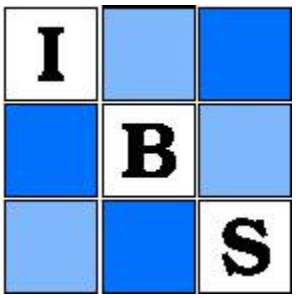

BIOMETRICS *32*, 805–816
DECEMBER 1976

Lung Cancer Incidence in Cigarette Smokers: Further Analysis of Doll and Hill's Data for British Physicians

ALICE WHITTEMORE[1] and BERNARD ALTSHULER

Institute of Environmental Medicine, New York University Medical Center,
550 First Avenue, New York, New York, 10016, U.S.A.

Summary

Doll's analysis of lung cancer incidence (mortality) in cigarette smokers is refined in more detail in this paper. His conclusion that incidence is approximately proportional to rate of smoking and the fifth power of years of smoking is shown to hold in each of several age and dose groups giving additional evidential support. The data are equally well fitted by a Weibull distribution and by a lognormal distribution with a constant geometric standard deviation. There is, however, a trend away from these fits which needs to be examined in other studies.

1. Introduction

The best dose-response data for human cancer is that obtained by Doll and Hill for lung cancer and cigarette smoking in British physicians (Doll [1971]). In his analysis, Doll reported that incidence (age-specific lung cancer mortality rate) is approximately proportional to rate of smoking and to the fifth power of smoking duration, as shown by age and dose standardized averages. There was no indication of a dose threshold (zero risk at smaller doses) or a guaranteed time delay (zero risk at earlier times).

Mathematical models of the dose-response relationships have not been reliable enough to allow an accurate prediction of cancer risks from low levels of environmental carcinogens (Hoel *et al.* [1975]). Consequently, such risk assessment has been dominated by over-conservative methods which give excessively high upper limits in the following sense: given a confidence coefficient, the probability that the upper limit is above and not below the actual risk is greater than the given confidence coefficient by some unknown amount. These methods consist of the conservative one-hit model in which, at low levels, response is proportional to dose (Hoel *et al.* [1975]; FDA [1971]) and the log-probit model with a ten-fold reduction in response per probit unit, chosen as a shallower slope than the generally observed slopes (Mantel *et al.* [1975] and Mantel and Bryan [1961]). It is, therefore, a critical matter to examine the fits of the related temporal distributions, Weibull and lognormal. The Weibull distribution is associated with the concept that the primary effect of dose is on the rate of transformation of cells and the unpredictability of time-to-cancer is due to a stochastic process taking place within the individual. The lognormal distribution is associated with the concept that time-to-cancer is determined by the dose for each individual and the unpredictability of time-to-cancer is due to unrecognized differences among individuals.

[1] At N.Y.U. Institute of Environmental Medicine, on leave from Hunter College, C.U.N.Y., with the SIAM Institute for Mathematics and Society under the support of the Alfred P. Sloan Foundation.

Key Words: Lung cancer; Cigarette smoking; Dose-response; Weibull distribution; Lognormal distribution; Risk extrapolation.

In the Doll and Hill study, smoking histories were obtained from over 34,000 male doctors in Britain in 1951, and changes in smoking habits were recorded in 1958 and at the termination of the study in 1966. Doll's paper gave the number of lung cancer deaths and number of man-years at risk among the cigarette smokers in several age and dose groups after culling out those who reported changes in smoking habits during the 17 years of observation (Table 1). Since there was incomplete recording for ages over 80 years, these data have not been used in this note. Though data were obtained on the start of smoking, which occurred between 15 and 25 years of age, they were not published. Here the data are treated as if everyone started smoking at age 20, and time is measured exclusively in years of smoking. The correction required for background (non-smoker) incidence is small and has been neglected. Indeed the lung cancer incidence in nonsmokers at a particular age t years is at most 0.015 times the incidence, standardized for amount smoked, of those who have been smoking for t years. If background exposure is equivalent to cigarette smoke, this corresponds to a smoking rate of 0.3 cigarettes per day, according to the analysis given in this paper.

2. *Definitions and General Approach*

With epidemiological data of this type, corrections for causes of mortality other than the primary cancer are absorbed in the data giving man-years at risk. Thus, we are concerned here with estimating the corrected distribution of time to lung cancer (death) for the theoretical situation where there are no other competing causes of death. Let T_x represent time to lung cancer at dose rate x, and let $F(t; x) = \Pr\{T_x \leq t\}$ denote the probability distribution function which gives the probability that T_x is less than or equal to t. Equivalent to $F(t; x)$ is the cumulative hazard

