Environmental Health Perspectives
Vol. 58, pp. 385–392, 1984

Statistical Issues in the Design, Analysis and Interpretation of Animal Carcinogenicity Studies

by Joseph K. Haseman*

Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies are discussed. In the area of experimental design, issues that must be considered include randomization of animals, sample size considerations, dose selection and allocation of animals to experimental groups, and control of potentially confounding factors.

In the analysis of tumor incidence data, survival differences among groups should be taken into account. It is important to try to distinguish between tumors that contribute to the death of the animal and "incidental" tumors discovered at autopsy in an animal dying of an unrelated cause. Life table analyses (appropriate for lethal tumors) and incidental tumor tests (appropriate for nonfatal tumors) are described, and the utilization of these procedures by the National Toxicology Program is discussed. Despite the fact that past interpretations of carcinogenicity data have tended to focus on pairwise comparisons in general and high-dose effects in particular, the importance of trend tests should not be overlooked, since these procedures are more sensitive than pairwise comparisons to the detection of carcinogenic effects.

No rigid statistical "decision rule" should be employed in the interpretation of carcinogenicity data. Although the statistical significance of an observed tumor increase is perhaps the single most important piece of evidence used in the evaluation process, a number of biological factors must also be taken into account. The use of historical control data, the false-positive issue and the interpretation of negative trends are also discussed.

Introduction

Although the overall evaluation of the results of a carcinogenicity study in laboratory animals is a complex process involving scientific judgment with all its frailties, the process is made less difficult if the experiment has been properly designed, the data appropriately analyzed and certain interpretative issues adequately addressed.

The purpose of this paper is to consider in detail each of these important areas. Statistical design issues include proper randomization of animals, sample size considerations, dose selection and animal allocation issues, and the control of potentially confounding factors such as littermate and caging effects. Data analyses should employ methodology that takes survival differences into account and ideally makes use of cause of death information. Issues of interpretation include the use of historical controls, awareness of the false-positive issue and the consideration of negative trends.

*Biometry and Risk Assessment Program, National Institute of Environmental Health Sciences, P. O. Box 12233, Research Triangle Park, NC 27709.

Experimental Design

If a study has been inadequately designed, no amount of careful histopathological evaluation or elegant statistical analysis can salvage the experiment. Many design issues are "nonstatistical" and will not be dealt with here. They include species and strain of animal employed in the experiment, route of administration, diet, study duration, caging, intercurrent infectious diseases and chemical stability. If one or more of these factors is compromised, the interpretability of the study becomes more difficult. The "statistical" design issues are discussed below.

Randomization

Animals should be randomly assigned to treated or control groups to insure that there is no selection bias in the formulation of these groups. In some instances, a stratified random sampling scheme may be appropriate: for example, where animals are first stratified by body weight and then assigned at random to treated or control groups. In any case, some formal randomization scheme should be employed rather than relying on

Table 1. Underlying tumor incidence in the high dose group that can be detected with 50%, 70% and 90% power by using Fisher's exact test with 50 animals per group.

Spontaneous tumor rate, %	Underlying tumor incidence, %					
	$p < 0.05$ test			$p < 0.01$ test		
	50%	70%	90%	50%	70%	90%
0.1	9.5[a]	11.8	15.8	13.5	16.2	20.5
1.0	11.0	13.8	18.4	15.1	18.2	23.4
3.0	14.0	17.4	22.9	18.9	22.8	29.0
5.0	17.0	20.8	27.0	22.5	26.8	33.3
10.0	24.2	28.8	35.7	30.2	34.9	41.9
20.0	36.8	41.7	49.0	43.2	48.4	56.0
30.0	48.1	53.6	61.1	54.8	59.9	67.0

[a]Tumor incidence; for example, if the spontaneous tumor rate is 0.1%, a one-sided Fisher's exact test comparing control and high dose groups of 50 animals each would have a 50% chance of detecting an underlying tumor incidence of 9.5% in the high dose group. Exact power calculations for Fisher's exact test were obtained by the method described by Haseman (*53*).

subjective judgment for assigning animals to treated and control groups.

Sample Size Considerations

A sufficient number of animals should be included in the experimental design to insure reasonable power for detecting carcinogenic effects. The "standard" NCI design employs 50 animals in each of three groups: control, low dose and high dose. Table 1 shows the approximate power of this particular design for detecting carcinogenic effects. The actual number of animals per group will depend upon the specific study objectives, but one should always consider the power of the design before the experiment is begun. If interim kills are to be employed, allowance for these extra animals must be made in the overall experimental design.

Dose Selection and Allocation of Animals

Perhaps no single issue related to the design of laboratory animal carcinogenicity studies has generated as much discussion as has dose selection. Since small numbers of rodents are serving as surrogates for a large human population, and since even the most carefully documented human carcinogens (e.g., cigarette smoking) do not produce tumors in the majority of subjects at risk, there is general acceptance of the basic principle that animal testing must be carried out at doses that exceed a typical human exposure (*1–5*). Large doses must be employed to insure reasonable power for detecting carcinogenic effects. However, there continues to be debate regarding the actual magnitude of doses that should be employed in these investigations.

