*Journal of Epidemiology and Community Health*, 1978, **32**, 303-313

# Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers

RICHARD DOLL AND RICHARD PETO
*From the Radcliffe Infirmary, University of Oxford*

SUMMARY In a 20-year prospective study on British doctors, smoking habits were ascertained by questionnaire and lung cancer incidence was monitored. Among cigarette smokers who started smoking at ages 16–25 and who smoked 40 or less per day, the annual lung cancer incidence in the age range 40–79 was :

$$0{\cdot}273 \times 10^{-12}.\ (\text{cigarettes/day}+6)^{2}.\ (\text{age}-22{\cdot}5)^{4{\cdot}5}.$$

The form of the dependence on dose in this relationship is subject not only to random error but also to serious systematic biases, which are discussed. However, there was certainly some statistically significant ($P<0{\cdot}01$) 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 by smoking. If a higher than linear dose-response relationship exists between dose per bronchial cell and age-specific risk per bronchial cell, this may help explain why bronchial carcinomas chiefly arise in the upper bronchi, for dilution effects might then protect the larger areas lower in the bronchial tree.

In essence, the 'multistage model' hypothesis about carcinoma induction is that a few changes, each heritable when somatic cells divide in the tissues, are needed to alter an ordinary epithelial cell into the progenitor of a carcinoma. As well as these 'stages', other processes may of course also be relevant—for example, partially transformed epithelial cells may already have some selective advantage over their unaltered neighbours, thus eventually increasing the number of such cells at risk of further change, while the host may have some defences against partially or even fully altered cells, thus reducing the number of such cells. There is not as yet sufficient knowledge of these other processes to allow their proper incorporation into the mathematical formulation of multistage models, but, as the various qualitatively different stages and processes involved in the production of one single carcinoma come to be understood separately, the need will arise for some synthesis of them into an overall sequence of events. Multistage models seem at present to offer the most promising framework for such an eventual synthesis (Peto, 1977), even if current knowledge is too sparse for such models to be tested critically.

Successive stages in the transformation of one cell may be separated from each other by several years, and the biologic nature of the early stages may be completely different from that of the later stages. If so, different stages may have different causes. Epidemiological data on smoking and bronchial carcinoma are more extensive than for any other cause of human carcinomas, so it may be profitable to ask which stage or process smoking affects most strongly. This has already been done (Doll, 1971), but the epidemiological evidence was difficult to fit together plausibly (Armitage, 1971). Smoking in early adult life seemed to have a substantial effect on the risk of cancer in old age, suggesting that smoking affected at least one early stage. Giving up smoking in later adult life seemed to have a substantial effect on the risk five or 10 years later, suggesting that smoking also affected at least one late stage or process. The simplest assumption would be that if, comparing two people with different smoking habits, the dose-rate of smoke to the target cells in one person was double that in the other person, then the rate of occurrence of both the early stage and the late stage or process would be approximately doubled, thereby multiplying the final age-specific incidence rate of lung cancer by about four. (This is not a firm prediction, of course,

for the two separate occurrence rates may not be even approximately proportional to the dose-rate of smoke). Armitage (1971), however, pointed out that most epidemiological data suggested that the age-specific risk was approximately proportional to the reported daily consumption rather than to the square of the reported consumption, and he found it surprising that the observed relationship seemed linear rather than quadratic, if two different aspects of carcinoma induction were indeed affected.

The reported cigarette consumption is, however, certainly an inaccurate measure of the current (let alone the past) extent of exposure of the bronchial epithelial stem cells to the carcinogenic agents in cigarette smoke, because of differences between smokers in the condition of the bronchial tree (deciliation, airflow obstruction, phlegm production, etc.); type of cigarette; number and size of puffs; butt length and depth of inhalation*.

The aims of this paper are (1) to present in detail the age-specific lung cancer incidence data from the 20-year follow-up in the prospective study of male British doctors undertaken in 1951 (Doll and Hill, 1964) in as accurate a form as possible, and (2) to point out that even if the remaining inaccuracies are fairly random, they will tend to conspire to make the exponent of dose in the observed relationship between lung cancer incidence and daily cigarette consumption lower than that in the true relationship between lung cancer incidence and extent of exposure of the bronchial stem cells to the relevant components of smoke. Thus, for example, even if the true biological dose-response relationship is quadratic, the observed epidemiological dose-response relationship might be roughly linear.

Taken together, (1) and (2) may allow circumvention of the difficulty discussed by Armitage.

## DATA ON SMOKING HABITS: 'REGULAR' CIGARETTE SMOKERS

The main features of the study we shall use have been described in detail elsewhere (Doll and Peto, 1976). The study period runs for 20 years, from 1 November 1951 to 31 October 1971, and we shall subdivide this period into 20 'study years' (study year 1, study year 2, . . . study year 20), recording only the study year in which deaths and lung cancer onsets happen, rather than the exact dates of these events. On 1 November 1951, postal inquiries (questionnaire Q1) were made into the current and past smoking habits of all men on the 1951 British medical register who were thought to be resident in Britain in October of that year. 34 440 men (69% of those then alive) replied, almost all very promptly. This paper chiefly concerns those who reported in 1951 that they were lifelong non-smokers, or who reported in 1951 that they had smoked cigarettes regularly since early adult life (which we defined as ages 16–25) and who did not report that they had ever given up smoking or smoked anything other than cigarettes. During the seventh and fifteenth years of the study, further questionnaires (Q2 and Q3) were sent to the doctors, reminding them of their previously stated smoking habit and asking whether this had continued. If they did not reply they were reminded at least twice. Replies were received before the ends of study years 7 and 15 from 98·4% and 96·4% of the survivors.

Ideally, we would like to examine the incidence of lung cancer among lifelong non-smokers and among men who have been exposed to a constant daily dose of cigarette smoke since a common starting age. If, therefore, in response to questionnaire Q2 or Q3, any non-smoker reported current or previous smoking, or any smoker reported giving up smoking temporarily or permanently, or reported current or previous use of cigars or pipes, or reported a change in consumption of more than five cigarettes/day, then that person has been excluded from our present analysis as from the end of year 7 (for Q2) or year 15 (for Q3). The few men who did not reply to either or both of Q2 and Q3 were presumed not to have altered their habits. We have called those smokers who satisfied our criteria for inclusion in part or all of the study 'regular' smokers, and we have studied them and the non-smokers up to the time when they died, or were excluded from the study, or were diagnosed as having lung cancer.