$$H(t; x) = -\log(1 - F(t; x)) \tag{2.1}$$

Table 1

MAN-YEARS AT RISK, NUMBER OF CASES OF LUNG CANCER (NUMBER OF CASES IN PARENTHESES), AND STANDARDIZED INCIDENCES

Cigarettes per day	1-9 5.2	10-14 11.2	15-19 15.9	20-24 20.4	25-34 27.4	35+ 40.8	Total	Incidence, standardized for dose, per 100,000 man-years
Years of Smoking (Age less 20 yrs.)								
15-19	3121	3577	4317	5683	3042	670	20,410	0
20-24	2937	3286 (1)	4214	6385 (1)	4050 (1)	1166	22,038 (3)	13.8
25-29	2288	2546 (1)	3185	5483 (1)	4290 (4)	1482	19,274 (6)	30.0
30-34	2015	2219 (2)	2560 (4)	4687 (6)	4268 (9)	1580 (4)	17,329 (25)	134.9
35-39	1648 (1)	1826	1893	3646 (5)	3529 (9)	1336 (6)	13,876 (21)	129.3
40-44	1310 (2)	1386 (1)	1334 (2)	2411 (12)	2424 (11)	924 (10)	9,789 (38)	359.6
45-49	927	988 (2)	849 (2)	1567 (9)	1409 (10)	556 (7)	6,296 (30)	458.0
50-54	710 (3)	684 (4)	470 (2)	857 (7)	663 (5)	255 (4)	3,639 (25)	700.0
55-59	606	449 (3)	280 (5)	416 (7)	284 (3)	104 (1)	2,139 (19)	1132.9
Total	15,562 (6)	16,961 (14)	19,102 (15)	31,135 (48)	23,959 (52)	8,073 (32)	114,792 (167)	
Incidence, standardized for duration, per 100,000 man-yrs.	33.3	76.9	101.1	170.9	220.2	347.6		

and the hazard or incidence rate

$$h(t; x) = \frac{d}{dt} H(t; x). \tag{2.2}$$

For a given dose rate x and time t, $h(t; x)$ is estimated from the data in a small interval of time at t by the number of lung cancers per number of man-years at risk.

The Weibull distribution is most easily described in terms of the hazard function which is posited here to be of the form

$$h(t; x) = bx^{k_1}t^{k_2} \tag{2.3}$$

where b, k_1 and k_2 are constants that are independent of dose rate. Note hazard is positive at all values of time and dose rate other than zero. The Weibull probability distribution function is therefore

$$F(t; x) = 1 - \exp[-(b/(k_2 + 1))x^{k_1}t^{k_2+1}]. \tag{2.4}$$

The lognormal distribution used here is most easily described by stating that $\log T_x$ is normally distributed about a mean of $(1/n) \log (c/x)$ with a constant standard deviation of $\log \sigma_g$ where c, n and σ_g are constants that are independent of dose rate. Median cancer time t_{50} is thus assumed to be related to the dose rate by the equation

$$t_{50}{}^n x = c. \tag{2.5}$$

The lognormal probability distribution function is therefore

$$F(t; x) = \frac{1}{\sqrt{2\pi} \log \sigma_g} \int_{-\infty}^{\log t} (1/u) \exp\left[-(1/2)\left(\frac{\log u - \log t_{50}}{\log \sigma_g}\right)^2\right] du. \tag{2.6}$$

Because of empirical and theoretical considerations, these two distributions of cancer times have been extensively discussed in the literature (Albert and Altshuler [1973], Chand and Hoel [1974], Gehan [1969], Peto and Lee [1973], Peto, Lee and Paige [1972] and Pike [1966]).

Each of these distributions, which imply some risk at all positive times and dose rates, are fitted to the data in two steps. The first step uses a grouped form of the data in which the data points, because they represent estimates of hazard rather than cumulative hazard, are independent of each other. Plotting the data in this form, straight lines can be fitted by eye to obtain slopes that estimate the parameters, other than a constant of proportionality, and to obtain a visual appreciation of residuals. A least square solution could also have been used. In the second step, the constant of proportionality is determined by the requirement that the observed number of cases equals the number of cases predicted by the fit. With this simplified approach, interval estimation of the parameters of each of the two distributions can be taken as the range of statistical consistency determined by a χ^2 test of goodness-of-fit. The method avoids the difficulty that arises in plots of cumulative data (such as cumulative hazard) in which each point depends on preceding points and the lack of independence of data points interferes with the interpretation of the residuals and the fit.

Although one could also ask for best (by some statistical criterion) estimates of the parameters and confidence intervals, the procedures are not easily available, nor are they necessary for the conclusions obtained from this data. A fourth positive parameter w corresponding to a guaranteed time delay can be introduced in both the Weibull and lognormal probability distribution functions (2.4) and (2.6) by replacing the time variable t by $t - w$ for times t greater than or equal to w, and setting $F(t; x)$ equal to zero for times

t less than w. It will be seen, however, from the following analysis that both distributions provide adequate fits to the data without the use of the additional parameters. The credibility of the conclusions is reinforced by the separate fits of the data grouped by years of smoking and rate of smoking, and the validity of the approach is demonstrated by the graphical indication of a dose rate dependence of curvature that is inconsistent with the dose-rate dependence predicted by each of the fitted distributions.

