FUNDAMENTAL AND APPLIED TOXICOLOGY 6, 263–269 (1986)

An Empirical Comparison of Methods Used to Estimate Carcinogenic Potency in Long-Term Animal Bioassays: Lifetable vs Summary Incidence Data

LOIS SWIRSKY GOLD,*,[1] LESLIE BERNSTEIN,† JOHN KALDOR,*
GEORGANNE BACKMAN,* AND DAVID HOEL‡

*Biology and Medicine Division, Lawrence Berkeley Laboratory, Berkeley, California 94720; †Department of Preventive Medicine, University of Southern California School of Medicine, Los Angeles, California 90033; and ‡National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

An Empirical Comparison of Methods Used to Estimate Carcinogenic Potency in Long-Term Animal Bioassays: Lifetable vs Summary Incidence Data. GOLD, L. S., BERNSTEIN, L., KALDOR, J., BACKMAN, G., AND HOEL, D. (1986). *Fundam. Appl. Toxicol.* **6**, 263–269. Two methods for estimating carcinogenic potency from animal carcinogenesis bioassays (TD50-defined in the paper) are compared, one based on lifetable data and one based on summary incidence data. The lifetable analysis adjusts for the differential effects of toxicity among dose groups and for differences in the time pattern of tumor incidence, while summary incidence analysis does not. However, summary data are all that are usually available in the published results of animal cancer tests. Using NCI bioassay results which provide full lifetable data, we compare lifetable and summary estimates of potency and their statistical significance as well as the estimated shape of the dose–response curve. There is substantial agreement between these methods of analysis in terms of potency estimation, although lifetable estimates are usually more potent. But, there are some notable differences in the estimated shape of the dose–response curve, suggesting that both target site selection and method of analysis play an important role in risk estimation.

The major source of information for assessing the carcinogenicity of the increasing number of chemicals in the environment is the chronic exposure animal bioassay. Evidence of carcinogenicity is obtained from experiments in which a test compound is administered to groups of laboratory animals over a long time period, and the pattern of tumors in exposed (treated) animals is compared with that of nonexposed (control) animals.

Statistical analyses of bioassay results are usually based on summary incidence data. The percentage of animals developing a tumor of interest is calculated for each treatment and control group, and the relationship between these fractions and the administered dose is examined. Most animal bioassay data are published only in the form of summary incidences.

Recently, interest has focused on the importance of using time-to-tumor or lifetable data in the analysis of animal carcinogenicity studies in order to adjust for the differential effects of toxicity among dose groups and for differences in the time pattern of tumor incidence (Peto *et al.*, 1984). If the dose level administered to the animals is toxic, then premature death from nonneoplastic causes may prevent some animals that would have developed tumors from developing them. Summary incidence data will, in such cases, indicate a smaller proportion of tumor-bearing animals than would be obtained with actuarial adjustments, and hence may result in an underestimate of the carcinogenicity of the test agent.

Statistical methods based on lifetable data

[1] To whom correspondence should be addressed.

0272-0590/86 $3.00

are, however, far more complex than those based on summary incidence data, both from a conceptual and a computational viewpoint. The purpose of this paper is to compare the empirical estimates of carcinogenic potency and dose–response curve shape obtained from the two methods of statistical analysis. We have used the database of bioassays conducted and published by the NCI Bioassay Program prior to July 1980, for the comparison; this is the largest single source of animal bioassays providing full lifetable data.

NCI BIOASSAYS

In the "standard NCI bioassay" as described in Sontag *et al.* (1976), the test agent is administered to both sexes of mice and rats for most of the lifetime. For each sex–species studied, there are three groups of 50 animals each: a control group (vehicle where appropriate), a group administered the "maximum tolerated dose" (MTD), and a group administered one-half the MTD. The MTD is defined as the maximum level of exposure which is not expected to shorten the natural lifetime from nonneoplastic causes, and which is expected to result in no more than a 10% weight decrement in animals receiving this dose when compared to controls.

Within a bioassay each test of one sex in one species is considered an experiment. The actual conduct of the NCI bioassays published prior to July 1980 varies from one experiment to another, and details of each experiment are given in Gold *et al.* (1984).

We consider here bioassays of 185 chemicals conducted in both sexes of rats and mice, and bioassays of four chemicals conducted only in male and female rats. Among the 776 experiments, 85% of those in mice and 80% of those in rats had 50 or more animals in each dose group; however, 60% of the control groups in each species had no more than 20 animals (the majority of these had 20), and only one-third had 50 or more. In nearly all tests, two dose levels were used; 31 experiments had an additional dose level.