In the present analysis, we have related lung cancer onsets between the start of study year 1 and the end of study year 7 to the smoking habits described in Q1, irrespective of information about the dates of changes in habits gleaned from Q2. Likewise, we have related onsets between the start of study year 8 and the end of study year 15 to Q2, irrespective of Q3, while we have related onsets between the start of year 16 and the end of year 20 to Q3, irrespective of any change which in fact occurred during years 16–20.†

*The effects of these differences in smoking style are important, but difficult to quantify, as is demonstrated by the curious observation that among heavy smokers the inhalers seem able to get some smoke safely past the main danger area in the upper bronchi, and so actually have a somewhat *lower* risk of bronchial carcinoma than do non-inhalers with the same cigarette consumption (Doll and Peto, 1976).

†This may seem perverse, since we know that some of our 'smokers' had in fact given up a few years previously, but it is necessary in order to avoid bias. If, among men who gave up smoking soon after one questionnaire, those who then died before the next questionnaire were classified as smokers while those who survived were not, the death rates of smokers would be overestimated and those of ex-smokers would be underestimated. No information was sought after men had died about their smoking habits between the last questionnaire they answered and their death.

COMPLETENESS OF FOLLOW-UP

Of the 34 440 respondents in 1951 to Q1, 10 072 (29·2%) were known to have died before 31 October 1971, 24 265 were traced after 31 October 1971 and were known to have been alive on 1 November 1971, and 103 (0·3%) could not be traced and have been arbitrarily taken to be alive on 1 November 1971. Most of these 0·3% were known to be alive at the time of the questionnaire Q3 in study year 15. Of those alive on 1 November 1971, we have examined almost all the death certificates of those who were known to have died within the next two and a half years for mention of lung cancer, in case any such lung cancers may have had onset during the main study period.

DATA ON LUNG CANCER ONSETS:

We shall be concerned with the date of clinical onset of lung cancer (which we shall use to calculate incidence rates during the 20-year study period). We have therefore excluded all men with onset before the initial questionnaire Q1, and we do not count any onsets which occurred after the end of the main 20-year study period at 31 October 1971. The principal means of ascertainment has been by death certificates, 556 of which mentioned cancer of the trachea, bronchus, or pleura as an underlying or associated cause of death. In addition, we learned of 15 other cases by informal means, chiefly from the replies to questionnaires Q1, Q2, or Q3, or to a questionnaire which was distributed after the end of study year 20 for other purposes. We wished to exclude any of these 571 putative lung cancers which were not really lung cancer, or which had onset outside the main study period. To help us decide which to accept, we obtained information about the basis for the diagnosis of 567 of these 571 possible lung cancers from the doctor who had signed the death certificate or from the consultant to whom the patient had been referred. Thirty-two of the 567 (5·6%) were, on review, considered unlikely to have been cancers of the lung or trachea and have been discounted; these are listed in Doll and Peto (1976). The remainder, together with the four for which we could not obtain information about the basis for the diagnosis, were accepted. Where there was doubt whether or not to accept particular cases, we sought the advice of Dr. J. R. Bignall (consultant physician at the Brompton Hospital for Diseases of the Chest), who was not informed of the subject's smoking history.

For those 539 in whom the diagnosis was accepted, we also sought the date when the diagnosis was first made and regarded this as the date of onset of the disease. The original plan of the study had not required information about the date of diagnosis, and later inquiries failed to elicit the date in 60 cases, most of which were doctors who died during the first few years of the study. For these, we assumed that the diagnosis was made three months before the date of death, which is the median duration of survival from clinical onset in the cases for which we had complete information. Four hundred and eighty-three of the 539 accepted cases of bronchial carcinoma had estimated dates of onset within the 20-year period from 1 November 1951 to 31 October 1971.

Not all patients were equally thoroughly investigated. They were, therefore, classified according to the strength of the evidence. 'Category 1' cases were those diagnosed at necropsy or on (1) microscopic evidence of cancer compatible with a bronchial origin plus (2) macroscopic evidence of the primary site of origin at operation, bronchoscopy, or radiological examination. 'Category 2' cases were those diagnosed without one or other of these and without necropsy, while 'Category 3' cases were diagnosed only on history and clinical examination. Four cases were known to us on the basis of the death certificate alone.

Table 1 *Reported cases of lung cancer by date of onset, diagnostic category, and relation to death*

| *Estimated date of clinical onset* | *Diagnostic category* | *Underlying cause of death* | *Contributory cause of death* | *Unrelated to death* | *Total reported* |
|---|---|---|---|---|---|
| During main study period 1 Nov. 1951 to 31 Oct. 1971 | 1 | 259 | 13 | 5 | 277 |
| | 2 | 172 | 7 | 8 | 187 |
| | 3 | 15 | 0 | 0 | 15 |
| | Death certificate only | 4 | 0 | 0 | 4 |
| | Total accepted | 450 | 20 | 13 | 483 |
| Before 1 Nov. 1951 | Total accepted | 6 | 0 | 2 | 8 |
| After 31 Oct. 1971 | Total accepted | 44 | 4 | 0 | 48 |
| All periods | Total accepted | 500 | 24 | 15 | 539 |
| All periods | Not accepted* | 29 | 3 | NA | 32 |
| All periods | Total reported | 529 | 27 | 15 | 571 |

*Bronchial carcinoma mentioned on death certificate, but not accepted on investigation; excluded from analysis.

The distribution of the reported cases by date of onset, diagnostic category, and relation to death is shown in Table 1. Of the 483 accepted lung cancers with estimated date of onset within the main study period, only 215 affected the non-smokers and regular cigarette smokers who are the subject of the present paper.