3. *The Weibull Fit*

Duration-standardized incidence $h(\cdot\,; x_j)$ at dose rate x_j and dose-standardized incidence $h(t_i\,;\,\cdot)$ for smoking duration t_i are defined by

$$h(\cdot\,; x_j) = (1/L) \sum_i L_i h(t_i\,; x_j) \tag{3.1}$$

$$h(t_i\,;\,\cdot) = (1/L) \sum_j L^j h(t_i\,, x_j) \tag{3.2}$$

where the indices i and j range over the nine smoking durations and six dose rate levels into which the data were grouped, L_i and L^j represent the total number of man-years at risk for duration t_i and dose rate x_j, respectively, and $L = \sum_i L_i = \sum_j L^j$ is the total number of man-years at risk in the study. Substituting the posited Weibull form (2.3) into (3.1) and (3.2) yields

$$h(\cdot\,; x_j) = bC_t x_j^{k_1} \tag{3.3}$$

$$h(t_i\,;\,\cdot) = bC_x t_i^{k_2}, \tag{3.4}$$

where $C_t = (1/L) \sum_i L_i t_i^{k_2}$ and $C_x = (1/L) \sum_j L^j x_j^{k_1}$. Hence by (3.3) and (3.4) duration-standardized incidence $h(\cdot\,; x_j)$ versus dose rate x_j and dose-standardized incidence $h(t_i\,;\,\cdot)$ versus duration t_i plot on log-log scales as straight lines of slopes k_1 and k_2, respectively. The estimates $k_1 = 1.10$ and $k_2 = 4.68$ were obtained by eye fits of straight lines to the log-log plots of the standardized incidences as shown in Figure 1, using the man-years at risk in Table 1. The constant $b = 2.80 \times 10^{-12}$ was calculated so as to make the total predicted (expected) number of cancers equal to the total observed number of 167; specifically, b satisfies

$$b \sum_{ij} L_{ij} x_j^{1.10} t_i^{4.68} = 167, \tag{3.5}$$

where t_i is the midpoint of the ith time interval, x_j is the average daily cigarette consumption for the jth smoking group as given in Doll's paper and Table 1, and L_{ij} represents the number of man-years at risk in the ith time period and jth smoking group as shown in Table 1.

The observed and expected frequencies in Table 2 correspond to age and dose groups chosen so that there are approximately five or more expected cancers in each cell. A χ^2 test based on the differences between observed and predicted cancers in these enlarged cells gave a value of 16.6 with corresponding probability of 49 percent, showing a satisfactory fit.

Although not shown, Weibull fits with $k_1 = 1$ and $k_1 = 2$ were obtained (where $k_2 = 4.68$ and the values of b were adjusted to equate total observed and predicted cancers) and χ^2 tests were performed on these fits. For $k_1 = 1$ the probability was 45 percent that χ^2 is greater than the observed value while for $k_1 = 2$ this probability was less than two percent. Thus $k_1 = 1$ gave a satisfactory fit and $k_1 = 2$ was excluded. This statistical analysis means that the data is consistent with Doll's conclusion that human risk is proportional

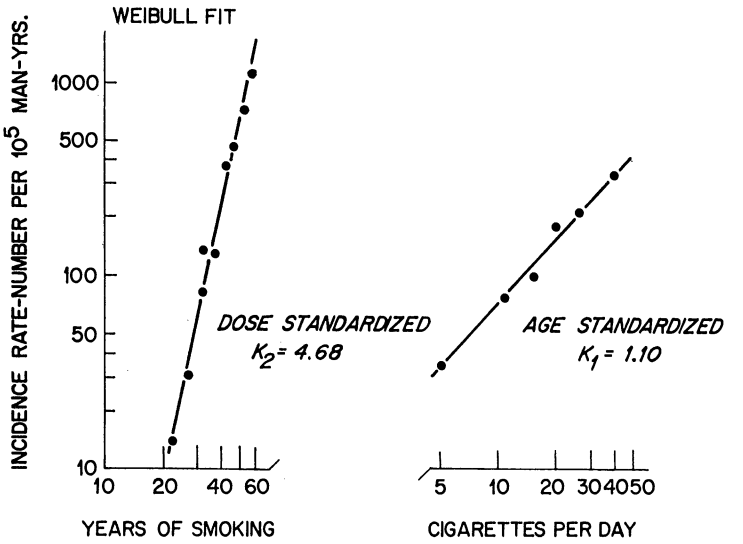

Figure 1

LOG-LOG PLOTS OF DOSE-STANDARDIZED LUNG CANCER INCIDENCE RATE VERSUS YEARS OF SMOKING t (LEFT) AND AGE-STANDARDIZED INCIDENCE RATE VERSUS SMOKING RATE x (RIGHT). THE STRAIGHT LINES WITH SLOPES 4.68 AND 1.10, RESPECTIVELY, DETERMINE THE PARAMETERS k_2 AND k_1 IN THE WEIBULL HAZARD (INCIDENCE RATE) FUNCTION $h(t; x) = bt^{k_1}x^{k_2}$.

