versions of the multistage model normally considered. They consider various modifications of the model that might help to fit the data better (allowing for variability in waiting times from malignancy to clinical endpoint; allowing for rates of progression through the stages to vary from person to and allowing for individual variation in susceptibility), person; but feel that "a more interesting idea is that the body can repair the lesions caused by smoking, and once the insult stops, the repair process is reasonably fast". They note that repair mechanisms are not compatible with the multistage model in standard form, but point out that the idea is incorporated into the model used by Gaffney and Altshuler (1988). Freedman and Navidi do not, however, consider the possibility of bias due to non-representativeness of ex-smokers.

As described in more detail below (section 5.4), Brown and Chu (1987) found that a multistage model in which the first and penultimate stages were affected by smoking predicted reasonably well the variation seen in the Lubin <u>et al</u> (1984) study in risk of lung cancer in ex-smokers by years since smoking stopped, given age and duration of smoking.

Lubin <u>et al</u> (1984) themselves present some less detailed analysis of these data. One table gives risks of lung cancer by number of years since smoking is stopped (0, 1-4, 5-9,  $\geq 10$ ) and duration of smoking habit (1-19, 20-39, 40-49,  $\geq$ 50). Another table gives risks by sex, number of years since smoking is stopped, and number of cigarettes a day (1-9, 10-19, 20-29, ≥30). There are some obvious limitations in these analyses. Firstly, duration of smoking habit, which is used directly in the first analysis, and as a standardizing variable in the second analysis, is not separated out into fine enough categories. Secondly, age at interview does not appear to have been adjusted for in any analysis. In a study where cases and controls are matched on age, such adjustment is necessary to avoid marked bias in estimating risk by duration. Patterns reported of variation in risk by time of giving up smoking are, however, similar to those described by Brown and Chu (1987) (vide supra).

Halpern et al (1993) presented detailed data based on over 4000 lung cancer deaths occurring over a six year follow-up period in the American Cancer Society Cancer Prevention Study II (see Table 4). The observed patterns were similar in both sexes. For those quitting smoking between ages 30 and 49 lung cancer death rate rose gradually with age at a rate slightly greater than that for those who had never smoked. For those quitting between ages 50 and 64 risk levelled off near to that attained at the time of quitting until around age 75, when it rose sharply. At age 75, compared with the risk for current smokers, relative risks were

-44-

approximately 0.45, 0.20, 0.10 and 0.05 for, respectively, those quitting in their early 60s, those quitting in their early 50s, those quitting in their 30s and those who had never smoked. The authors do not actually fit multistage models to their data, instead fitting a logistic model which contains terms in sex, education, cigarettes per day, years smoked and smoking age, status (and in some cases higher order terms and interactions). They note that the "plateau of risk in the age-at-quitting cohorts covering ages 50-64 is inconsistent with ... the Armitage-Doll multistage model, which predicts continuous increases" without pointing out that various forms of the multistage model predict approximate constancy of risk for a period after stopping. They also note that their results are "inconsistent with the results of Freedman and Navidi (1990) who suggest that the absolute risk declines for about 20 years after cessation of smoking". Looking at it is in fact notable that, in contrast to the data from Table 4. the US Veterans Study and the first ACS study, there appears to be no real evidence at all of a decline in absolute risk following stopping. For example compare the risk in continuing smokers of age 54-58 (156.8) with that of ex-smokers who had given up at ages 55-59 (which is 244.0, 270.5 and 353.6 at, respectively, ages 64-68, 69-73 and 74-80). A similar conclusion can be reached for other ages of stopping.

Sobue <u>et al</u> (1993) describe analyses of data from a Japanese case-control study involving 776 lung cancer cases (553 current and 223 former smokers) and 772 controls (490 current and 282 ex

-45-

smokers) all of whom started to smoke at age 18-22. Risk of lung cancer in ex-smokers according to the number of years given up was compared with that in continuing smokers, separate analyses being conducted for the overlapping age groups 55-64, 60-69, 65-74 and The decline in relative risk was more rapid in the younger 70-79. age groups (e.g. at age 55-64 RRs = 1.00, 0.85, 0.47 and 0.34 for current smokers and smokers giving up for 1-4, 5-9 and 10+ years) than in the older age groups (e.g. at age 70-79 RRs = 1.00, 0.85, 0.49 and 0.50), reflecting the fact that the smoking period as a fraction of total lifetime was greater at younger ages. Based on assumed values of risk by age for nonsmokers and continuing smokers (these could not be assessed directly as cases and controls had been matched on age), the authors used their relative risk estimates to compute estimates of absolute risk by age at cessation, age at admission and years since cessation. The pattern was of a clearly increasing absolute risk after stopping smoking, though to less of an extent than occurs if smoking is continued. In interpreting the results from this study one should note that no adjustment has been made for number of cigarettes smoked. Nor has any attempt been made to exclude patients who gave up smoking for health reasons. Nevertheless the results clearly seem to conflict with those of the studies considered by Freedman and Navidi (1990) which suggested a decline in absolute risk on giving up smoking.

Lee (1974) analyzed the results from a mouse skin painting experiment in which groups of mice were treated with 180 mg/wk cigarette smoke condensate (CSC), with 600 mg/wk Fraction G of CSC,

-46-

or with 36 or 60  $\mu$ g/wk on benzo[a]pyrene for life or for various periods of time ranging from 10 to 50 weeks. Lee compared the tumour incidence observed with that expected under three hypotheses: no effect of stopping; tumour rate remaining constant at the time of stopping painting; and tumour rate remaining constant in weeks after stopping painting. For all types of treatment, it was clear that stopping painting reduced the tumour incidence compared with continuing painting. It was also clear that the tumour rate did not remain constant after stopping, this being evident from the simple observation that the groups painted for only 10 weeks had a zero tumour rate at 10 weeks (and indeed at 30 weeks for CSC and G) and yet had an overall tumour yield far in excess of the untreated controls. In the benzo[a]pyrene treated groups incidence continued to rise after stopping painting but very much less steeply than it would have done had painting been continued. In the CSC and G groups painted for long enough for tumours to be seen before incidence declined somewhat for 20 or 30 weeks after painting, stopping and then rose, eventually markedly exceeding that seen at A multistage model in which the carcinogens the time of stopping. affected at least two stages of the cancer process, one early and fitted the observed results quite well. For all the one late, treatments the fitted effect relative to background was greater for the early stage than for the later stage, this being far more marked for benzo[a]pyrene than for CSC or G. It would be noted that the best fitted models for each treatment generally assumed that there was an effect on the final stage (as well as on other stages). Models in which only the first and penultimate stage were affected

-47-

did not explain the drop-off in incidence observed after stopping in the CSC and G groups. It is interesting to note that for continuous painting best fitted Weibull distributions of the form  $I = b(t - w)^k$ generally fit a positive value for w of about 10 weeks. This is consistent with the observation that, for benzo[a]pyrene, even at very high doses indeed, tumours are never seen before 11 or 12 The general interpretation of the w parameter is the time weeks. taken between the final mutation occurring and the tumour becoming clinically evident, and Lee carried out his model-fitting work under this assumption, i.e. he used the formulae in section 2 to estimate risk at time t + w resulting from exposure occurring up to time t. Lee actually points out that w may arise as the sum of constants  $w_1$  $+ w_2 + w_3 + \ldots$  representing fixed delays between a cell undergoing one mutation and being at risk of the next. He derived formulae for the risk in this more complex situation but never actually fitted due to the extensive and expensive nature of the computing them. involved. Such an extension of the model would seem required to try to reconcile the observation that there is a minimum time below which tumours cannot occur and the observation that risk may decline quickly after stopping.