The duration of dosing and the length of experiment to terminal sacrifice also varied widely. Most exposures were long term, with more than 88% of both species exposed for at least 18 months; however, in only 15% of the experiments were animals dosed for a full 2 years. Overall, the experiments in mice were shorter than those in rats: only one-third of the mouse tests lasted 2 years compared to 85% of the rat tests. The median length of mouse experiments was 94 weeks.

STATISTICAL ANALYSES FOR EACH EXPERIMENT

The lifetable methods which we have used to analyze the experimental data have been described elsewhere (Sawyer *et al.*, 1984). Briefly, a proportional hazards model (Cox, 1972) is assumed for the time-to-tumor data, in which $\lambda(t, d)$, the tumor-hazard rate at age t for a specific site, is linearly related to d, the administered dose rate of test chemical in milligrams per kilogram body weight per day, as

$$\lambda(t, d) = (1 + \beta d)\lambda_0(t). \qquad (1)$$

$\lambda_0(t)$ is the tumor-incidence rate at zero dose. The parameter β and the function λ_0 are estimated using maximum likelihood methods. The likelihood ratio statistic tests the hypothesis that the chemical has no carcinogenic effect (i.e., $\beta = 0$), and a χ^2 goodness-of-fit statistic tests the validity of the linear relationship between dose and tumor incidence expressed by Eq. (1). In fitting the model, no attempt is made to distinguish between tumors found in a fatal context and tumors found in an incidental context. Thus the time-to-tumor occurrence is taken to be the time to death of the animal, whether death results from the tumor of interest, or from some other cause, including terminal sacrifice (Sawyer *et al.*, 1984; Peto *et al.*, 1984).

For summary incidence data, we fit by maximum likelihood methods the comparable model

$$p_d = 1 - \exp\{-(a + bd)\}, \qquad (2)$$

where $a > 0$ and $b > 0$ and p_d is the probability that an animal exposed at dose d for its lifetime develops a tumor. This model is linear at low doses and is often referred to as the "one-hit model." Here, the number of animals developing tumors at dose d is assumed to follow a binomial distribution with parameters n_d and p_d, where n_d is the number of animals initially exposed at dose d. As with lifetable data, the likelihood ratio statistic is used to test whether the compound is carcinogenic, i.e., whether $b = 0$, and a χ^2 statistic tests the adequacy of the model.

From either type of analysis, we estimate carcinogenic potency as TD50: the dose rate in milligrams per kilogram body weight per day which would halve the probability of an animal remaining tumor free by the end of the standard lifespan for the species (Peto *et al.*, 1984). In other words, the TD50 is that daily dose which will induce tumors in half of the animals that would have remained tumor free at zero dose. One advantage of the TD50 is that the experimental dose range will often include it, which makes for statistically accurate estimation. Another advantage is that it takes the spontaneous tumor rate into account. The estimate of TD50 based on summary incidence data is simply $\log 2/b$, where b is the maximum likelihood estimate (MLE) of b. For lifetable data, the estimate is a more complex function of the MLEs of β and $\lambda_0(t)$ (Sawyer *et al.*, 1984). In our database we have estimated 99% confidence intervals for the TD50s calculated from lifetable data and for those based on summary incidence data. The method for calculating these intervals from lifetable data is described in Sawyer *et al.* (1984). For summary incidence data, 99% likelihood–ratio test-based confidence limits are obtained for b and are then transformed to limits for TD50.

We have applied both the summary and lifetable analyses to every NCI experiment in the Carcinogenic Potency Database. The results of the lifetable analyses are reported in full by Gold *et al.* (1984). In the summary calculations for NCI experiments, we have based the proportion of tumor-bearing animals on the number of animals *started* in the dose group, rather than an effective number (such as number at risk at the time of first tumor), thus maximizing the difference between the two methods of analysis.[2]

For either method of estimating TD50, if the χ^2 goodness-of-fit test indicated statistically significant departure from linearity ($p < 0.05$) and this departure was downward, the analysis was repeated eliminating the highest dose group. The purpose of this procedure was to remove the effects of toxicity in summary incidence analyses and to remove the effects of dose saturation in the lifetable analyses. If the goodness-of-fit test indicated an upward departure from linearity, no groups were eliminated when fitting the model.