Table 2

OBSERVED AND EXPECTED NUMBER OF CASES OF LUNG CANCER WITH ENLARGED INTERVALS

Weibull Fit: $b = 2.80 \times 10^{-12}$, $k_1 = 1.10$, $k_2 = 4.68$
Lognormal Fit: $n = 4.49$, $\sigma_g = 1.48$, $\ln c = 23.2713$

Cigarettes per day	1-19 (11.2)			20-24 (15.9)			25-34 (20.4)			35- (40.8)			Total		
Years of Smoking	Obs	Exp_W	Exp_L	Obs	Exp_W	Exp_L	Obs	Exp_W	Exp_L	Obs	Exp_W	Exp_L	Obs	Exp_W	Exp_L
15-39 (26.8)	8	6.25	4.31	8	7.93	7.33	14	9.05	9.89	4	4.93	6.14	34	28.16	27.67
35-39 (37.6)	1	4.93	5.03	5	6.54	7.75	9	8.81	10.90	6	5.14	6.71	21	25.42	30.39
40-44 (42.6)	5	6.52	6.95	12	7.77	8.92	11	10.81	12.45	10	6.38	7.12	38	31.48	35.44
45-49 (47.6)	4	7.39	7.71	9	8.50	8.90	10	10.57	10.68	7	6.46	6.01	30	32.92	33.30
50-59 (54.6)	17	15.23	13.77	14	12.95	10.94	8	13.16	10.33	5	7.69	5.18	44	49.03	40.22
Total	35	40.32	37.77	48	43.69	43.84	52	52.40	54.25	32	30.60	31.16	167	167.01	167.02

$\chi^2_W = 16.65$, $\text{Prob}\{\chi^2_{17} \geq 16.65\} = 0.49$

$\chi^2_L = 17.37$, $\text{Prob}\{\chi^2_{17} \geq 17.37\} = 0.43$

to dose rate (Doll [1971]) and inconsistent with mouse skin painting experiments in which incidence is proportional to the square of dose rate (Lee and O'Neill [1971] and Altshuler *et al.* [1971]).

We next examine the question whether more detailed examination of the data might suggest either other forms of the Weibull hazard function (e.g., the dose dependence of k_2), or else systematic trends in the data that would be obscured in the averaging process. For this purpose the data of Table 2 was plotted in the manner shown in Figure 2. The lines in Figure 2 represent the Weibull distributions which were fitted to these data as described at the beginning of this section. In plotting, the data for the ith time interval and jth dose interval were plotted at the value τ_i and y_j indicated beneath those intervals in Table 2, and not at the midpoints of those intervals. The time τ_i was chosen so that the value of the Weibull hazard at τ_i was equal to the weighted average of the Weibull hazard over the ith time interval. The weights were proportional to the number of man-years at risk in each of the subintervals of interval i. The dose values y_j were chosen analogously.

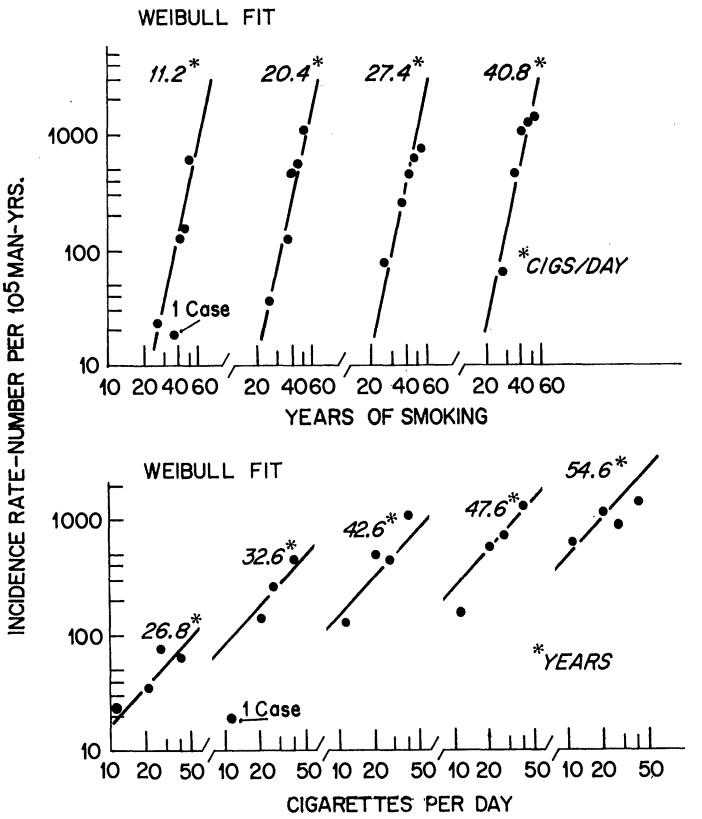

Figure 2

LOG-LOG PLOTS OF LUNG CANCER INCIDENCE RATE VERSUS YEARS OF SMOKING t FOR EACH OF FOUR SMOKING GROUPS (ABOVE), AND VERSUS SMOKING RATE x FOR FIVE SMOKING DURATION GROUPS (BELOW). STRAIGHT LINES INDICATE THE FITTED WEIBULL HAZARD (INCIDENCE RATE) FUNCTION $h(t; x) = 2.80 \times 10^{-12} x^{1.10} t^{4.68}$.