# 3.5 <u>Variation with age in relative risk associated with exposure</u>

Many epidemiological studies appear to show that the ratio of the risk of lung cancer of a smoker of a fixed number of cigarettes a day to that of a nonsmoker (or to that of a smoker of a different fixed number of cigarettes a day) is approximately invariant of age, and indeed the formula proposed by Doll and Peto (1978) (vide supra)

-48-

predicts exact invariance, with the terms in dose and age completely separable. However, inspection of formulae 6/1-6/3 shows that this simple relationship does not hold exactly. If, for example, one considers formula 6/3, taking  $u_1 = u_2 = 1$ , and  $v_1 = v_2 = d$  for a smoker, and  $v_1 = v_2 = 1$  for a nonsmoker, one can express the ratio of incidences at age T for a smoker (starting to smoke at age S) to that of a nonsmoker of the same age as

$$R = \frac{S^{k-1} + d(T^{k-1} - S^{k-1} - (T-S)^{k-1}) + d^{2}(T-S)^{k-1}}{T^{k-1}}$$
(21)

For S = 20 years, k = 5 and d = 5, for example, one can readily calculate R for various values of T

<u>T</u>	<u>R</u>
50	7.50
60	8.90
70	10.18
80	11.31

The fact that R increases with T is not dependent on the precise values chosen of S, k or d, but is a general property, reflecting the fact that the greater the proportion of time one is exposed ((T-S)/T) the greater the relative risk. The rapidity of the rise in R with increasing age does however depend on which stages are most affected. Lee (1979) presents results of some illustrative calculations for a model in which the first and penultimate stages are affected and in which the relative risk at age 60-64 is assumed constant, the only variation being in the relative contribution of the first and penultimate stage effects ( $v_1$  and  $v_2$ ). Where  $v_1$  is relatively small and  $v_2$  relatively large, the increase in R with increasing age is quite modest, but as  $v_1$  increases and  $v_2$  decreases the increase in R with increasing age becomes relatively steep. This is illustrated by further calculations showing the rise in R with increasing T for S = 20, k = 5, d = 20 using formulae 6/1 (first stage only affected) and 6/2 (penultimate stage only affected)

<u>First stage affected</u>	Penultimate stage affected
3.46	19.51
4.75	19.77
5.95	19.87
7.01	19.93
	First stage affected 3.46 4.75 5.95 7.01

There is rather little published data showing how the relative risk for smokers/nonsmokers varies with increasing age. Hammond (1966) did observe some increase, with relative risks of 7.17 at age 35-54, 9.84 at age 55-69, and 10.67 at age 70-84, but Kahn (1966) did not, with relative risks of 11.30 at ages 55-64, and 7.03 at ages 65-74. However considerable sampling variation (due especially to relatively small numbers of lung cancer deaths among younger subjects) and failure to standardize for smoking duration (at that time the older men would certainly have tended to start smoking later than the younger men) makes these results difficult to interpret. The findings certainly do not seem inconsistent with the predictions of the multistage model, but they may be inconsistent with versions of the model in which the main effect of cigarette smoking arises from an early stage.

- 50 -

### 3.6 <u>Effects of joint exposures</u>

For continuous exposure to two agents, the joint dose response relationship will be very different depending on whether the agents affect the same or different stages of the cancer process. If the agents affected the same stage then the relationship should be additive, with the effect of a dose x of one agent being interchangeable with the effect of a dose y of the other, the ratio x/y reflecting the relative effectiveness of the different agents. If the agents affect different stages, however, the joint dose response should have a multiplicative component, the relationship becoming more multiplicative with higher doses as background effects become relatively weaker.

Evidence in favour of there being more than two stages comes from a number of studies which have shown multiplicative (or at least super additive) relationships between incidence and exposure to two agents. Selikoff and Hammond (1975) have reviewed some of the evidence on multiple risk factors in environmental cancer. Factors which show evidence of a multiplicative relationship with lung cancer include smoking and uranium mining, smoking and exposure to radiation from atomic bombs, and smoking and asbestos. The evidence for smoking and asbestos exposure is quite strong, with Hammond et al (1979) reporting lung cancer relative risks of 1, 5.2, 10.9 and 53.2 for exposure to, respectively, neither asbestos nor smoking, smoking only, or both asbestos and smoking (though asbestos only, small numbers of deaths in the group exposed to neither asbestos nor smoking may mean the apparent very multiplicative relationship was to some extent a chance finding). It would be interesting to see multistage models fitted to detailed joint exposure data but I am not aware that this has been attempted. One reason may be the lack of large studies providing detailed data on level, time of start and time of cessation of exposure.

Although, as noted below (see section 4.1), there is good animal evidence for some combinations of exposures that agent A followed by agent B elicits far more tumours than agent B followed by agent A, there appears to be little or no relevant epidemiological evidence here. Peto (1984) in fact notes that the initiation/promotion phenomenon has never actually been observed directly in human carcinogenesis.

### 3.7 Effect of changing the type of cigarette smoked

Lee (1993a) recently reviewed the available epidemiological evidence relating risk of lung cancer to type of cigarette smoked. Although evidence relating to smoking cigarettes of tar 12 mg or less is still very sparse, there is quite substantial evidence that switching from plain to filter cigarettes or from higher to lower tar cigarettes is associated with some reduction in risk of lung cancer. Of 38 relative risk estimates associated with tar reduction or the plain/filter switch, 32 are less than 1.0, with the median 0.65. The fact that an apparent reduction in risk has been seen, despite the fact that in many studies smoking of the filter or lower tar cigarettes has only been for a relatively short period, is consistent with other evidence that smoking affects a late stage of

-52-

the cancer process. As far as I am aware, however, no-one appears to have carried out formal multistage model fitting to such data.

# 3.8 <u>Relationship of dose to age of onset of exposure</u>

Passey (1962) noted that in a sample of hospital patients, age of onset of lung cancer appeared to be the same almost irrespective of their daily cigarette consumption, and argued that this provided evidence that cigarette smoke does not act as a carcinogen. That this line of reasoning was wrong was made clear by Pike and Doll (1965) in a paper which emphasized how misleading a statistic average age at onset of a disease may be. While it is true that in animal experiments involving different doses of a strong carcinogen (which causes cancer in all or virtually all the exposed animals) increasing dose will lead to decreasing average age of tumour onset, this is not so for a weak carcinogen which leaves overall survival of the exposed population materially unaffected. If the function relating incidence rate to dose and time can be separated into terms dependent on dose and terms dependent on time, and the overall survivorship is similar in the various dose groups, it is apparent that the distribution of time of onset will be essentially Separability of dose and time is a independent of dose. characteristic of the Weibull expression  $I = bd^{c}t^{k}$  and similarity of average age of onset in different dose groups is therefore consistent with this. In fact, two additional points which act in opposite directions need to be taken into account. The first is especially at higher ages, the proportion of heavy smokers that. surviving will be less than the proportion of lighter smokers,

leading to some reduction in age of onset with increasing dose. The second is that, using a proper multistage formulation, and not the Weibull approximation, relative risk for heavy to light smokers increases with increasing age (see section 3.5), leading to some increase in age of onset with increasing dose. It should also be realized that variation in age distribution between heavy and light smokers and variation in age in the difference in mean age of starting to smoke between heavy and light smokers may upset any simple relationship.

Generally approximate similarity of mean age of onset of lung cancer in smokers of differing amounts is broadly consistent with the predictions of a multistage model, but the statistic is a difficult one to interpret and its use should be avoided if possible.

# 3.9 Other issues

Gaffney and Altshuler (1988) point out that, assuming a multistage model with the first and penultimate stages affected, the relative risk of heavy and lighter smokers will increase with increasing duration. Based on a best fit (six stage) to the Doll and Peto (1978) British Doctors data they point out that the relative risk comparing two packs a day and one pack a day smokers should increase from 2.5 at age 42.5 (smoking for 20 years) to 3.3 at age 72.5 (smoking for 50 years). In fact they noted that this prediction was not supported by the data. For smokers of, respectively, 17.5-27.5, 27.5-37.5, 37.5-47.5 and 47.5-57.5 years

-54-

the relative risk of smokers of 25-40 cigarettes a day compared with smokers of 10-24 cigarettes a day was 2.5, 2.2, 2.5 and 1.6, i.e. there was no evidence of an increase in relative risk and indeed, in the highest duration category, some evidence of a decrease.