For each experiment, the most potent target site (based on the lowest value of TD50) was determined separately for the lifetable and summary incidence analyses. When the p value associated with the test for carcinogenicity was less than 0.01, target sites within an experiment were ordered according to the magnitude of TD50, and the smallest TD50 was defined as the most potent site. If no p values were less than 0.01, this process was repeated using all sites with $p < 0.1$. (See Gold *et al.*, 1984 for details of the selection of the most potent site.) Our comparisons between the lifetable and summary analyses are based on the most potent TD50 from each experiment because this is a conservative procedure from a human risk assessment standpoint and also provides one TD50 per method of analysis

[2] Throughout the Carcinogenic Potency Database, when estimating summary TD50 values for experiments in the general literature, we have based the proportion of tumor-bearing animals on the number alive at the appearance of the first tumor in the experiment or on the number examined histologically, whenever these are reported (see Gold *et al.*, 1984). Such a reduced denominator does make allowance for the effects of premature deaths on the numbers of animals that develop tumors. Use of starting number for the NCI experiments in this comparison of lifetable and summary incidence methods is intended to maximize the difference between the two methods.

TABLE 1

COMPARISON OF LIFETABLE AND SUMMARY INCIDENCE ANALYSES BY p-VALUE ASSOCIATED WITH THE TEST FOR CARCINOGENICITY IN THE MOST POTENT SITE FOR 776 EXPERIMENTS

Lifetable p value	Summary incidence p value		
	$p < 0.01$	$0.01 \leq p < 0.05$	$p \geq 0.05$
$p < 0.01$	229	14	8
$0.01 \leq p < 0.05$	25	46	15
$p \geq 0.05$	9	22	408

to summarize each experiment. For the majority of lifetable-summary comparisons (92%) the target organ(s) and histopathology for the most potent site were found to be the same.

CARCINOGENIC POTENCY: TD50

Table 1 compares the statistical significance of the most potent TD50 estimated by lifetable data and summary incidence methods. As noted above, this value indicates the significance associated with testing whether the slope of the dose–response curve is different from zero, and is therefore a measure of the carcinogenicity of the compound. For the vast majority of experiments, the two analyses produce similar levels of significance. Using the 0.01 level of significance, 720/776 experiments (93%) produce concordant results. This high degree of consistency is important because summary data are all that are available for most experiments.

We compared the estimates of TD50 from the lifetable and summary analyses for the experiments with a statistically significant carcinogenic effect ($p < 0.01$) in the lifetable analysis. Differences between the methods are maximized by including in these comparisons those cases (1) where the p value for the summary TD50 was greater than 0.01, and (2) where, for only one method, the TD50 was calculated after eliminating the highest dose group. A histogram of the ratios of the TD50 from the most potent lifetable site to the TD50 from the equivalent summary incidence site is shown in Fig. 1. The lower the ratio, the more potent the lifetable estimate of TD50 compared to the summary estimate.

As expected, the lifetable TD50 is nearly always more potent than the summary TD50 (i.e., ratio < 1.0); however, the overall differences are not large. For about half the cases the effect of using lifetable data is to reduce the TD50 (to increase potency) by less than 30%. The median ratio is 0.72, and 90% of the ratios lie between 0.30 and 1.30. In only 3.5% of the cases did the TD50s differ by an order of magnitude (ratio < 0.10). In five of these nine cases, no TD50 could be calculated by the summary incidence method.

As expected, the magnitude of the ratio of the lifetable to summary TD50 was related to

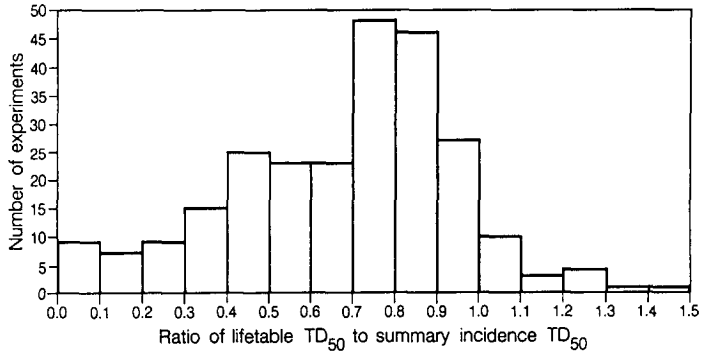

FIG. 1. Frequency distribution of ratio of lifetable TD50 to summary incidence TD50 in most potent sites for statistically significant experiments (lifetable $p < 0.01$).

the occurrence of chemical toxicity in an experiment. For 15/25 cases where the lifetable estimate is considerably more potent than the summary incidence estimate (lifetable-summary ratio < 0.30), the NCI Technical Report noted survival problems among dosed animals; this compares to only 39/226 for the remaining cases.