## Results

The basic data are presented in Tables 2 and 3. Because of digit preference in reporting cigarette consumption, the mean consumption (Table 2) is often not in the middle of the range, as many doctors stated that they smoked exactly 20, 30, or 40 cigarettes, while few reported adjacent numbers. For example, among men smoking 35 or more cigarettes/day, 34% reported smoking 35–39, 46% reported smoking exactly 40, and 20% reported smoking more than 40. Because of this, the top two dose-groups which we have used are '35–40' and 'more than 40', rather than '35–39' and '40 or more'. This has given us a reasonable amount of observational material throughout the dose-range 0–40,

Table 2 *Distribution of man-years by current age and by amount smoked**

| *Age group (years)* | *CIGARETTES/DAY (RANGE AND MEAN)* | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Never smoked* | *1–4 (2·7)* | *5–9 (6·6)* | *10–14 (11·3)* | *15–19 (16·0)* | *20–24 (20·4)* | *25–29 (25·4)* | *30–34 (30·2)* | *35–40 (38·0)* | *More than 40 (50·9)* | *All amounts (11·0)* |
| 20–24 | 378 | 19½ | 38 | 91½ | 91 | 57 | 7½ | 2 | 2½ | 0 | 687 |
| 25–29 | 5 099½ | 400 | 701½ | 1 529 | 1 427 | 1 424 | 304½ | 153 | 46 | 10½ | 11 095 |
| 30–34 | 10 838 | 914 | 1 762½ | 3 270 | 3 343 | 3 966½ | 1 042½ | 582½ | 224½ | 32½ | 25 976 |
| 35–39 | 15 105 | 1156½ | 2 178 | 3 819½ | 4 649½ | 6 003½ | 1 991½ | 1 108½ | 545½ | 110½ | 36 668 |
| 40–44 | 17 846½ | 1216 | 2 041½ | 3 795½ | 4 824 | 7 046 | 2 523 | 1 715½ | 892½ | 234 | 42 134½ |
| 45–49 | 15 832½ | 1000½ | 1 745 | 3 205 | 3 995 | 6 460½ | 2 565½ | 2 123 | 1150 | 305½ | 38 382½ |
| 50–54 | 12 226 | 853½ | 1 562½ | 2 727 | 3 278½ | 5 583 | 2 620 | 2 226½ | 1281 | 335½ | 32 693½ |
| 55–59 | 8 905½ | 625 | 1 355 | 2 288 | 2 466½ | 4 357½ | 2 108½ | 1 923 | 1063 | 284 | 25 376 |
| 60–64 | 6 248 | 509½ | 1 068 | 1 714 | 1 829½ | 2 863½ | 1 508½ | 1 362 | 826 | 183½ | 18 112½ |
| 65–69 | 4 351 | 392½ | 843½ | 1 214 | 1 237 | 1 930 | 974½ | 763½ | 515 | 120 | 12 341 |
| 70–74 | 2 723½ | 242 | 696½ | 862 | 683½ | 1 055 | 527 | 317½ | 233 | 52 | 7 392 |
| 75–79 | 1 772 | 208½ | 517½ | 547 | 370½ | 512 | 209½ | 130 | 88½ | 18½ | 4374 |
| 80–84 | 1 185½ | 173 | 281 | 314 | 180½ | 188 | 81 | 37 | 36 | 2½ | 2 478½ |
| 85+ | 870½ | 77½ | 149 | 123½ | 61½ | 67½ | 28 | 1 | ½ | 0 | 1 379 |
| All ages | 103 381½ | 7788 | 14 939½ | 25 500 | 28 437 | 41 514 | 16 491½ | 12 445 | 6904 | 1689 | 259 089½ |

*On average, a group of men of a particular stated age at the start of the study will have a mean age exactly halfway through that year of age. Likewise, on average, men who suffer lung cancer onset or death from any cause in a particular study year will do so exactly halfway through that study year. These 'average' assumptions underlie this Table.

Table 3 *Numbers of lung cancer onsets during 20-year study period by current age and by amount smoked*

| *Age group (years)* | *CIGARETTES/DAY* | | | | | | | | | | *Per cent in diagnostic category* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Never smoked* | *1–4* | *5–9* | *10–14* | *15–19* | *20–24* | *25–29* | *30–34* | *35–40* | *More than 40* | *All amounts* | |
| 20–24 | — | — | — | — | — | — | — | — | — | — | — | |
| 25–29 | — | — | — | — | — | — | — | — | — | — | — | |
| 30–34 | — | — | — | — | — | — | — | — | — | — | — | |
| 35–39 | 1/1 o | — | — | — | — | — | — | — | — | — | 1/1 o | |
| 40–44 | — | — | — | 1/1 u | — | 0/1 | — | 1/1 o | — | — | 2/3 o, u | 85% |
| 45–49 | — | — | — | 0/1 | 1/1 a | 0/1 | 2/2 o, u | 2/2 2s | — | — | 5/7 o, a, 2s, u | |
| 50–54 | 1/1 s | — | — | 2/2 a, s | 3/4 s, u | 5/6 2o, 3s | 3/3 3s | 3/3 a, 2s | 3/3 o, s, u | — | 20/22 3o, 2a, 12s, 2u | |
| 55–59 | 2/2 o, u | 1/1 s | — | 1/1 o | — | 6/8 3a, 3s | 4/5 3s, u | 5/6 a, 2s, 2u | 3/4 o, s, u | 1/1 a | 23/28 3o, 5s, 10s, 5u | 82% |
| 60–64 | — | 1/1 a | 1/1 s | 1/1 s | 2/2 a | 7/13 a, 6s | 3/4 3s | 5/11 3s, 2u | 6/7 o, a, 2s, u | 1/1 o | 27/41 2o, 4a, 16s, 3u | 66% |
| 65–69 | — | — | 1/1 s | 1/2 o | 0/2 | 9/12 2a, 5s, u | 3/5 o, a, s | 5/9 3o, 2s | 3/9 3s | 0/1 | 22/41 5o, 3a, 12s, u | 54% |
| 70–74 | 0/1 | 1/1 s | 1/2 s | 2/4 s, u | 2/4 s, u | 3/10 3s | 4/7 a, 2s, u | 0/2 | 2/5 s, u | 1/2 o | 16/38 o, a, 10s, 4u | 42% |
| 75–79 | 0/2 | — | — | 1/4 u | 3/5 a, s, u | 3/7 s, u | 2/4 s, u | 2/2 2s | 0/2 | — | 11/26 a, 5s, 4u | |
| 80–84 | — | — | — | 0/1 | 1/1 | 1/1 a | — | 1/2 u | 0/1 | — | 3/6 a, u | 41% |
| 85+ | — | — | — | — | 0/1 | 0/1 | — | — | — | — | 0/2 | |
| All ages | 4/7 2o s, u | 3/3 a 2s | 3/4 3s | 9/17 2o, a 3s, 3u | 12/20 3a 3s, 3u | 34/60 2o, 7a 21s, 2u | 21/30 2o, 2a 13s, 4u | 24/38 4o, 2a 13s, 5u | 17/31 3o, a 8s, 4u | 3/5 2o, a | 130/215 17o, 18a 67s, 22u | 60% |
| Per cent in diagnostic category 1 | 57% | | 62% | | 57% | | | 62% | | | 60% | |