### 4. <u>LIMITATIONS OF THE MULTISTAGE MODEL</u>

### 4.1 <u>Stages undefined</u>

One obvious limitation of the multistage model is that it assumes that a number of stages must occur before the onset of cancer, but does not given any direct indication of what the stages might be. Although no clear evidence of what all the stages are has yet emerged (if indeed there are such stages and the model is not just a convenient mathematical approximation), there has been direct evidence for a long time that there are sequential aspects to It is over 50 years since it was demonstrated that carcinogenesis. the cocarcinogen croton oil was found capable of enhancing skin tumour induction when applied after a subeffective dose of carcinogenic hydrocarbon but not when applied beforehand. Such so-called "initiation/promotion" experiments led to the idea of "the See Berenblum (1982) for a comprehensive two-stage hypothesis". review of the evidence relating to sequential aspects of chemical carcinogenesis in the skin, where much of the work has been conducted. It is interesting to note that for many years it was unclear whether cocarcinogens of tumour promoter type were actually Recent observations by Hecker (1984) in the relevant to man. Caribbean island of Curaçao are of particular interest here. On this island the black and Creole population have an extremely high rate

-55-

of oesophageal cancer and, as part of the local diet, the fresh green leaves of the aromatic bush known as "welensali" are commonly used to prepare a "bush tea". One cup of tea prepared from this bush, <u>Croton flavens</u> L, contains very high levels indeed of known tumour promoters, and Hecker makes a strong case for this being responsible for the high oesophageal cancer rate.

It is possible that molecular genetic studies may help to identify the stages required for tumorigenesis. Renan (1993), in a paper attempting to answer the question as to how many mutations are required, notes that "molecular studies have strongly supported the idea that multiple genetic changes are required". He cites the example of colorectal malignancies, "which involve genetic alterations on chromosomes 5q, 12q, 18q and 17p and possibly other lesions as well".

# 4.2 <u>Reversibility of effects may occur</u>

As specified, the multistage model does not allow for reversibility of any of the stages. Over time the numbers of cells that have passed through the various stages can only increase. Conceivably, for some stages at least, damage may be repaired. Though, for continuous exposure, taking the possibility of reversibility into account should not affect the mathematical approximations (the transition probabilities can be viewed as differences between probability of damage minus probability of repair), this need not be the case for discontinuous exposure. Clear evidence that incidence declines in absolute terms after stopping

-56-

would suggest reversibility and indicate the assumptions behind the multistage model are too simplistic.

# 4.3 <u>Transition probabilities may vary from individual to individual for</u> <u>a given exposure</u>

For a given exposure it is assumed by the multistage model that the transition probabilities for each stage do not vary from individual to individual. For a disease with a large genetic component this may be an inappropriate assumption. If the population actually consists of two groups of individuals, a susceptible group with non-zero transition probabilities for each stage, and a non-susceptible group with zero transition probabilities for one or more stages, then it is easy to see that one will not observe the simple relationship between incidence rate and age (formula 2) predicted for continuous exposure. Rather the incidence rate, instead of rising continuously with age, will fall off past a given point in time as the susceptibles are depleted, perhaps eventually reaching zero when only non-susceptibles remain. Sellers et al (1990), using segregation analysis, reported finding that lung cancer patients could be divided into three groups, one with a much higher risk of early onset disease (given smoking habits and occupation) than the other. This suggestion of a genetic component is supported by evidence (summarized by Lee, 1993b) that family history of lung cancer is an independent risk factor for lung The extent to which such genetic variation will modify cancer. predictions from the multistage model is not clear at this point in time.

In their analyses relating incidence of cancer (I) of 31 types in 11 populations to age (t), Cook <u>et al</u> (1969) found that in 54% of cases there was evidence of downward curvature from the theoretical straight line relationship predicted by the Weibull formula  $\log_e I =$  $\log_e b + k \log_c t$ . One possible explanation that they considered for this (apart from underdiagnosis in old age or differences in exposure between different age-cohorts) was that only a proportion of the population might be susceptible to cancer. If the initial proportion of susceptibles is C, it can be shown that instead of the simple relationship given above, the relationship will be of the form

$$\log_{e} I = b + k \log_{e} t - \log_{e} [C + (1 - C)e^{F/C}]$$
(22)  
where F =  $e^{a}t^{k+1}/(k+1)$ .

They presented a graph showing that the extent of downward curvature is very small indeed for C even as low as 0.1 or 0.05. Only for C = 0.01 did substantial downward curvature occur with incidence falling off after age 60. They pointed out that if susceptibility were the explanation for the downward curvature one would expect to see an increased amount of curvature with increasing levels of incidence in genetically similar populations. However the data did not appear to support this. They concluded that there was "no evidence ..... to suggest that the shape of the observed relationship could be attributed to attenuation of a limited pool of susceptibles".

Peto et al (1985) cite data of Parish (1981) to support the idea that there is considerable variation among outbred mice in susceptibility skin cancer induced by chronic their to benzo[a]pyrene treatment. A figure was presented comparing the new tumour incidence rate/time relationship of mice who had respectively 0, 1, 2 or 3 tumours already. There was a clear tendency for incidence at a given time to increase with the number of tumours already present, and for the log incidence/log(duration of exposure - 15) relationships to show downward curvature from a straight line. Peto et al note that their results are consistent with substantial heterogeneity of susceptibility with risk varying 100-fold between the upper and lower 95% extremes of the distribution. As they note, the more susceptible an animal is, the more tumours it is likely to thus explaining the higher risk with increasing have already, numbers of tumours present. They also note that failure to take variation in susceptibility will lead to into account underestimation of the true number of stages of the cancer process. Elsewhere, Doll (1978) makes it clear that substantial variation in susceptibility is not inconsistent with relatively small differences in risk associated with family history of cancer. Consider, for a recessive gene that increases the risk of a particular example, cancer 50-fold in homozygotes. The relative risk in the siblings of probands would then be just over 4-fold if the population frequency of the gene was approximately 10%.

- 59 -

One possibility apparently not considered in the literature is that, within an individual, all the cells capable of being transformed to cancer of a particular type may not be equally susceptible.

# 4.4 <u>The model may be inaccurate if the transition probabilities are not</u> <u>small</u>

Consider a two stage process in which both transition probabilities are equal, having the value a. The probability,  $1-G_T^*$ , of a cell surviving tumour free at time T is then given by the expression:

$$1-G_{T}^{*} = e^{-aT} + 2 \int_{0}^{T} ae^{-au} e^{-a(T-a)} du \qquad (23/1)$$

$$= e^{-aT}(1+aT)$$
 (23/2)

The probability,  $1-G_T$ , of the organism, with N cells, surviving tumour free at time T is then given by:

$$1 - G_{T} = (1 - G_{T}^{*})^{N}$$
 (24)

The incidence rate of cancer at time T,  $I_{T}$ , is then given by:

$$I_{T} = \frac{dG/dT}{1-G} = \frac{Na^{2}T}{1+aT}$$
(25)

This compares with the standard approximate form of the incidence rate given by formula 1 in section 2, of:

$$I = Na^2 T$$
 (26)

The exact form of the incidence rate would show some downward curvature when log I is plotted against log t, whereas the approximate form would not. This would also be true for the more general situation of a k stage process, with differing transition probabilities from stage to stage (see Hakama (1971) for the more general exact formulae).

The question arises as to how adequate the approximate form of the incidence rate formula actually is. In discussion on Hakama (1971),Moolgavkar (1977) noted the approximate Armitage-Doll formula can be viewed as the first term in an infinite (Taylor) series expansion of the solution, and that retention of additional terms in the power series would give a better approximation and might explain some of the deviations from the theoretical incidence curve noted by Cook et al (1969). Peto and Doll (1977) and Hakama (1977), in reply to Moolgavkar's letter, point out that in practice the Armitage-Doll approximation is extremely good, and that downward curvature in the lung cancer incidence rate curve is much more likely to result from underdiagnosis of lung cancer in the elderly, from cohort effects or from selective mortality, than it is to result from a poor approximation of the formula.

This can be illustrated by considering the two stage process above. Suppose we consider incidence at age 70. The annual incidence rate of lung cancer will not exceed 1 in 100. Given a fairly conservative number of cells at risk of 10,000, one can readily calculate that the annual transition probability per cell is about  $1.2 \times 10^{-4}$ . The difference between 1+aT = 1.008 and 1 is really then quite small compared with other sources of variation. A similar conclusion can be reached using higher numbers of stages.

-61-

The approximateness of the formula does not seem to be a problem in practice.

#### 4.5 Other problems

As noted above, genetic heterogeneity may have the effect of altering the observed power of time, so that evidence of a kth power relationship between incidence and time (or duration of exposure) does not necessarily imply there are k+1 stages of cancer. Peto (1984) notes that other factors, including selective proliferation and diagnostic delay may also have this effect by altering the observed power of time.

Although the multistage model has been expressed in terms of mutations occurring since birth, it is possible that cancer may arise in individuals who are born with one (or more) of the mutations already present. See for example the retinoblastoma model proposed by Knudson (1971).