One trend in the data that appears in the upper graphs representing time dependence at different dose rates is the apparent change in curvature from upward to downward concavity as dose increases. (In evaluating curvature, the point representing one case of cancer has been ignored as too unreliable.) Though small, this change is monotonically progressive, and therefore a simple rank statistic shows that this would happen by chance with a probability of $1/4! = 0.04$ under the null hypothesis of no change in curvature. The Weibull hazard starting at zero time has the constant curvature of zero and so is contradicted by the observed trend in concavity which is significant at the five percent level. The lognormal is similarly contradicted but the argument is somewhat more difficult because the downward concavity of the lognormal decreases for smaller values of incidence. As will be seen in Figure 5, the downward concavity of the lognormal changes only slightly in the range of data, and thus there is also a significant difference between the observed trend in downward concavity and that predicted by the lognormal.

The relatively low incidence at high doses and durations could be an artifact due to the same incompleteness of data that was reported to affect those over 80 years of age (Doll [1971]). On the other hand, it could be due to cell toxicity at higher smoking rates. It points to either a higher value of k_2 at the lower doses, or else at higher doses to a time delay before tumors can appear, indicating that heavier smoking somehow decelerates progression toward tumor. If k_2 increases as dose decreases, then at very low doses the consequences of the above model would be modified; in particular, there would be a higher probability of cancer and a later average cancer time among cancer cases. Because trends in this range may be magnified when extrapolating to much lower dose rates, the presence or absence of this change in curvature should be examined in the analysis of other and future studies.

4. The Lognormal Fit

In this section the lognormal distribution, with a slope of one probit for a six-fold change in dose, is shown to fit the data as well as the Weibull distribution discussed in Section 3. As stated in Section 2, we have assumed for the lognormal distribution that $\log T_x$ is normally distributed with a mean equal to the logarithm of median cancer time t_{50} given by

$$\log t_{50} = (1/n) \log (c/x) \tag{4.1}$$

with a standard deviation of $\log \sigma_g$. We first show how the parameters n and σ_g were estimated. For a fixed dose rate x the standard normal deviate $z(x, t)$ corresponding to percent cancer by time t is given by

$$z(x, t) = (\log t - \log t_{50})/\log \sigma_g \tag{4.2}$$

so that $\log t$ plots against $z(x, t)$ as a line of slope $\log \sigma_g$. Using the procedure described in the appendix, observed incidence was cumulated over dose for each of the eight duration groups with non-zero incidence (see Table 1). From this cumulative incidence a percent cancer was calculated by (2.1) for each duration group. The standard normal deviates corresponding to these values were then plotted against the logarithm of smoking duration, as shown in the upper graph of Figure 3. Calculating the slope of a line eye fitted to the data yielded $\sigma_g = 1.48$. Substitution of the right-hand side of (4.1) for $\log t_{50}$ in (4.2) yields, after taking the factor $1/n$ outside of the parenthesis,

$$z(x, t) = [\log x - \log (c/t^n)]/(n \log \sigma_g). \tag{4.3}$$

Thus, for fixed time t, log x plots against the standard normal deviate representing percent cancer for dose rates up to x as a line of slope n log σ_g. Incidence was cumulated over time for each smoking group and from these values percent cancer at each dose was computed and displayed in the lower graph of Figure 3. By eye fitting a line to this data and using the estimate $\sigma_g = 1.48$ obtained above, n was estimated at 4.49. Like the Weibull parameter b, the parameter $c = e^{23.2713}$ was obtained not from these lines but rather calculated so that total predicted number of cancers equals the total observed number. Specifically, c was obtained by linear interpolation to satisfy

$$\sum_{ij} L_{ij} h_c(t_i \ ; x_j) = 167, \tag{4.4}$$

where L_{ij} , x_j and t_i are given as in (3.5), and where $h_c(t_i \ ; x_j)$ represents the lognormal hazard at t_i and x_j corresponding to the parameters $n = 4.49$ and $\sigma_g = 1.48$.

If this lognormal fit were used to extrapolate incidence down to lower doses, the value $\sigma_g^n = 5.75$ indicates that one would use a line of slope one probit per approximately six-fold

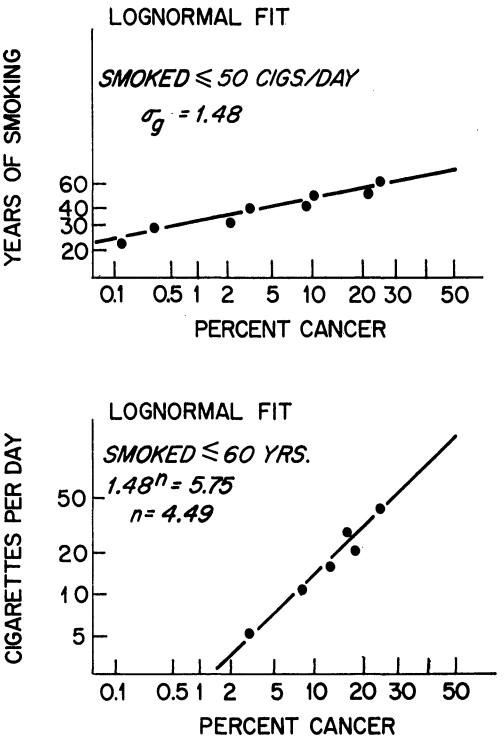

Figure 3

PERCENT CANCER ON A PROBIT SCALE VERSUS LOGARITHM OF YEARS OF SMOKING FOR ALL SMOKERS (ABOVE) AND VERSUS SMOKING RATE FOR ALL DURATION GROUPS (BELOW). THE SLOPES OF THE LINES SHOWN WERE USED TO DETERMINE THE VALUES 1.48 FOR THE GEOMETRIC STANDARD DEVIATION σ_g AND 4.49 FOR THE PARAMETER n IN THE LOGNORMAL MODEL THAT ASSUMES $t_{50}{}^n x$ = CONSTANT, WHERE t_{50} IS MEDIAN CANCER TIME.