Key: x/y, where x=number in diagnostic category 1 and y=total number of confirmed onsets.
The cell type was determined histologically for 124 of the 130 cancers of diagnostic category 1, and for these the numbers of oat cell (o), squamous cell (s), undifferentiated (u), and adenocarcinomas (a), are indicated below x/y.

while the topmost group, 'more than 40', with which we shall deal separately by discussion rather than by model-fitting, contains under 1% of the whole study population. The percentages of cases that were diagnostic Category 1 are also cited in Table 3. It appears that over 80% of lung cancer patients aged under 60 underwent a thorough diagnostic investigation, while only about 40% of lung cancer patients aged over 70 did so. There does not appear to be any marked tendency for the thoroughness of diagnostic investigation to be related to smoking habits, but although this is reassuring it does not guarantee that diagnostic errors will be unrelated to dose. However, we consider it unlikely that any material diagnostic biases exist which are dose related.

### HISTOLOGY

The numbers in the four separate histological categories (oat, adeno, squamous, and undifferentiated) are too small to allow us to estimate the shapes of the four separate dose-response relationships; moreover, the histologic criteria used must have varied from hospital to hospital. However, in spite of, (or, perhaps, partly because of!) these difficulties, we did observe statistically significant (1-tailed $P<0{\cdot}001$) trends with respect to the ten dose-groups in Tables 2 and 3 in the age-standardised incidence of each separate histologic category (oat, adeno, squamous and undifferentiated), contrary to the report of Doll and Hill (1964), based on the 10-year follow-up of this study, that there was no material dependence on smoking for adenocarcinomas.

### RESTRICTED ANALYSIS

In what follows we shall restrict our attention wholly to Tables 2 and 3 (lifelong non-smokers, or men who started smoking between the ages of 16 and 25 and who never reported stopping, changing by more than 5/day, or smoking any form of tobacco other than cigarettes). We shall concern ourselves only with those lung cancers where the diagnosis was accepted and where the estimated date of onset lay within the main study period, and we shall analyse these irrespective of histological type or diagnostic category. Furthermore, we shall exclude from our analysis those men who reported smoking over 40 cigarettes/day, studying only the dose range 0–40 cigarettes/day, over the whole of which we have appreciable amounts of data. (Because the numbers of lung cancers arising among men smoking 1–4 and 5–9/day were both small we have, when plotting our data, merged these into one single group, with mean consumption 5·3/day.) Finally, because there were no lung cancers among smokers aged under 40, and because there was very little data on smokers aged over 80, we shall restrict our analysis to lung cancer onsets in the age range 40–79. This avoids those extremes of old age in which diagnostic errors are likely to be most severe. The range of doses and ages that will henceforth concern us are delimited by the lines in Tables 2 and 3.

### DOSE-RESPONSE RELATIONSHIP

The proportion of heavy smokers in old age differs from that in middle age, and this must be allowed for when studying the dependence of incidence rates on dose. For each 5-year age group we have therefore computed the numbers of lung cancers that would have been expected in each dosage category if the total number of onsets in that age group had been distributed (in proportion to the man-years in that age group) irrespective of dose. Summation of these expected numbers for all eight age groups within one dosage category then yields an 'indirectly age-standardised' overall expected number for that dosage category. These overall numbers are given in Table 4, together with the

Table 4 *Dose-response standardised for age among men aged 40–79 smoking 0–40/day*

| | *Cigarettes/day* Range | Mean | *O* Observed number of onsets | *E* Indirectly age-standardised expected | *Relative risk estimated by* (a) O/E | (b) ML Maximum likelihood | *Age-standardised onset rate/$10^5$ man-years** |
|---|---|---|---|---|---|---|---|
| | 0 | 0·0 | 6 | 74·20 | 0·081 | 0·080 | 9 |
| | (1–4) | (2·7) | (3) | (6·38) | (0·470) | (0·458) | (50) |
| | 1–9 | 5·3 | 7 | 21·02 | 0·333 | 0·321 | 36 |
| | (5–9) | (6·6) | (4) | (14·64) | (0·273) | (0·263) | (29) |
| | 10–14 | 11·3 | 16 | 20·47 | 0·782 | 0·769 | 85 |
| | 15–19 | 16·0 | 18 | 19·66 | 0·916 | 0·972 | 102 |
| | 20–24 | 20·4 | 58 | 31·11 | 1·864 | 1·891 | 209 |
| | 25–29 | 25·4 | 30 | 15·09 | 1·988 | 2·032 | 224 |
| | 30–34 | 30·2 | 36 | 11·97 | 3·008 | 3·125 | 344 |
| | 35–40 | 38·0 | 30 | 7·49 | 4·005 | 4·134 | 456 |
| Total | 0–40 | 10·7 | 201 | 201·00 | 1·000 | 1·000 | 112 |

*Maximum likelihood (ML) relative risk × 112, where 112 = crude onset rate/$10^5$ man-years in all men aged 40–79 smoking 0–40/day = 201 × $10^5$/179273.

ratio of observed to expected. This ratio gives an indication of the relative risks associated with different habits. Estimation of these relative risks by an iterative maximum likelihood procedure* is in principle slightly preferable to estimation of them by O/E, but as may be seen by comparing columns (a) and (b) in Table 4, no material differences exist between the results of these two methods of estimation. These statistical methods would be ideal if the age-specific incidence rates at different dose levels were simple multiples of each other. Although this might be approximately true among the different groups of smokers, there is good evidence (Doll, 1971) that the incidence among non-smokers is not a multiple of that among smokers. However, since we have only six lung cancers among non-smokers aged 40–79, these statistical methods will be sufficiently accurate for these data, although the formulae we shall derive from these relative risk estimates will only be wholly satisfactory for predicting lung cancer risks which are dominated by the effects of smoking.