In his paper on multistage models, Peto (1977) points out that though they "hold out the most promise of being a useful framework for describing the process of neoplastic transformation, there are various observations which do not appear to fit naturally into the multistage formulation". These include:

 (i) The fact that given age and dose of carcinogen, an animal is more likely to get a tumour if it already has a tumour of the same type than if it does not;

(ii) The existence of tumours of mixed cellularity; and

-62-

(iii) The fact that when mutagens are applied to cells <u>in vitro</u> it is much easier to cause neoplastic transformation than it is to cause gene mutation.

For all the problems, and a discussion, the interested reader should refer to Peto (1977).

# 5. APPLICATIONS OF THE MULTISTAGE MODEL

# 5.1 Using data on prevalence of smoking at different ages

Section 2 gives formulae, based on the multistage model, for one continuous period of smoking (formulae 7 and 8 and for two continuous period of smoking (formula 10). Formulae can also readily be derived for multiple periods. In cohort (or case-control) studies, where data are available on an individual basis concerning a person's lifetime smoking history, these formulae can be derived directly. However a number of coworkers have attempted to fit multistage (or other) models to national age-specific lung cancer incidence data where the only data available are cohort-specific percentages of smokers each year or each five years (sometimes accompanied by data on average consumption levels).

In order to convert these percentages into estimates of the frequency of people smoking for different periods of time (and hence use the multistage model formulae) it is necessary to make some assumptions. For example, if there were two time periods with 30% smokers in the first and 40% in the second there are various possibilities, including:

- (i) 30% smoking throughout, 10% smoking only in the second period.
- (ii) 30% smoking only in the first period, 40% smoking only in the second.
- (iii) 20% smoking throughout, 10% smoking only in the first period,20% smoking only in the second.

The first possibility maximizes the proportion of long duration smokers, the second minimizes it. The third is one of many intermediate possibilities.

When attempting to get round this problem, Townsend (1978) assumed that smokers can be ordered from "hard core" to "highly capricious", so that the frequency of longer duration smokers is maximized. If, for example, the percentages of smokers at six successive time periods are 20, 30, 45, 40, 50 and 35, one can divide the population into 20% smoking throughout, 10% (= 30-20) smoking at all times except in the first period, 5% (= 35-30) smoking at all times except the first, second and sixth, and so on.

An alternative approach was used by Swartz (1992). Here it was assumed that smokers, once they give up, never start again. If, for example, the percentages of smokers at six successive time periods are 10, 20, 10, 20, 10, 20, Swartz would assume there are four groups of people, 10% who smoke throughout, 10% who smoke only in period 2, 10% who smoke only in period 4, and 10% who smoke only in period 6. This contrasts with Townsend's assumptions, which would

-64-

involve only two groups, 10% who smoke throughout and 10% who smoke only in periods 2, 4 and 6. Hakulinen and Pukkala (1981) appear to make similar assumptions to Swartz. It should be noted that the Swartz assumption may, with certain data, lead to more than 100% of the subjects being classified into smoking groups.

In theory it would be possible to investigate the validity of either approach using data from a study in which detailed lifetime smoking histories were collected, but no such investigation appears to have been carried out. On general grounds it seems that both approaches are likely to be incorrect, the first probably overestimating risk, the second probably underestimating it.

# 5.2 Applications to cohort data

Mazumdar et al (1991) describe techniques for fitting to cohort data multistage models dose-related stages. with two Their methodology and software allow for exposure to vary over intervals during the person's life as may be needed for occupational mortality The method is illustrated studies with detailed exposure data. using lung cancer mortality data for a cohort of non-white male coke oven workers exposed to coal tar pitch volatiles and shown to fit adequately. This group at the University of Pittsburgh are extending their software to fit alternative models proposed by Moolgavkar and his colleagues. Those intending to do detailed fitting of such complex data would do well to approach the authors, though note that the computing was done on a CRAY Super Computer!

-65-

### 5.3 <u>Whittemore (1988)</u>

Whittemore (1988) used data from three sources to test the fit of two functions relating lung cancer incidence to smoking habits. The first two sources, the British Doctors Study (Doll and Peto, 1978) and the US Veterans Study (Kahn, 1966), presented data on risk for current smokers and for lifelong nonsmokers. The third source, a case-control study of non-Hispanic white men in New Mexico, data for which were provided by Prof J Samet, had detailed lifetime smoking histories, and so provided a more rigorous test. The first function used, the packs function  $g_1$ , specified that the excess death rate at age t depended linearly on the cumulative amount smoked

$$g_1 = 2.01 \times 10^{-12} (t - 5)^{4.5} (1 + \alpha P)$$
 (27)

where P is the total number of packs of cigarettes smoked by age (t - 5) and  $\alpha$  is a constant to be specified. The second function used, the multistage function  $g_2$ , specified that the death rate at age t is of the form

$$g_{2} = 2.01 \times 10^{-12} [(t - 5)^{4.5} + pc(1 + 2pc)(t_{1} - t_{0})^{4.5}$$
(28)  
+ 2pc(t\_{1}^{4.5} - t\_{0}^{4.5})] (28)

where c is the number of cigarettes per day and p is a constant to be specified.

Whittemore found that both functions fitted the British Doctors data with best-fitting parameters  $\alpha = 1.13 \times 10^{-3}$  and p = 0.207, there being little to choose between the functions.

With the US Veterans' data, best-fitting parameters were lower,  $\alpha = 0.59 \times 10^{-3}$  and p = 0.128, but neither function fitted the data very adequately, there being a notable tendency to overestimate risk at age 65-74 (624 deaths expected vs. 576 observed using  $g_2$ ), and to underestimate it at age 55-64 (477 E vs. 547 0). For the New Mexico data,  $g_2$  fitted markedly better than  $g_1$ . However there was some tendency for  $g_2$  to overestimate risk in ex-smokers (68 E vs. 45 0) and to underestimate it in current smokers (166 E vs. 179 0). Both functions, however, explained substantially more variation in the New Mexico data than did any of several logistic regression models involving categorical variables for age and smoking.

Some points to note about this work are as follows:

- (i) The function g<sub>1</sub> is stated to indicate excess risk. However as it is not zero for P = 0 it presumably was actually intended to indicate actual risk. The function is in any case not of a form predicted by the multistage model.
- (ii) The function  $g_2$ , stated to be based on a multistage model in which the first and penultimate stages are affected, the penultimate stage being twice as strongly affected as the first, is actually incorrectly derived (or has been misreported). As noted elsewhere (see section 2), the term  $pc(t_1 t_0)^{4.5}$  should be replaced by  $pc[(t t_0)^{4.5} (t t_1)^{4.5}]$ . This does not affect the fit for continuous

exposure, where  $t_1 = t$ , but gives different predictions for ex-smokers. The fit to the New Mexico data will therefore be in error.

- (iii) The nonsmoker part of the function,  $2.01 \times 10^{-12} (t 5)^{4.5}$ , was based on a fit to nonsmokers' data from the American Cancer Society CPS I study. Since these subjects are unrepresentative, and since there are a multitude of risk factors in nonsmokers, this function may not be fully appropriate for other data. It is surprising that Whittemore apparently did not at least try the effect of fitting constants other than 2.01.
- (iv) When fitting the New Mexico data, Whittemore tried using  $\alpha$  and p values fitted to either the British Doctors data or the US Veterans data. The values for the US Veterans study fitted much better and were used in her main work. It was surprising that Whittemore did not try to determine the parameter values which best fitted the New Mexico data.
- (v) Commenting on the lack of fit of the models to the US Veterans' data, Whittemore notes that this may be due to inadequate smoking data. Numbers smoked were determined only at the start of the study and may have changed both before and after.