We found substantially similar distributions of the ratio of lifetable to summary incidence TD50 for a variety of subsets of the NCI database: for each sex of rat and mouse examined separately, as well as for the group of experiments which were evaluated in the text of the Technical Reports as showing evidence of carcinogenicity. We also found substantially similar distributions of this ratio when we used in the analysis the least potent TD50 ($p < 0.01$) for those experiments with more than one statistically significant target site, and for the group of TD50s estimated after elimination of the high-dose group in either or both methods of analysis.

We also compared the range of TD50 values encompassed by 99% confidence intervals estimated by lifetable and summary incidence analyses. The intervals overlap in all cases where the ratio is greater than 0.53, and do not overlap in 21/25 cases where the lifetable summary ratio is less than 0.30.

Because estimates of TD50 for the NCI experiments range over approximately seven orders of magnitude, we would expect little difference in the ranking of chemicals by TD50 if either summary incidence or lifetable TD50s were used. We calculated the Spearman rank correlation coefficient between the two measures for each sex of rat and mouse: all correlations were greater than 0.87.

DOSE–RESPONSE SHAPE

We classified the shape of the dose–response curve for each method of analysis using results from the respective model goodness-of-fit tests to determine whether the two statistical approaches were consistent in fitting the data. The lifetable goodness-of-fit test compares observed tumor counts with expected tumor counts, from the fitted linear model [Eq. (1)], which have been adjusted for differential effects of toxicity among dose groups and the time pattern of tumor occurrence. Expected counts in the summary incidence goodness-of-fit test make no such adjustment. Thus, the two methods could result in conflicting classifications. For example, if the lifetable adjustment for toxicity increases the ratio of observed to expected tumors in the high-dose group, lifetable analysis could classify the data as curving significantly upward ($p < 0.05$) whereas the summary analysis might not.

To compare dose–response curve shapes, we have again considered all experiments with a statistically significant carcinogenic effect ($p < 0.01$) based on the lifetable calculation, using the most potent site to summarize each experiment. The curve shape for the summary incidence method is based on the same target site(s) as the curve shape for the lifetable method. The analysis has been restricted to experiments in which two nonzero doses were tested, since only in this case can dose–response curves which fail the goodness-of-fit test be classified as to a significant upward or downward departure from linearity.

Table 2 presents the nine possible combinations of lifetable and summary incidence curve shapes. For two-thirds of the experiments lifetable and summary methods agree on the shape of the dose–response curve. As expected, among curves estimated to be nonlinear and curving downward in summary incidence analysis, a majority are either linear or curving upward in lifetable analysis. Lifetable analysis classifies 53% of the curves as linear, 36% as nonlinear consistent with upward curvature, and 11% as nonlinear consistent with downward curvature. Since there is often more than one statistically significant target site within an experiment, we repeated the curve shape analysis in experiments with two nonzero doses using (1) all statistically significant (lifetable $p < 0.01$) TD50s in the database for NCI experiments ($N = 850$), and

TABLE 2

COMPARISON OF LIFETABLE AND SUMMARY INCIDENCE ANALYSES BY DOSE–RESPONSE CURVE SHAPE IN 216 EXPERIMENTS WITH LIFETABLE $p < 0.01$ AND TWO NONZERO DOSE GROUPS

	Summary incidence curve shape			
Lifetable curve shape	Not consistent with linear, upward curvature	Consistent with linear	Not consistent with linear, downward curvature	Total
Not consistent with linear, upward curvature	26	46	7	79
Consistent with linear	1	101	12	114
Not consistent with linear, downward curvature	0	7	16	23

then (2) only those sites evaluated in the NCI Technical Reports as providing evidence for carcinogenicity of the compound ($N = 308$). The results in both cases are substantially similar to those shown in Table 2.

These results on the estimation of curve shape represent a difference between the two methods of analysis. There is disagreement in the curves in 76 experiments. For 53 of these, the curve in the lifetable analysis was upward while in the summary incidence analysis it was either linear or downward. The more frequent classification of the most potent site as upward curving when using lifetable methods does not mean, however, that other statistically significant target sites within an experiment are not consistent with linearity. We subsequently examined all of the statistically significant target sites within each of these 53 experiments. Thirty-two experiments have more than one statistically significant target site, and 24 of these have a different target site with a linear dose–response estimated by lifetable methods. Of the 29 experiments which do not have a linear dose–response in the same experiment, 18 have a statistically significant site with a linear curve in another sex-species group tested with the same compound. Thus for 42 of the 53 cases, there is a significant lifetable curve which is consistent with linearity in at least one sex-species group administered the test compound.