We have found it easier to grasp the medical significance of these relative risks if we multiply each by the overall lung cancer incidence rate in the whole population (for men aged 40–79 smoking 0–40/day this is $112/10^5$), to obtain age-standardised incidence rates/$10^5$. These are therefore displayed in the last column of Table 4 and are plotted in Fig. 1.

In Fig. 1 we also plot the best-fitting straight line (incidence$=9$ (dose$+1)/10^5$) and the best-fitting second order polynomial (incidence$=0{\cdot}26$ (dose$+6)^2/10^5$). Although the indicated 90% confidence intervals for the upper points are large, the curved line does appear to fit considerably better than the straight line, and this visual impression may be confirmed by a statistical test ($\chi^2\dagger=7{\cdot}5$ on one degree of freedom, $P<0{\cdot}01$ for curvature).

## RESULTS AMONG MEN WHO REPORT SMOKING MORE THAN 40/DAY

It should be noted that if we had not excluded the 100 or so men who reported smoking more than 40/day, the results of this whole analysis would have been materially different. Although their average reported consumption was 50 cigarettes/day, their observed lung cancer risk is similar to that among men reporting smoking only 30 cigarettes/day. (In Fig. 1, the point for men smoking more than 40/day is plotted but has not influenced the fitted lines). Only five of these self-reported heavy smokers developed lung cancer, which is about two-thirds of the number predicted by the straight line and barely one-third of the number predicted by the curve in Fig. 1. Chance may play some part in this shortfall, but several other explanations may be considered, although we cannot test them directly.

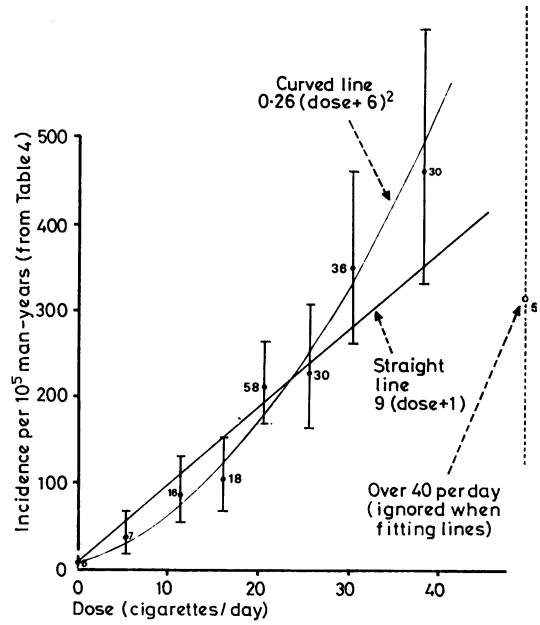


Fig. 1 *Dose-response relationship, standardised for age. The numbers of onsets in each group are given, and 90% confidence intervals are plotted.*

Firstly, it may be that the only men who can really stand smoking 50 or 60 cigarettes/day are those who, because of some aspect of their constitution or their average dose per cigarette, are less affected than the average smoker by the noxious components of smoke.

Secondly, the amount stated to be smoked may for various reasons be an especially misleading measure of the amount actually smoked among those reporting the most extreme habits. For example, one man claimed to have smoked 120 cigarettes/day since the age of three. This was obviously bogus, but a few doctors might enjoy a certain bravado in reporting substantial but not obviously bogus habits to a study of smoking and mortality, and lesser degrees of exaggeration would be undetectable. (Alternatively, a reported extreme consumption at the initial survey might have been a temporary aberration for a few days, weeks, or months only). Occasional misreporting would have little effect on the observed outcomes in the common smoking categories

*In this procedure, joint effects of dose and age were fitted simultaneously by one independent parameter per age group and one per dose group.

†Based on double the improvement in log likelihood as we go from the best linear fit to the plotted curve.

The fact that $(\text{dose}+6)^2$ fits the data significantly better than does a straight line is, however, not proof that exactly 2 stages or processes are strongly affected by smoking, nor even proof that the true relationship is quadratic. It is merely an indication that, as has been hypothesised on other grounds, the dose-response relationship does exhibit some upward curvature in the range 0–40 cigarettes/day. Even if we ignore the very substantial effects of various biases, the algebraic form of the dose-response relationship cannot be uniquely inferred from these data. For example, $(\text{dose}+37)^8$ fits just as well as $(\text{dose}+6)^2$, and indeed any exponent of 2 or more apparently allows adequate fit if a suitable 'background' is first added to the dose.

Table 5 *Age-specific incidence standardised for dose*

| | *Age group (years)* | *O Observed number of onsets* | *E Indirectly dose-standardised expected* | *Relative risk estimated by (a) O/E* | *(b) ML* | *Dose-standardised onset rate/$10^5$ man-years (as in Table 4)* |
|---|---|---|---|---|---|---|
| | 40–44 | 3 | 41·24 | 0·073 | 0·072 | 8 |
| | 45–49 | 7 | 41·06 | 0·170 | 0·170 | 19 |
| | 50–54 | 22 | 38·85 | 0·566 | 0·566 | 63 |
| | 55–59 | 27 | 31·46 | 0·858 | 0·859 | 96 |
| | 60–64 | 40 | 22·50 | 1·778 | 1·787 | 199 |
| | 65–69 | 40 | 14·48 | 2·762 | 2·780 | 309 |
| | 70–74 | 36 | 7·67 | 4·694 | 4·753 | 529 |
| | 75–79 | 26 | 3·74 | 6·952 | 7·097 | 790 |
| Total | 40–79 | 201 | 201·00 | 1·000 | 1·000 | 112 |

but the lung cancer rates among the small proportion of men who really smoke more than 40/day could be appreciably biased by just a few dozen exaggerated claims, or by a few dozen men who do actually light more than 40 cigarettes a day, but who then take rather few puffs per cigarette, either leaving them smouldering for long periods or stubbing them out while still quite long. If the actual intake of smoke by the men who claimed to smoke more than 40/day is really rather similar to that of men who only reported 30/day then no anomaly remains, and it is possible that some such biases do operate. After all, it must be physically quite an achievement to smoke each of 60 cigarettes thoroughly in one single day. Sixty cigarettes/day is about one every 15 minutes of waking life; apart from any other effects, the carboxyhaemoglobin accumulation alone under standard smoking conditions would be nearly sufficient to poison the subject.