# 5.4 Brown and Chu (1987)

Brown and Chu (1987) carried out detailed analyses relating cigarette smoking to lung cancer based on the large multicentre West European prospective study of Lubin <u>et al</u> (1984) involving 6920 male

-68-

patients and 13460 male controls. They compared the risk of lung cancer in smokers who had given up for 3, 4, 5-6, 7-8, 9-11, 12-15, 16-20, 21-26 or 27+ years of smoking with those who had continued to smoke (including those who had given up for 1 or 2 years in this group), after adjustment for reason for quitting, study area, age at interview. number of cigarettes smoked, duration of smoking, frequency of inhalation, and percent of time smoking nonfiltered cigarettes. The pattern of relative risks, 0.99, 0.78, 0.71, 0.69, 0.48, 0.47, 0.39, 0.44 and 0.40 for the nine ex-smoking groups, was shown by the authors to be quite well predicted by a multistage model in which the penultimate stage only was affected, and somewhat better predicted by a model in which both the first and penultimate stages were affected, the latter predicting a flattening out and eventual slight increase in the relative risk many years after The authors emphasized the importance of giving up smoking. adjustment for duration of smoking in their analyses. Had no adjustment been made, the fitted pattern of decline in relative risk with years given up smoking would have been much steeper, declining to 0.17 after 27+ years. Two features of the study design should be noted. One feature is the very large number of deaths, which means that the relative risk estimates have small sampling error (e.g. the estimate of 0.69 for having given up 7-8 years has 95% confidence limits of (0.56 - 0.84). The other feature is the fact that cases and controls were age matched. This means that comparisons cannot be made of risk of subjects in different age groups, so that one cannot compare risk in ex-smokers with that in smokers at the time they gave up.

Brown and Chu also carried out analyses relating risk in smokers who had started to smoke at ages  $\leq 14$ , 15, 16, 17, 18, 19-20 and  $\geq 21$  with that in nonsmokers after adjustment for study area, age at interview, number of cigarettes smoked, frequency of inhalation and percent of time smoking nonfiltered cigarettes. The relative risks in general showed a declining pattern with increasing age of start (3.6, 4.1, 4.0, 4.0, 3.6, 3.4, 2.9 - 95% confidence limits are about  $\pm 0.8$  on each estimate) with the exception of the group starting at age  $\leq 14$ . The pattern of decline was found by the authors to be much better fitted by a multistage model in which the first and penultimate stages were affected than by models in which only the first, or only the penultimate stage was affected.

The authors also fitted the overall data to try to determine the relative effect of smoking on the first and penultimate stages, for smokers of 1-10, 11-20, 21-30 and 31+ cigarettes per day. The best fit values for all four smoking categories were found to indicate a higher penultimate stage than first stage effect (2.8 vs. 0.7 for 1-10 cigs/day, 5.0 vs. 2.5 for 11-20, 6.3 vs. 3.5 for 21-30, and 7.0 vs. 4.0 for 31+). On average smoking appeared to have about twice the effect per unit dose on the penultimate stage than on the first stage. This work was the basis of the assumption used by Whittemore (1988) that smoking had twice the effect on the penultimate stage that it had on the first stage. Especially as the various relationships seen were found to be consistent over subsets

-70-

of the data by age, duration of smoking and number of cigarettes smoked, the results appear to provide quite strong support for the multistage model.

# 5.5 Other authors

Brown and Chu (1983a,b) analyzed the incidence of lung cancer during the period 1938 to 1973 in a cohort of men occupationally exposed to arsenic and other contaminants. After adjustment for duration of exposure they found a clear tendency for risk to of starting employment. increase with increasing age They interpreted their findings as indicating that arsenic appeared to exert a definite effect on a late stage of the carcinogenic process, although their analyses could not conclusively rule out a possible additional effect on the initial stage. The data were found to be adequately fitted by a multistage model in which occupational exposure affected the penultimate stage. No data were available for cigarette smoking on this cohort, but evidence from other studies was cited by the authors in support of the view that this would not materially have biassed the results.

Day (1984) is a review paper demonstrating that a wide range of epidemiological phenomena can be described in terms of simple multistage models of carcinogenesis. He notes "the relationship of cancer risk with the different time variables considered corresponds closely with the behaviour predicted by theories of multistage process. Furthermore, the different behaviour associated with different agents enables one to attempt some classification as to

-71-
how an agent is acting". Day considers evidence inter alia on asbestos and mesothelioma and lung cancer, on ionizing radiation and cancer of various sites, on arsenic and lung cancer, on nickel and nasal sinus cancer, on chloromethylethers and lung cancer, on various risk factors for breast cancer, and on exogenous oestrogen exposure and endometrial cancer. The last is interesting in that it is the only well documented occasion in cancer epidemiology of a <u>last</u> stage agent, absolute excess risk disappearing after exposure stops.

An earlier review paper, reaching similar conclusions, is that by Day and Brown (1980). Included in this paper are some analyses of the Tobacco Research Council Stopping painting experiment, from which they concluded that Fraction G of smoke condensate T57 behaved like a carcinogen affecting predominantly a late stage, in contrast to benzo[a]pyrene which behaved more like a carcinogen predominantly affecting an early stage. These conclusions are not dissimilar from those by Lee (1974) described in section 3.4.

#### 6. MODIFIED VERSIONS OF THE MULTISTAGE MODEL

Some authors have attempted to fit models based on the multistage model but using formulae not actually predicted by it.

#### 6.1 <u>Doll and Peto (1978)</u>

Doll and Peto (1978) fitted the function

$$I = 0.273 \times 10^{-12} (cigarettes/day + 6)^{2} (age - 22.5)^{4.5}$$
(29)

-72-

to 20 year follow-up data from the British Doctors, restricting attention to men aged 40-79, and to lifelong nonsmokers or to subjects who reported same amount of 40 or less per day at each interview. The fit was found to be adequate, but it should be realized that the functional form is not strictly multistage (it should contain terms in duration<sup>k</sup> and in age<sup>k</sup>), although it may be a fairly close approximation. The issues relating to exclusion of subjects smoking more than 40 cigarettes per day and of subjects aged 80+, justified by Doll and Peto at length in their paper, have already been discussed. One limitation of the British Doctors study is that it contains no data on age of starting to smoke.

#### 6.2 <u>Townsend (1978)</u>

Another attempt to use a function related to the multistage model, but not actually predicted by it is that by Townsend (1978). Her model, described in detail in the original paper, was expressed in terms of the sum of three components:

- (a) a product of a length of smoking effect and a level of smoking effect for cigarette smokers,
- (b) a similar product for smokers of other products, and
- (c) an effect for nonsmokers.

The length of smoking effect was of the form

$$\sum_{i} (e_{i} z_{i}^{k}) / \sum_{i} e_{i}$$
(30)

the population being divided into i groups of smokers with frequency  $e_i$  who had smoked for duration  $z_i$ .

-73-

The level of smoking effect was of the form

$$\sum_{t} ((t - w)^{\beta} e_{t} l_{t} f_{t}) / \sum_{t} (t - w)^{\beta}$$
(31)

where t is age, w is age of starting to smoke, and  $e_t$ ,  $l_t$  and  $f_t$  are respectively the values at time t of the proportion of smokers, the number smoked and a cigarette effect parameter (depending on weight of tobacco, tar content and plain/filter status). The function is a weighted mean of smoking levels at each age, the weight (t - w)<sup> $\beta$ </sup> indicating the importance of recent relative to past smoking, recent smoking being more important for  $\beta > 0$ .

Using national annual age and sex specific data on percentage of smokers, generated partly by extrapolation, and other data on type of cigarette, Townsend fitted the model to England and Wales lung cancer data from 1935 to 1970 by five-year age and time periods. The model tended to overestimate rates for 1935-1945 and to fit male data much better than female data. Even after putting in terms to account for likely greater underdiagnosis of lung cancer, the model did not fit the data well for females, predicting downturns in mortality at higher ages in the latter half of the period that were not seen.

The model, although intended to be based on multistage principles, is clearly not a true multistage model. <u>Inter alia</u>, the effects of length and of level of smoking are not separable, and the

-74-

effects of cigarette smoking, cigar/pipe smoking and nonsmoking are not independent. There are also problems with the extrapolated smoking data, detailed surveys only being carried out annually from 1948. This work does not really add to any conclusions regarding adequacy of the multistage model.

#### 7. DISCUSSION AND CONCLUSIONS REGARDING THE MULTISTAGE MODEL

As a mathematical model for describing variation in lung cancer incidence rate by age, dose and duration of exposure, there is no doubt that the multistage model has proved useful and popular. Certainly its properties have been more widely discussed and are more widely understood than any of the other models which we will consider in a later document. The multistage model has a lot going for it: it is flexible, reasonably tractable, and in broad terms its predictions fit in with a number of observed facts. These include:

- (i) the approximate power law relationship of incidence with duration of exposure when exposure is continuous;
- (ii) evidence that age <u>per</u> <u>se</u> does not affect incidence of many cancers;
- (iii) direct evidence from initiation/promotion studies that some cancers require multiple exposures in a specific order for cancer to arise;
- (iv) the observation that tumour incidence may be increased as a result of exposure that has long since ceased;
- (v) evidence of quadratic dose-response relationships for some carcinogens;
- (vi) explaining why the joint effect of two carcinogens is often

-75-

multiplicative, or at least markedly super-additive; and (vii) describing reasonably well patterns of incidence following cessation of exposure.