reduction in dose, which is less than the more conservative ten-fold reduction of the Mantel-Bryan procedure (Mantel and Bryan [1961]).

Table 2 also shows the comparison between observed and expected numbers of cancer using this lognormal fit. A χ^2 test was calculated indicating, as shown at the bottom of Table 2, essentially the same level of agreement as for the Weibull fit.

Figure 4 is analogous to Figure 2 in displaying time dependence for each of the dose groups and dose dependence for each of the time groups; it differs from Figure 3 in that the latter displays time and dose dependence over all the dose and time groups, respectively. In the graphs at the left of Figure 4, incidence was cumulated over dose for each of the duration groups by the procedure outlined in the appendix, converted to percent and plotted against log dose rate. The lines represent the lognormal fit. The graphs on the right show percent cancer as computed from incidence which was cumulated over time for each smoking group. The same enlarged intervals were used throughout the analysis. For comparison with the Weibull fit, as shown in Figure 2, Figure 5 shows the lognormal fit to observed incidence rates on a log-log plot with the fitted incidence rates given by the dotted lines. As was stated in the previous section, the difference between the observed concavity and that of the fitted curve in the upper plot is monotonically progressive with change in smoking rate, indicating a significant trend away from the lognormal fit.

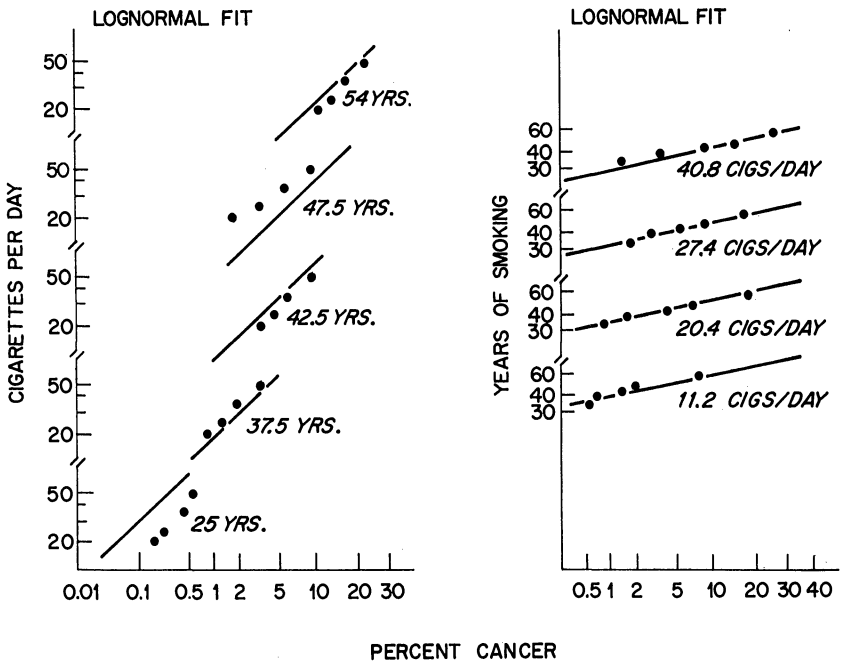

Figure 4

LOG-PROBIT PLOTS OF PERCENT CANCER AGAINST SMOKING RATE FOR EACH OF FIVE DURATION GROUPS (LEFT) AND AGAINST YEARS OF SMOKING FOR FOUR SMOKING GROUPS (RIGHT). THE STRAIGHT LINES REPRESENT THE LOGNORMAL DISTRIBUTION FUNCTION WITH CONSTANT GEOMETRIC STANDARD DEVIATION $\sigma_g = 1.48$ AND WITH MEDIAN CANCER TIME t_{50} GIVEN BY $t_{50}^{4.49}x = e^{23.2713}$.