Thirdly, there are paradoxical effects of inhalation. Among men smoking 25 to 40/day, those who say they inhale actually get less lung cancer than those who say they do not (Doll and Peto, 1976). The reasons for this are not known, but most lung cancers do arise in the upper bronchi and it might be that deep inhalation actually carries most smoke particles well past this danger area. Whatever the reason may be, the observed protective effect does exist, and although the proportions of self-reported inhalers were the same (70·3% and 70·5%) in those smoking more than 40 and in those smoking 1–40, it might still be that some paradoxical aspect of smoking style reduces the risk for the 0·7% of men who smoke more than 40/day to the risk for men smoking 30 to 40/day.

It is unlikely that the true relationship between cancer risk and exposure of the target cells to smoke flattens off and decreases with increasing dose. Although there is an animal model for such a turnover in leukaemia induced by acute doses of radiation (Major and Mole, 1978), we know of no model for it in animal carcinoma induction by chronic exposure to carcinogens.

## RELATIONSHIP OF INCIDENCE TO AGE

An analysis in which the incidence rates at different ages are indirectly standardised for dose in nine groups (0, 1–4, 5–9, etc., to 35–40) may be performed, yielding Table 5. Again, the relative risk estimates derived by simple indirect standardisation are very similar to those derived more correctly by iterative maximum likelihood, and again we have preferred to multiply these relative risks by 112 to estimate dose-standardised incidence rates in each age group.

Fig. 2 gives three alternative plots of these data against time on a log/log scale, plotting the standardised rates against age, against (age-22·5), or against (age-34). Although the fit of the observed data to a straight line is marginally better if incidence is plotted against (age-22·5) than if it is plotted against (age-34) or just against (age)*, it is clear that, as in animal experiments (Peto and Lee, 1973), a reasonably straight line would result no matter what number of years (between 0 and 34 in these data) we subtract from the age before plotting such graphs. Since for inclusion in the present analysis all the smokers had to start smoking when they were between 16 and 25 years old (mean=19·2 years old), and since once it starts growing a lung cancer probably usually takes only a few years to become clinically evident, (age-22·5) approximately represents the duration of smoking when the lung cancer finally emerged as a truly neoplastic focus and is therefore a physiologically reasonable measure of time.

If the cancers being studied nearly all arose from cells which were completely normal until acted on by cigarette smoke (rather than from cells which suffered early preneoplastic changes spontaneously and which then suffered further changes due to smoking), simple multistage models (Armitage and Doll, 1961; Doll, 1971; Peto, 1977) predict that incidence rates should rise approximately as a power of duration of smoking. If we estimate the duration of

*Relative to the log-likelihood value for the best-fitting straight line, incidence rates proportional to age$^{7\cdot2}$, (age—22·5)$^{6\cdot6}$ and (age-34)$^8$ imply log-likelihood values of —0·5, 0·0 and —0·4 respectively. For any w in the range $0<w<34$, an assumption of incidence rates proportional to (age-w) $^{(7\cdot2-0\cdot12w)}$ would imply a log-likelihood between 0·0 and —0·5 for these data, so statistical considerations alone cannot indicate definitely which range of w is acceptable. Given w, the coefficient of variation of the ML estimate of the exponent of (age-w) is approximately 7% in these data.

smoking at the time of emergence of a truly neoplastic focus as (age-22·5) then the central straight line in Fig. 2 (which has slope 4·49 with standard error 0·31) clearly fits the data excellently. Whether or not the men in this study really started smoking exactly when they claimed to have done, it is clear that the only whole-number exponents of smoking duration which can possibly fit these data are 4 or 5.

In each of these six cases, the constants of proportionality have been estimated from the 195 lung cancers among men aged 40–79 smoking 1–40/day, so the estimates have coefficients of variation of ±7%. (The non-smokers are known to obey a different age distribution (Doll, 1971).)

The age-specific incidence rate and the dose-response relationship have both been obtained on

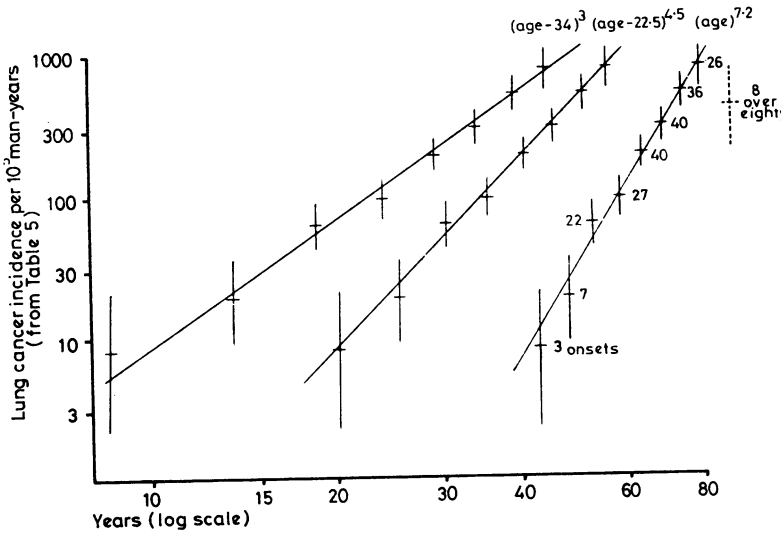


Fig. 2 *Age-specific incidence rates, standardised for dose. The numbers of onsets in each group are given, and 90% confidence intervals are given as vertical lines.*

## SIMULTANEOUS RELATIONSHIP OF INCIDENCE TO AGE AND DOSE

If we take (age-22·5) as a measure of duration of smoking, we have shown that incidence may be proportional to $(\text{age-22·5})^{4 \text{ or } 5}$ but cannot be proportional to any substantially different power of duration of smoking, and we have compared the linear relationship in which incidence is proportional to (dose+1) (where dose=cigarettes/day) with the curved relationship in which incidence is proportional to $(\text{dose}+6)^2$, and found the latter preferable. Combining our dose and time analyses, we have preferred the relationship:

incidence proportional to $(\text{dose}+6)^2 \cdot (\text{Age-22·5})^{4,\ 4\cdot5,\ \text{or}\ 5}$,

and, depending on whether the exponent is 4, 4·5, or 5, the coefficient of proportionality will be estimated from our data as $1{\cdot}74\times10^{-12}$, $0{\cdot}2730\times10^{-12}$, or $0{\cdot}0423\times10^{-12}$.