It would be asking too much of any model to describe adequately all aspects of the variations seen in lung cancer incidence rate. Even in a carefully controlled animal experiment in which precisely defined doses are given at predetermined points in time and animals are randomized to different groups there will inevitably be some sources of variation that will not be completely accounted for. Animals and cells within animals are unlikely to be totally homogeneous in susceptibility for example, so that the multistage similarly exposed animal is effectively assumption that each identical, containing an identical number of identical cells, can at best only be an approximation to reality. That, however, need not be an important limitation if models are seen in the light in which they are put forward, namely as a means of approximately explaining known facts and of making reasonable approximate predictions.

In judging the usefulness of a model, one has to consider whether its predictions materially break down in any circumstances. Much of the testing of the multistage model has been carried out on data from epidemiological studies, and it is important to be aware that such data are limited in a number of ways. These include:

(i) inaccuracy of diagnosis of disease;

(ii) inaccurate quantification of average extent of exposure;(iii) inadequate details on changes in exposure;

-76-

(iv) inadequate information on other causes of the disease which may confound the smoking/lung cancer relationship. In this respect it is important to realize that nonsmokers, light smokers, heavy smokers and ex-smokers are not randomly selected and are likely to be systematically different in many respects. Comparison of ex-smokers with continuing smokers is a particular problem in this respect, since the decision to give up smoking may be related to several factors (including illness and increased health awareness) that are themselves linked to risk of disease.

difficulties Bearing in mind these in interpreting epidemiological data, are there any features of the smoking/lung cancer data that the multistage model notably fails to predict? Certainly, providing it is assumed that smoking affects two distinct probably the first and penultimate stage, stages of the process, the multistage model does not in general do too badly. There are, however, three aspects of the data where it appears that it may have some difficulty.

The first of these is the dose-response relationship, some studies indicating an apparent linear relationship of incidence with number of cigarettes smoked when the requirement for smoking to affect early and late stages of the process (needed to explain relationships of incidence to age at starting to smoke and to time since stopping smoking) would suggest a quadratic relationship. When one bears in mind that a multistage model with two stages moderately affected only actually predicts a relationship that has only a modest quadratic component, and when one realizes that inaccuracies in measuring exposure are likely to reduce the slope of the dose-response relationship, it is not at all clear that this objection undermines the validity of the model. The evidence presented by Doll and Peto (1978) based on the British Doctors data and the arguments they put forward can be seen as a reasonable defence of the model.

A second apparent difficulty of the multistage model that has been referred to is the fact that in the British Doctors data there is no evident tendency for the ratio of risks of heavy to light smokers to increase with increasing age. Gaffney and Altshuler (1988) draw attention to this, pointing out that an increase with age in this ratio would be predicted by the multistage model. Bearing in mind the following facts:

(i) the predicted rise is not very large anyway;

(ii) the data on number smoked may not be completely reliable;

- (iii) ability to smoke a large number in an old man may be an indicator of reasonable health (put another way, symptomatic smokers may cut down); and
- (iv) the lack of data in the Doctors study on age of starting to smoke;

I would not regard this point as a major one. It would be valuable, however, to see additional analyses from other studies to try to confirm whether in fact the overall evidence does or does not indicate a rise in relative risk with increasing age.

The final, and most serious, apparent difficulty relates to the data on giving up smoking. Under a multistage hypothesis in which any stages are affected except the last, the incidence rate of lung cancer will continue to increase on giving up smoking, though the slope of the increase will depend dramatically on which stages are As shown in section 3.4, the rise will be much greater affected. if the first stage is affected than if the penultimate stage is affected. Even if both the first and penultimate stages are only relatively modest for some affected the rise be may considerable time, provided the penultimate stage is affected more than the first stage.

A decline in absolute risk can occur if the last stage is affected, but this will be immediate and not a gradual decline. Freedman and Navidi (1990) have claimed that the epidemiological evidence indicates that absolute risk of lung cancer declines on giving up smoking and that this is inconsistent with the predictions of the multistage model. Gaffney and Altshuler (1988) have also argued that the multistage model is inadequate because it cannot simultaneously fit the incidence in smokers and ex-smokers. They argue that the best fit to the data for continuing smokers predicts that excess incidence will greatly increase in ex-smokers whereas the data indicate no change or a decrease.

In interpreting this evidence a number of important points should be made:

-79-

- (i) Freedman and Navidi, and Gaffney and Altshuler, pay little attention to the problems of bias caused by the non-representativeness of ex-smokers. Some studies, but not all, attempt to get round the bias due to some smokers giving up because of severe illness. If ignored, this might give the false impression that giving up smoking markedly increases risk of lung cancer in the short term. More difficult to adjust for is the bias in the reverse direction resulting from the likelihood that those who give up, because they have less inherent desire to smoke than those who continue, are more likely to have been smokers who have smoked in a way that predicts less risk regardless of whether they give up. They may have smoked less, inhaled less, smoked to a longer butt, smoked lower tar brands, etc., facts which are difficult, if not impossible, to adjust for completely.
- (ii) The available data on risk in continuing smokers by age and number of cigarettes smoked do not actually permit reliable estimation of the relative effect of smoking on the first and penultimate stages to be made. Contrast, for example, Gaffney and Altshuler's best six-stage fit, based on the British Doctors data, which estimated the first stage effect to be almost three times stronger than that on the penultimate stage, with the work of Brown and Chu (1987) based on the Lubin study which estimated that the penultimate stage effect was about twice that on the first. While these estimates make different predictions about the pattern of risk on giving up smoking, neither should be relied upon. As regards the British

Doctors data, the absence of information on age at starting to smoke should particularly be noted, as taking it into account may have affected the predictions considerably.

(iii) Neither Freedman and Navidi, nor Gaffney and Altshuler, consider all the relevant data on ex-smoking (albeit some have appeared since their papers were published). Gaffney and Altshuler's analysis was based solely on the 20 year follow-up of the British Doctors data, which did not involve a large number of lung cancer deaths in ex-smokers. The "freezing" of the rate on stopping is clearly at best only an approximation. Doll (1978) in fact notes the data suggest a slight fall followed by an increase. Freedman and Navidi's analysis was based on two data sets for ex-smokers; the <u>US Veterans</u> data which appeared to show a slight decline in absolute risk on giving up smoking and the ACS CPS I data which appeared to show a more marked decline. Neither study, however, is based on a large number of lung cancer deaths in ex-smokers (169 in the Veterans, and 294 in the ACS study), and the numbers are particularly low regards longer term as ex-smokers. Thus the Veterans Study only has 21 deaths for ex-smokers who have given up for 20 years or more, while the ACS CPS I study only has 6 deaths for ex-smokers who have given up for 10 years or more (and this group remarkably shows a lower absolute risk than in nonsmokers - a fact that would not be explained by any model). More recent data, based on much larger numbers of lung cancer deaths in ex-smokers, show a very different pattern. Particularly noteworthy are the case-control study of Lubin et al (1984) which involved almost 2000 lung cancer deaths in ex-smokers and the ACS CPS II prospective study (Halpern et al, 1993) which involved over 1000. The pattern of response in ex-smokers in the Lubin study was found by Brown and Chu (1987) to be well described by a multistage model, though the fact that the case-control study was age matched makes it impossible to determine trends in absolute risk from the time of giving up. The most interesting data set in this respect is that from the ACS CPS II study. As shown in Table 4, it is quite clear when one looks at trends in risk over a long period in time that risk does not decline or freeze, it clearly increases with age. Whether one the increase in risk with considers absolute or excess risk, increasing age in ex-smokers is clearly evident. It seems though this has not formally been tested, likely, that the pattern of risk in Table 4 could be fitted quite well by a Certainly it would not fit the suggested multistage model. alternative "two-stage model with clonal growth" of Gaffney and Altshuler (1988) which predicts constant excess risk in ex-smokers on giving up. The rise in risk between ages 69-73 and 74-80 in smokers giving up at age 60-64 from 409 to 607 per 100,000 per year is clearly vastly greater than the corresponding rise for lifelong nonsmokers from 31 to 39 per 100,000 per year (each of these rates being highly stable since they are based on about 100 lung cancer deaths).