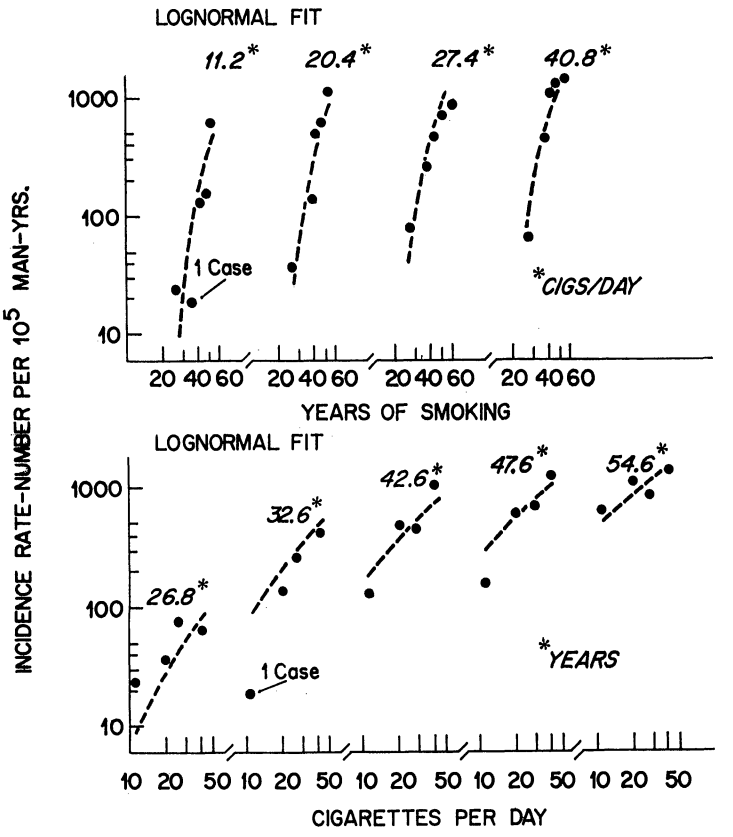

Figure 5

LOG-LOG PLOTS OF LUNG CANCER INCIDENCE RATE VERSUS YEARS OF SMOKING FOR FOUR SMOKING GROUPS (ABOVE), AND VERSUS SMOKING RATE FOR FIVE DURATION GROUPS (BELOW). THE DOTTED LINES REPRESENT THE FITTED LOGNORMAL HAZARD (INCIDENCE RATE) FUNCTION WITH $\sigma_g = 1.48$ AND WITH t_{50} DETERMINED BY $t_{50}^{4.49}\, x = e^{23.2713}$

5. *Conclusion*

Both the Weibull and lognormal distributions approximate the data equally well in the given dose and time ranges. However, the consequences of using them to extrapolate to very low doses are quite different (Hoel *et al.* [1975]). One consequence is that the probability of cancer at a given low dose is considerably higher for the Weibull model than for the lognormal model. Another consequence for low incidence in the population is that the average age of cancer among cancer cases increases as dose decreases according to the lognormal model, while under the assumptions of the fitted Weibull distribution it remains independent of dose. If, in the log-log plot of incidence versus time, the observed trend to increasing upward concavity as dose rate decreases is real, the lognormal distribution with constant geometric standard deviation, which is always concave down, will become a worse approximation than the Weibull with zero guarantee time at very low doses. This observed trend needs to be verified in other studies and with the additional data being accumulated in the continuing study of lung cancer incidence in British physicians.

Cigarette smoking may be a special type of carcinogenic agent and not generally representative of other agents. It promotes in animal skin painting and has a strong synergistic effect when combined with radiation exposure in uranium miners and asbestos exposure. Nevertheless, these quantitative relationships are important for their relevance to the general extrapolation problem. They are by far the best dose-response results on human cancer and are consistent with the one-hit model for risk extrapolation. The one-hit model is conservative in that it predicts higher risks than other extrapolation models and should not be ruled out without positive evidence to the contrary.

Acknowledgments

This investigation was part of center programs supported by Grant ES-00260 from the National Institute of Environmental Health Sciences and Grant CA-13343 from the National Cancer Institute.

Incidence sur le Cancer du Poumon pour les Fumeurs de Cigarette: Nouvelle Analyse des Données de Doll et Hill pour des Physiciens Britaniques

Résumé

L'analyse de Doll pour l'incidence du cancer du poumon (mortalité) des fumeurs de cigarette est affinés dans cet article. On confirme sa conclusion que l'incidence est approximativement proportionnelle au taux de tabagie et à la cinquième puissance du nombre d'années de tabagie; cette conclusion mieux confirmée est valable dans chaque classe d'âge et pour chaque groupe de fumeurs. Les données sont aussi bien ajustées à une distribution de Weibull qu'à une distribution lognormale avec un écart type géométrique constant. Il y a cependant une tendance à s'écarter de ces ajustements qui demande à être examinée dans d'autres études.

References

Albert, R. and Altshuler, B. [1973]. Considerations relating to the formulation of limits for unavoidable population exposures to environmental carcinogens, Appendix. *Radionuclide Carcinogenesis*, AEC Symposium Series, Conf-72050, National Technical Information Service, 233–53.

Altshuler, B., Klassen, A., Troll, W. and Orris, L. [1971]. Tumor response in mouse skin from repeated applications of 1, 2, 5, 6–dibenzanthracene. *Proceedings of American Association for Cancer* Research *12*, 49.

Chand, N. and Hoel, D. G. [1974]. A comparison of models for determining safe levels of environmental agents. *Reliability and Biometry*. F. Proschan and R. J. Serfling, (ed.), SIAM, 681–700.

Doll, R. [1971]. The age distribution of cancer: implication for models of carcinogenesis. *Journal of the Royal Statistical Society, Series A, 134*, 133–66.