DISCUSSION

We have compared two methods of estimating potency from animal cancer bioassay results and found that determination of carcinogenic effect and estimation of an index of carcinogenic potency by the two methods produced very similar results. The lifetable method usually produced a more potent estimate. The two methods differ, however, in classifying the shape of the dose–response curve.

These analyses represent two extremes in terms of utilization of animal carcinogenicity data. The summary incidence analysis ignores the time at which tumors are found and is based on the starting number of animals, whereas lifetable analysis, which incorporates time-to-tumor information, is based on actuarially adjusted proportions of animals at risk. Both adjust for tumor incidence in control animals. Several alternative approaches for analysis could have been considered. These include the use of "effective" number of animals at risk at the time of first death with tumor, rather than the number initially exposed for summary incidence analysis; or the fitting

of parametric survival distributions, such as the Weibull, to the time-to-tumor (lifetable) data. However, the comparison presented here between the models specified in Eqs. (1) and (2) should be expected to maximize differences in results if such differences exist.

The similarity in the TD50 estimates by the two methods of analysis suggests that summary incidence data can be used to estimate carcinogenic potency. Summary estimates can be improved if experimental results are published for the number of animals with the tumor(s) of interest as a proportion of the number alive at the time of the first tumor in the experiment, rather than as a proportion of the number initially exposed. Such information adjusts for early mortality, and removes from the potency calculation those animals which were not alive and at risk of tumor at the time of tumor occurrence.

Estimates of the shape of the dose–response curve by summary and lifetable methods are consistent for two-thirds of the most potent sites in statistically significant experiments. While more dose–response curves are classified as curving upward by lifetable methods, statistically significant sites with linear curves are also usually found within the same experiment or in other experiments of the same test agent. It is not clear what overall effect utilization of "effective" number of animals would have on estimates of the shape of the dose–response curve by summary methods.

To summarize, we have determined that there is substantial agreement between lifetable and summary incidence methods of analysis in terms of the statistical significance of TD50 and its estimated value. Since summary incidence data are usually all that is available for most experiments, we can view the potency analyses based on such data with a high degree of confidence. However, it is often the case that we reject linearity of the dose response for a given target site with one method of analysis and not the other. Similarly, we have observed that the shape of the dose–response curve may differ for different target sites in experiments with the same test agent. This suggests that estimates of risk at lower doses, which incorporate information on the shape of the dose–response curve, may differ depending on the method of analysis and target site(s) used.

ACKNOWLEDGMENTS

This work was supported by NIEHS/DOE Interagency Agreement 222-Y01-AS-10066, EPA-NCI/DOE Interagency Agreement Y01-CP-15791, and Grant SIG-2 from the American Cancer Society. We want to thank Dr. Ken Brown for his helpful comments on earlier drafts.

REFERENCES

COX, D. R. (1972). Regression models and lifetables. *J. R. Stat. Soc. Brit.* **32,** 187–220.

GOLD, L. S., SAWYER, C. B., MAGAW, R., BACKMAN, G. M., DE VECIANA, M., LEVINSON, R., HOOPER, N. K., HAVENDER, W. R., BERNSTEIN, L., PETO, R., PIKE, M., AND AMES, B. N. (1984). A carcinogenic potency database of the standardized results of animal bioassays. *Environ. Health Perspect.* **58,** 9–319.

PETO, R., PIKE, M. C., BERNSTEIN, L., GOLD, L. S., AND AMES, B. N. (1984). The TD50: A proposed general convention for the numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments. *Environ. Health Perspect.* **58,** 1–8.

SAWYER, C., PETO, R., BERNSTEIN, L., AND PIKE, M. C. (1984). Calculation of carcinogenic potency from long-term animal carcinogenesis experiments. *Biometrics* **40,** 27–40.

SONTAG, J. A., PAGE, N. P., AND SAFFIOTI, U. (1976). Guidelines for carcinogen bioassay in small rodents. *Carcinog. Tech. Rep. Ser. 1—USNCI,* DHEW Pub. No. (NIH) 76–801.