Although a more certain evaluation could perhaps be reached by a further simultaneous detailed investigation of all the data, one must conclude that the multistage model remains a very useful one. There appears no obvious reason at this point in time why predictions based on it should not be quite reliable.

#### 8. <u>REFERENCES</u>

Armitage P. A note on the time-homogeneous birth process. J R Statistic Soc Series B 1953;15:90-1.

Armitage P. Discussion on Doll's paper. J Royal Statistic Soc A 1971;134:155-6.

Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 1954;8:1-12.

Berenblum I. Sequential aspects of chemical carcinogenesis. In: Becker FF (ed). Cancer. A comprehensive treatise, Vol 1. 2nd ed. Plenum Press, New York, 1982. 451-84.

Brown CC, Chu KC. Implications of the multistage theory of carcinogenesis applied to occupational arsenic exposure. JNCI 1983a;70:455-63.

Brown CC, Chu KC. A new method for the analysis of cohort studies: implication of the multistage theory of carcinogenesis applied to occupational arsenic exposure. Environ Health Perspec 1983b;50:293-308.

Brown CC, Chu KC. Use of multistage models to infer stage affected by carcinogenic exposure: example of lung cancer and cigarette smoking. J Chronic Dis 1987;40(Suppl 2):171-9. Cook PJ, Doll R, Fellingham SA. A mathematical model for the age distribution of cancer in men. Int J Cancer 1969;4:93-112.

Crump K, Howe RB. The multistage model with a time-dependent dose pattern: application to carcinogenic risk assessment. Risk Analysis 1984;4:163-76.

Davies RF, Lee PN, Rothwell K. A study of the dose response of mouse skin to cigarette smoke condensate. Br J Cancer 1974;30:146-56.

Day NE, Brown CC. Multistage models and primary prevention of cancer. JNCI 1980;64:977-89.

Day NE. Epidemiological data and multistage carcinogenesis. In: Models, mechanisms and etiology of tumour promotion. IARC Scientific Publications No 56, Lyon, 1984:339-57.

Doll R. The age distribution of cancer; implications for models of carcinogenesis. J Royal Statistic Soc A 1971;134:133-55.

Doll R. An epidemiological perspective of the biology of cancer. Cancer Res 1978;38:3573-83.

Doll R, Gray R, Hafner B, Peto R. Mortality in relation to smoking: 22 years' observations on female British doctors. BMJ 1980;1:967-71.

Doll R, Peto R. Mortality in relation to smoking: 20 years' observation on male British doctors. BMJ 1976;2:1525-36.

Doll R, Peto R. Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers. J Epidemiol Community Health 1978;32:303-13.

-84-

Doll R, Peto R. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. JNCI 1981;66:1191-308.

Druckrey H. Quantitative aspects in chemical carcinogenesis. In: Truhaut R, editor. Potential carcinogenic hazards from drugs. UICC Monograph Series Vol 7, 1967:60-78.

Fisher JC, Holloman JH. A hypothesis for the origin of cancer foci. Cancer 1951;4:916-8.

Fisher RA, Tippett LHC. Limiting forms of the frequency distribution of the largest or smallest number of a sample. Proc Cambridge Phil Soc 1928;24:180-90.

Fréchet M. Sur la loi probabilité de l'Écart maximum. Ann de la Soc polonaise de Math (Cracow) 1927;6:93-116.

Freedman DA, Navidi W. On the risk of lung cancer for ex-smokers. Technical Report No 135, University of California, 1987.

Freedman DA, Navidi W. Ex-smokers and the multistage model for lung cancer. Epidemiology 1990;1:21-9.

Gaffney M, Altshuler B. Examination of the role of cigarette smoke in lung carcinogenesis using multistage models. JNCI 1988;80:925-31.

Gumbel EJ. Statistics of extremes. Columbia University Press, New York, 1958.

Hakama M. Epidemiologic evidence for multi-stage theory of carcinogenesis. Int J Cancer 1971;7:557-64.

Hakulinen T, Pukkala E. Future incidence of lung cancer: forecasts based on hypothetical changes in the smoking habits of males. Int J Epidemiol 1981;10:233-40.

Halpern MT, Gillespie BW, Warner KW. Patterns of absolute risk of lung cancer mortality in former smokers. JNCI 1993;85:457-64.

Hammond EC. Smoking in relation to the death rates of one million men and women. In: Haenszel W, editor. Epidemiological approaches to the study of cancer and other chronic diseases. Natl Cancer Inst Monogr 1966; 19:127-204.

Hammond EC, Selikoff IJ, Seidman H. Asbestos exposure, cigarette smoking and death rates. Ann NY Acad Sci 1979;330:473-90.

Hecker E. Cocarcinogens of the tumour-promoter type as potential risk factors of cancer in men. A first complete experimental analysis of an etiological model situation and some of its consequences. In: Models, mechanisms and etiology of tumour promotion. IARC Scientific Publications No 56, Lyon, 1984:441-63.

Hegmann KT, Fraser AM, Keaney RP, Moser SE, Nilasena DS, Sedlars M et al. The effect of age at smoking initiation on lung cancer risk. Epidemiol 1993;4:444-8.

International Agency for Research on Cancer. Tobacco smoking. IARC Monographs on the evaluation of the carcinogenic risk of chemicals to humans. Volume 38, IARC, Lyon 1986.

Kahn H. The Dorn study of smoking and mortality among US veterans: report on 8½ years of observation. In: Haenszel W, editor. Epidemiological approaches to the study of cancer and other chronic diseases. Natl Cancer Inst Monogr 1966;19:1-126.

Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci. USA, 1971;68:820-3.

-86-

Lee PN. Final analysis - Experiment 1.1.1.9. The effect of stopping painting. Tobacco Research Council document M432, 1974.

Lee PN. Cigarette smoking and lung cancer. A new mathematical model. Tobacco Advisory Council document TA 1243, 1979.

Lee PN. A summary of evidence relating risk of lung cancer to type of cigarette smoked. Unpublished, 1993a.

Lee PN. Epidemiological studies relating family history of lung cancer to risk of the disease. Indoor Environment 1993b;2:129-42.

Lee PN, O'Neill JA. The effect both of time and dose applied on tumour incidence rates in benzopyrene skin painting experiments. Br J Cancer 1971;25:759-70.

Lee PN, Rothwell K, Whitehead JK. Fractionation of mouse skin carcinogens in cigarette smoke condensate. Br J Cancer 1977;35:730-742.

Lijinsky W. Life-span and cancer: the induction time of tumours in diverse animal species treated with nitrosodiethylamine. Carcinogenesis 1993;14:2373-5.

Likhachev A, Anisimov V, Montesano R, editors. Age-related factors in carcinogenesis (IARC Scientific Publication No 58) Lyon, International Agency for Research on Cancer, 1985.

Lubin JH, Blot WJ, Berrino F, Flamant R, Gillis CR, Kunze M, Schmahl D, Visco G. Modifying risk of developing lung cancer by changing habits of cigarette smoking. BMJ 1984;288:1953-56.

Major IR, Mole RH. Myeloid leukaemia in X-ray irradiated mice. Nature 1978;272:455-6.

-87-

Mazumdar S, Redmond CK, Costantino JP, Patwardhan RN, Zhou SYJ. Recent developments in the multistage modelling of cohort data for carcinogenic risk assessment. Environ Health Perspec 1991;90:271-7.

Moolgavkar SH. The multistage theory of carcinogenesis (Letter). Int J Cancer 1977;19:730.

Nordling CO. A new theory on the cancer-inducing mechanism. Br J Cancer 1953;1:68-72.

Parish SE. Exploiting animal tumour data using multistage models. D Phil Thesis, University of Oxford, 1981.

Passey RD. Some problems of lung cancer. Lancet 1962;ii:107-12.

Peto J. Early and late-stage carcinogenesis in mouse skin and man. In: Models, mechanisms and etiology of tumour promotion. IARC Scientific Publications No 56, Lyon, 1984:359-71.