Food and Drug Administration Advisory Committee on Protocols for Safety Evaluation. [1971]. Panel on carcinogenesis report on cancer testing in the safety evaluation of food additives and pesticides. *Toxicology and Applied Pharmacology 20*, 419–38.

Gehan, E. A. [1969]. Estimating survival functions from the life table. *Journal of Chronic Diseases 21*, 629–44.

Hoel, D. G., Gaylor, D. W., Kirschstein, R. L., Saffiotti, U. and Schneiderman, M. A. [1975]. Estimation of risks of irreversible, delayed toxicity. *Journal of Toxicology and Environmental Health 1*, 133–51.

Lee, P. N. and O'Neill, J. A. [1971]. The effect both of time and dose applied on tumor incidence rate in benzpyrene skin painting experiments. *British Journal of Cancer 25*, 759–70.

Mantel, N. and Bryan, W. R. [1961]. "Safety" testing of carcinogenic agents. *Journal of the National Cancer Institute 27*, 455–70.

Mantel, N., Bohidar, N., Brown, C., Ciminera, J. and Tukey, J. [1975]. An improved "Mantel-Bryan" procedure for "safety" testing of carcinogens. *Cancer Research 35*, 865–72.

Peto, R. and Lee, P. N. [1973]. Weibull distributions for continuous-carcinogenesis experiments. *Biometrics 29*, 457–70.

Peto, R., Lee, P. N. and Paige, W. S. [1972]. Statistical analysis of the bioassay of continuous carcinogens. *British Journal of Cancer 26*, 258–61.

Pike, M. C. [1966]. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics 22*, 142–61.

Received May 1975., *Revised December* 1975

Appendix

Computations Used to Determine the Lognormal Parameters n and σ_g

Let T_x represent single-risk time to cancer when continuously exposed to dose rate x, and let $F(t, x) = \Pr\{T_x \leq t\}$. For a fixed dose rate x the age-specific incidence rate at time t, given by the number of cancers in an interval Δt per number of man-years at risk in Δt, is an estimate of

$$h(t, x) = \frac{d}{dt}(-\log(1 - F(t, x))).$$

The cumulative hazard $H(t, x) = \int_0^t h(\tau, x)\, d\tau$ is approximated from the data by summing the products of the age-specific incidence rates and corresponding time intervals Δt over all times less than or equal to t. Under the assumptions of the lognormal distribution stated in Section 2,

$$\frac{\partial H}{\partial x}(t, x) = h(t, x)t/(nx);$$

thus

$$H(t, x_0) = \int_0^{x_0} \frac{\partial H}{\partial x}(t, x)\, dx = \int_0^{x_0} h(t, x)t/(nx)\, dx$$

may also be approximated from the data by summing the products of age-specific incidence rates scaled by $t/(nx)$ and corresponding dose intervals Δx over all dose rates x less than or equal to x_0.

Since $\log T_x$ is assumed to be normally distributed about a mean of $(1/n) \log (c/x)$ with a standard deviation of $\log \sigma_g$, it follows that for a given dose rate x, sample points from this distribution of the form $(t, F(t, x))$ where $F(t, x)$ represents percent cancer for durations less than or equal to t, will plot on log-probit scales as a line of slope $\log \sigma_g$. Moreover, for fixed exposure duration t, the standard normal deviate $(\log t - \log t_{50})/\log \sigma_g$, with $t_{50}{}^n x = c$, can also be written as $(\log x - \log x_{50})/(n \log \sigma_g)$, where x_{50} is the 50 percent effective dose rate given by $t^n x_{50} = c$; see (4.3). Hence sample points from this distribution of the form $(x, F(t, x))$ plot on log-probit scales as a line of slope $n \log \sigma_g$.

The parameters n and σ_g were estimated as follows:

i) A trial value of n was obtained from the Weibull parameters $k_1 = 1.10$ and $k_2 = 4.68$ as $n = (k_2 + 1)/k_1 = 5.16$. The relation $n = (k_2 + 1)/k_1$ is obtained from the Weibull distribution by using (2.4) to determine median time t_{50} and requiring as in (2.5) that $t_{50}{}^n x$ is constant.

ii) Using this value of n, observed incidence rates were summed as described above over dose rates less than or equal to $x_0 = 50$ cigarettes per day to estimate $H(t, 50)$ and thus $F(t, 50) = 1 - \exp(-H(t, 50))$ for each duration group t.

iii) Log σ_g was estimated as the slope of a line eye fitted to the plot of $\log t$ versus the standard normal deviate of $F(t, 50)$.

iv) Observed incidence rates were summed as described above over durations up to 60 years to estimate $H(60, x)$ and then $F(60, x)$ for each dose rate x.

v) $n \log \sigma_g$ was estimated as the slope of a line eye fitted to the plot of $\log x$ versus the standard normal deviate of $F(60, x)$ and a value for n was then determined using the value of $\log \sigma_g$ obtained from (iii).

vi) Steps (ii)–(v) were repeated with the new value of n. The process was repeated until convergence for n was obtained, with the final values $\sigma_g = 1.48$, $n = 4.49$. The corresponding lines are shown in Figure 3.