If we had fitted a linear dose-response relationship, then the fitted linear relationship would have been:

incidence proportional to $(\text{dose}+1)\cdot(\text{Age-22·5})^{4,\ 4\cdot5,\ \text{or}\ 5}$,

and, depending on whether the exponent is 4, 4·5, or 5, the coefficient of proportionality will be estimated as $60{\cdot}4\times10^{-12}$, $9{\cdot}46\times10^{-12}$, or $1{\cdot}46\times10^{-12}$.

the assumption that the age-specific incidence rates for men smoking different amounts are parallel. To test this assumption, we have looked for an 'interaction' between $(\text{dose}+6)^2$ and $(\text{age-22·5})^{4\cdot5}$ among the smokers of 1–40/day aged 40–79. This yielded a chi-square on one degree of freedom of 0·14, indicating that the model fits excellently and suggesting that the age-specific incidence rates for light and heavy smokers are indeed adequately described as being parallel.

## Discussion

As was already known (for example, Doll, 1971), incidence rates in the age range 40–79 are proportional to a power of duration of smoking, in excellent conformity with the predictions of simple multistage model theory. To estimate this power, we must calculate smoking duration by subtracting a chosen quantity (we have chosen 22·5 years) from the age. Having done this, we find that the best-fitting exponent is 4·5±0·3. The adequate fit of this model is not a critical test of simple multistage model theory, of course, but it is nevertheless

gratifying. The three features of note in this analysis are, firstly, that (as with the animal data) the data themselves cannot define the quantity to subtract from age; secondly, that the only two whole number exponents of duration that fit the data naturally are 4 and 5; and thirdly, that if the whole analysis is repeated, including the men aged over 80, then the smooth increase in age-specific rates which we have seen between the ages of 40 and 79 suddenly reverses at the age of 80; even after standardisation for dose, men aged 80–84 or 85+ both have lung cancer incidence rates which are only about half the rates for men aged 75–79. Whether this shortfall is due entirely to under-diagnosis, to selective survival* and to unreported cohort differences in smoking habits during early life, or whether in addition some aspect of the biology of extreme old age does indeed reduce the risk of carcinoma induction, is unclear. (Although under each of our fitted formulae we would expect about two lung cancers among smokers aged 20–39, we actually observed none. We attribute this shortfall entirely to chance, as national lung cancer death rates exhibit no sudden changes at the age of 40.)

The upward curvature of the dose-response relationship is intriguing, and runs rather counter to the linear dose-response relationship estimated by Whittemore and Altshuler (1976) when studying some less carefully restricted data from this study. The minimal claim that we make for it is that it indicates a need to re-examine data from other studies to see if they, too, exhibit upward curvature in the range 0–40 cigarettes/day when attention is restricted to the age range 40–79 among men who have since the age of about 20 smoked only cigarettes. It also indicates a need to discover some way of estimating cigarette smoke dosage to the stem cells of the upper bronchi more accurately than is possible by simple smoking questionnaires. Dosage to the alveoli can of course be estimated by the uptake of CO or nicotine into the blood, either by measuring CO or nicotine directly, or by measuring nicotine breakdown products in the blood; and it would probably be helpful to be able to relate the lung cancer risk to such objective measurements of dose to the periphery of the lung. It is, however, more difficult to imagine reliable methods of assessing the dose deposited on the upper bronchi. Phlegm production in response to cigarette smoke chiefly originates from the upper bronchi, and so might perhaps help us to assess bronchial dose. It has already been reported (Rimington, 1971) that among men who say they smoke similar amounts of tobacco, those with chronic phlegm production from their upper bronchi are at considerably greater age-standardised risk of lung cancer, but it is not clear how much this is due to differences in the effective bronchial dose per cigarette and how much it is due to effects of nature or nurture which predispose to both mucus hypersecretion and cancer.

*Smokers reporting a given daily cigarette consumption differ in the effective amounts they really smoke and perhaps they differ in other correlates of lung cancer in their life-styles or genotypes. Consequently, some are at less risk of lung cancer than the average for such smokers, and those at lower risk are presumably more likely to survive beyond the age of 80. Despite the lack of evidence among men aged 75 to 79 of any such effects, at least some small part of the shortfall in lung cancer beyond the age of 80 may thus be due to selective survival.

## SYSTEMATIC BIASES DUE TO NON-SYSTEMATIC RANDOM ERRORS

The number of cigarettes/day that we record may be a very inaccurate measure indeed of the extent to which bronchial epithelial stem cells are affected, which we shall refer to as the 'true insult'. It is therefore possible that the shape of our graph of

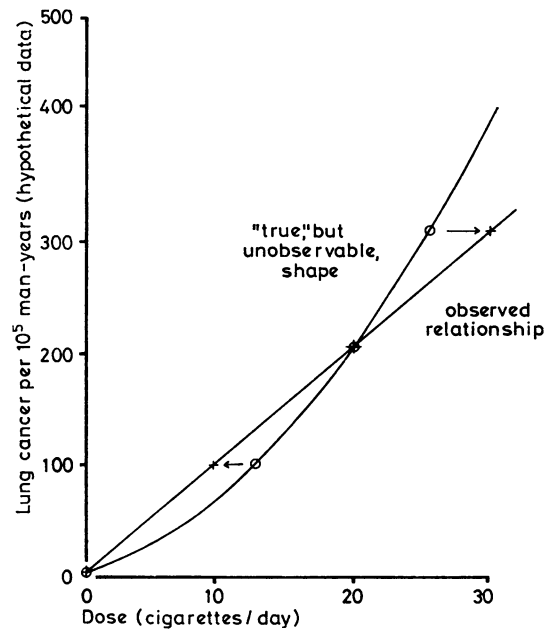


Fig. 3 *Hypothetical dose-response data, with shape distorted.*

After men have been divided on the basis of their current smoking habits into non-smokers, light smokers (mean 10/day), medium smokers (mean 20/day), and heavy smokers (mean 30/day), suppose the mean 'true insult' rate (in arbitrary units) in the four groups is 0, 1·5x, 2x, and 2·5x, and suppose the relation between mean cancer incidence and mean true insult (open circles) is as given by the curved line. The relation between cancer incidence and reported cigarette consumption (plus signs) will then be misleadingly observed as a straight line.

incidence against cigarettes/day might be materially different from the shape of a graph of incidence against 'true insult'. A particular example of different shapes is illustrated in Fig. 3. At first sight, such systematic discrepancies between reported consumption and 'true insult' seem extremely implausible, but in fact they can be shown to arise from quite general and plausible assumptions.