Peto J, Seidman H, Selikoff IJ. Mesothelioma mortality in asbestos workers: Implications for models of carcinogenesis and risk assessment. Br J Cancer 1982;45:124-135.

Peto R. Epidemiology, multistage models and short-term mutagenicity tests. In: Hyatt H, Watson J, jWnston S, editors. Origins of human cancer. Cold Spring Harbor NY: Cold Spring Harbor Laboratory, 1977:1403-28.

Peto R, Parish SE, Gray RG. There is no such thing as ageing, and cancer is not related to it. IARC Scientific Publications No 58, Lyon, 1985:43-53.

Peto R, Roe FJC, Lee PN, Levy L, Clack J. Cancer and aging in mice and men. Br J Cancer 1975;32:411-26.

Peto R, Doll R. Comment on the letter by Moolgavkar. Int J Cancer 1977;19:731.

Pike MC, Doll R. Age at onset of lung cancer: its significance in relation to the effects of smoking. Lancet 1965;1:665-8.

Pike MC. A method of analysis of a certain class of experiments in carcinogenesis. Biometrics 1966;22:142-161.

Renan MJ. How many mutations are required for tumorigenesis? Implications from human cancer data. Molecular Carcinogenesis 1993;7:139-49.

Rogot E. Smoking and mortality among US veterans. J Chronic Dis 1974;27:189-203.

Seidman H. Age at exposure versus years of exposure. Natl Cancer Inst Monogr 1985;67:205-9.

Selikoff IJ, Hammond EC. Multiple risk factors in environmental cancer. In "Persons at high risk of cancer: An approach to cancer etiology and control" (ed J.F.Fraumeni, Jr.), Academic Press, New York 1975, 467-83.

Sellers TA, Bailey-Wilson JE, Elston RC, Wilson AF, Elston GZ, Ooi WL, Rothschild H. Evidence for Mendelian inheritance in the pathogenesis of lung cancer. J Natl Cancer Inst 1990;82:1272-9.

Sobue T, Yamaguchi N, Suzuki T, Fujimoto I, Matsuda M, Doli O <u>et al</u>. Lung cancer incidence rate for male ex-smokers according to age at cessation of smoking. Jpn J Cancer Res 1993;84:601-7.

Stenbäck F, Peto R, Shubik P. Initiation and promotion at different ages and doses in 2200 mice. II Decrease in promotion by TPA with ageing. Br J Cancer 1981;44:15-23. Swartz JB. Use of a multistage model to predict time trends in smoking induced lung cancer. J Epidemiol Community Health 1992;46:311-5.

Szende B, Kendrey G, Lapis K, Lee PN, Roe FJC. Accuracy of admission and pre-autopsy clinical diagnosis in the light of autopsy findings: a study conducted in Budapest. Human and Experimental Toxicology 1994, in press.

Townsend J. Smoking and lung cancer. A cohort study of men and women in England and Wales 1935-1970. J Royal Statistic Soc A 1978;141:95-107.

US Surgeon General. Smoking and health. A Report of the Surgeon General. US Department of Health, Education and Welfare. Publication No (PHS) 79-50066, 1979.

US Surgeon General. The health consequences of smoking. Cancer. A report of the Surgeon General. US Department of Health and Human Services. Publication No (PHS) 82-50179, 1982.

US Surgeon General. Reducing the health consequences of smoking. 25 years of progress. A report of the Surgeon General. US Department of Health and Human Services. Publication No (CDC) 89-8411, 1989.

Whittemore AS. Effect of cigarette smoking in epidemiological studies of lung cancer. Stat Med 1988;7:223-8.

ς.

# Observed male lung cancer death rates per 100,000 per year (numbers of deaths) in relation to age, age of starting and number of cigarettes smoked (from Kahn, 1966)

	Age of starting to smoke							
Age	<15		15-19		20-24		25+	
<u>All cigarette smokers</u>								
55-64	251	(70)	168	(293)	99	(133)	53	(30)
65-74	478	(65)	350	(259)	241	(138)	162	(70)
<u>1-9 cigs/day</u>								
55-64	NE <sup>*</sup>	(1)	27	(5)	42	(6)	15	(2)
65-74	NE	(2)	108	(7)	99	(8)	52	(5)
<u>10-20 cigs/day</u>								
55-64	156	(16)	118	( 81)	78	(47)	43	(13)
65-74	321	(17)	322	(100)	186	( 54)	152	(29)
<u>31-39 cigs/day</u>								
55-64	323	(32)	217	(133)	135	(55)	58	(10)
65-74	744	(30)	435	( 89)	363	( 49)	282	(25)
<u>&gt;39 cigs/day</u>								
55-64	366	(15)	341	( 49)	177	(14)	182	(3)
65-74	NE	(12)	578	( 32)	NE	( 16)	296	(6)

\*NE: rate not estimated

TABLE 1

Age of start	Age	Duration	Duration <sup>4</sup> (divided by 10 <sup>6</sup> )	Population (scaled)	Deaths observed	Deaths expected
25+	55-64	33	1.19	0.566	30	31.2
20-24	55-64	38	2.08	1.343	133	129.6
15-19	55-64	43	3.42	1.744	293	276.6
25+	65-74	43	3.42	0.432	70	68.5
<15	55-64	48	5.31	0.279	70	68.7
20-24	65-74	48	5.31	0.573	138	141.1
15-19	65-74	53	7.89	0.740	259	270.8
<15	65-74	58	11.32	0.136	65	71.4
Total	- <u></u>				1058	1058.0

Fit of a fourth power law relationship of duration of smoking to risk of lung cancer (using data of Table 1 for all cigarette smokers)

TABLE 2

NB. Scaled population estimated by deaths/rate per 100,000 per year Expected deaths calculated by multiplying population x duration<sup>4</sup> x scaling factor

Scaling factor =  $\Sigma$  observed deaths /  $\Sigma$ (population x duration<sup>4</sup>).

-92-

# TABLE 3 Dose relationships under various hypotheses

## Hypothesis A - equal effects on stages 1 and 6

Dose	(proportional to numbers	<u>Stage_effects</u>		Relative Risk	
. <u> </u>	of cigarettes smoked)	<u>a</u>	<u></u>	<u>at age 70-74</u>	<u>Linear fit</u>
	0	1	1	1176	1176
	1	2	2	2665	3446
	2	3	. 3	4466	5715
	4	5	5	9005	10255
	6	7	7	14794	14794
	8	9	9	21833	19333
	10	11	11	30122	23873

### <u>Hypothesis B - greater effect on stage 6 than stage 1</u>

	<u>Stage e</u>	ffects	<u>Relative risk</u>	
Dose	<u>α</u>	<u></u>	<u>at age 70-74</u>	<u>Linear fit</u>
0	1	1	1176	1176
1	1.25	3.875	4708	5270
2	1.5	6.75	8465	9363
4	2	12.5	16652	17550
6	2.5	18.25	25737	25737
8	3	24	35721	33924
10	3.5	29.75	46603	42111

i

Lung cancer incidence rates per 100,000 per year (numbers of deaths) in relation to age, and time of giving up smoking (from Halpern <u>et al</u>, 1993)

Smoking	Age								
habits	40-43	44-48	49-53	54-58	59-63	64-68	69-73	74-80	
Never smoker	0.000*	3.62	4.69	6.93	13.28	18.99	31.23	39.48	
	(0)	(9)	(20)	(33)	(61)	(75)	(91)	(93)	
Current smoker	10.72	45.75 (62)	82.24 (195)	156.8 (398)	272.0	430.9 (622)	643.0 (518)	858.7 (332)	
Former smoker	(0)	(02)	(2007	(0)0)	(002)	()	(/	()	
Age at cessation	ı								
30-39	-	7.73	18.46	27.70	19.29	57.39	68.49	42.76	
		(4)	(18)	(27)	(13)	(22)	(14)	(4)	
40-49	-	-	-	52.21	73.59	106.8	109.2	114.4	
				(53)	(74)	(72)	(30)	(20)	
50-54	_	-	-	-	134.8	133.8	170.9	241.5	
					(66)	(54)	(45)	(33)	
55-59	-	-	-	-	-	244.0	270.5	353.6	
						(89)	(64)	(48)	
60-64	-	-	-	-	-	-	409.2	607.4	
							(100)	(97)	
65-69	-	-	-	-	-	-	-	724.8	
								(91)	

\*Based on 82,335 person years