### INDIVIDUAL DIFFERENCES IN MEAN DOSE PER CIGARETTE

For example, the mean 'true insult' per cigarette is certainly very different for different individuals, and if people who usually get less nicotine per cigarette tend to smoke a few more cigarettes to make up for this, then some of them will be removed from the light into the medium group, or from the medium into the heavy group, causing exactly the effects postulated in Fig. 3. Such effects can likewise be expected if smokers who usually smoke a certain number of cigarettes by a certain time of day are at all influenced in the number they usually smoke thereafter by their usual blood nicotine at that time. (Indeed, the biases indicated in Fig. 3 will only be avoided if, implausibly, pharmacologic factors are entirely irrelevant to cigarette consumption.)

### VARIATION DURING LIFE OF DAILY CIGARETTE CONSUMPTION

If men were divided into light, medium, and heavy smokers at the age of 25, then the division would obviously not classify everyone exactly as they would be classified at the age of 65; some of those who are light smokers at 65 were once heavy or medium smokers, while some of those who are heavy smokers at 65 were once light or medium smokers. Thus, the ratio of the mean *lifelong* smoking of those who are light smokers at 65 to that of those who are heavy smokers at 65 is likely to be less extreme than the ratio of the mean current (at 65) consumption of these two groups of men. Since the 'observed consumption' for those who die of lung cancer in an epidemiological study is usually recorded in middle or old age, the biases invoked in Fig. 3 again follow.

### HETEROGENEITY OF THE TARGET POPULATION

If, in the low dose groups, most of the lung cancers arise among people with some special constitutional or occupational synergy with smoking (or with some especially harmful manner of inhaling), and if most such people in the high dose groups die of early lung cancer, then in old age the survivors in the high dose group will be more 'cancer-proof' than average, reducing the upward curvature of the dose-response relationship. (However, such effects might also cause downward curvature in the relationship of incidence with age, and this was not observed in the age range 40–79).

## Conclusion

The general effect of random errors in dosimetry is likely to make the relationship between risk and measured dose less extreme*—that is, to bias a higher powered relationship with dose (for example, incidence proportional to the square of 'true insult') into a lower powered one (for example, incidence directly proportional to recorded consumption). This is a true bias, in that it is not in expectation reduced by doing larger and larger studies. The fact that we have observed statistically significant upward curvature in spite of this bias suggests that with perfect dosimetry we would have obtained even greater upward curvature.

If correct, our suggestion of a quadratic (or higher powered) dose-response relationship in the range 1–40 cigarettes/day has two mechanistic implications. Firstly, the shape of the dose-response relationship ceases to be evidence against the suggestion that two (or more) stages in the production of lung cancer may be strongly affected by smoking. Secondly, it may help explain the well-known fact that lung carcinomas arise chiefly in the upper bronchi. The surface area of the lower bronchi so greatly exceeds that of the upper bronchi that if the total dose deposited from the smoke droplets onto the bronchi between (say) bronchial divisions 1 and 8 were of the same order as that deposited between 8 and 16, then the dose per unit area in the lower bronchi would be much less. If the risk per cell were simply proportional to dose, this would not affect the total risk, for although there would be a far lower dose per cell in the lower bronchi, there would be correspondingly far more cells at risk. If, however, the risk per cell were proportional to dose$^2$, then this dilution of the total dose deposited over a larger area would greatly reduce the total risk, and might partly or wholly account for the fact that bronchial carcinomas chiefly arise in the upper bronchi.

Reprints from Sir Richard Doll, Regius Professor of Medicine, Radcliffe Infirmary, University of Oxford.

The study was started in collaboration with Sir Austin Bradford Hill. Barbara Hafner collected

*There is an analogy here with the fact that in the standard least squares regression of y on x, random errors in x will bias the regression coefficient towards zero, if we imagine that we are estimating the exponent of dose to which incidence is proportional by examining the regression coefficient of y=log (incidence) on x=log (recorded consumption)=log (true insult)+error.

and maintained the records, assisted by Mrs. Sutherland, Mrs. Norton, and Mrs. Thompson. Richard Gray assisted with computing, and Ruth Rohrbasser typed this report.

**References**

Armitage, P. (1971). Discussion of 'The Age Distribution of Cancer'. *Journal of the Royal Statistical Society*, series A, **134**, 155–156.

Armitage, P., and Doll, R. (1961). Stochastic models for carcinogenesis. *Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 19–38. University of California Press: Berkeley.

Doll, R. (1971). The Age Distribution of Cancer: implications for models of carcinogenesis (with discussion). *Journal of the Royal Statistical Society*, series A, **134**, 133–166.

Doll, R., and Hill, A. B. (1964). Mortality in relation to smoking: ten years observations of British doctors. *British Medical Journal*, **1**, 1399–1410 and 1460–1467.

Doll, R., and Peto, R. (1976). Mortality in relation to smoking: 20 years observations on male British doctors. *British Medical Journal*, **2**, 1525–1536.

Major, I. R., and Mole, R. H. (1978). Myeloid leukaemia in X-ray irradiated mice. *Nature*, **272**, 455–456.

Peto, R. (1977). Epidemiology, multistage models and short-term mutagenicity tests. In: *Origins of Human Cancer*, pp. 1403–1428. Cold Spring Harbor Publications: New York.

Peto, R., and Lee, P. N. (1973). Weibull distributions for continuous-carcinogenesis experiments. *Biometrics*, **29**, 457–470.

Rimington, J. (1971). Smoking, chronic bronchitis and lung cancer. *British Medical Journal*, **2**, 373–375.

Whittemore, A., and Altshuler, B. (1976). Lung cancer incidence in cigarette smokers: further analysis of Doll and Hill's data for British physicians. *Biometrics*, **32**, 805–816.