Dose selection for the chronic study is generally based on the results of a series of subchronic (single-dose, two-week, and 90-day) toxicity studies. Data from these experiments that are factored into the dose selection process include weight gain and survival information, pharmacokinetic and metabolism data, and the results of a thorough histopathological examination.

Sontag et al. (*6*) recommend that the highest dose employed should be the "maximum tolerated dose" (MTD) which they define as "The highest dose of the test agent during the chronic study that can be predicted not to alter the animals' normal longevity from effects other than carcinogenicity." This definition formed the basis for dose selection in the NCI carcinogenicity studies and continues to be used by the NTP as well.

Table 2. Recommended designs for carcinogenicity studies.[a]

Group	Dose	Number of animals
Control	0	50
Low	20–30% MTD	30–40
Middle	50% MTD	60–70
High	MTD	50

[a]From Portier and Hoel (*13*) assuming a total sample size of 200 and a four dose design. MTD = Maximum tolerated dose.

Some investigators object to the use of maximum tolerated doses in carcinogenicity testing. They argue that these doses are often "too high" and may produce tumors due to metabolic overloading of the body's natural detoxification mechanisms or may result in carcinogenic effects "secondary" to recurrent cytotoxicity and tissue damage. Several recent papers (*7,8*) present a comprehensive discussion of these and other important issues involved in the selection of doses for carcinogenicity studies.

Certain disagreements regarding the use of high doses in carcinogenicity testing seem more related to the basic definition and estimation of the maximum tolerated dose rather than to major philosophical differences. Clearly, doses that produce excessive mortality (apart from that related to chemically induced carcinogenicity) are undesirable. On the other hand, the high dose must elicit some signs of toxicity or some biological effect (*2,3,7–11*), or the animals may not have been sufficiently challenged by the test chemical. The NTP has recently incorporated an additional lower dosed group into its study design and this group provides for a margin of safety should the high dose be overly toxic. NTP also has increased its efforts to obtain pharmacokinetic and metabolism data for the test chemical that might be factored into the dose selection and study evaluation processes.

Traditionally, equal allocations of animals have been employed in dosed and control groups. However, in some instances unequal allocation may be preferable. Portier and Hoel (*12,13*) investigated various allocations of animals to derive an "optimal design," where optimality was defined in terms of the power of the design, the variability of low-dose risk estimates and the accuracy of estimating the dose-response curve in the experimental region. Their final recommended design is given in Table 2 and involves unequal allocation of animals to dosed and control groups.

Confounding Factors

If certain factors can be identified that would otherwise confound the interpretation of the study, then the experimental design should be modified to take these factors into account. One example is cage location, which in several studies (*14,15*) has been shown to be related to increased incidences of cataracts and retinopathy (because of the proximity of the animals to the fluorescent light source). This particular problem can be dealt with by a systematic rotation of cages. In addition to cage location, the housing of more than one animal per cage may introduce "cage effects" that require consideration in the subsequent statistical analysis (*16*).

Littermates may also be a potentially confounding factor. Littermate information is generally not available when the animals are received from the supplier and it is implicitly assumed (reasonably in most cases) that any "litter effects" have been "balanced out" by the random assignment of animals to treated or control groups. Specialized methods of analysis have been proposed (*17,18*) if it is desired to employ a formal litter-matched design.

A more subtle potential source of bias is a difference in slide preparation of tissues for histopathological examination. Ideally, identical procedures should be employed for dosed and control groups with regard to the number of slides prepared and the methodology employed. Occasionally, additional slides may be made for a particular animal because of suspicious "lumps and bumps" noted at gross necropsy, but the investigator must be careful that these additional slides do not introduce bias into the diagnostic process. For example, if for a particular organ consistently more sections are examined for dosed animals than for controls, one would expect to find more dosed-group tumors by chance alone.

A related issue is "blind" pathology, which has traditionally been the source of debate between statisticians and pathologists. Most statisticians recommend "blind" pathology (i.e., histologic examination without knowing the source of the tissue) to insure the total objectivity of the tumor diagnosis. Otherwise, subjective bias could effect the overall interpretation of the results.

On the other hand, most pathologists feel that the disadvantages of "blind" pathology outweigh the advantages (*19*). They assert that control animals must be examined first to determine the naturally occurring incidence and severity of neoplastic and nonneoplastic lesions. They further contend that blind pathology will require additional time and effort and will introduce the possibility of coding errors, and that an experienced pathologist can generally distinguish between tissues from control and treated animals in any event.

Perhaps a reasonable compromise (currently employed by the NTP) is to have the original pathologist diagnose lesions in a nonblind fashion. Then, if apparent treatment-related effects are found, these particular tissues can then be reviewed blindly to determine whether or not these lesions can be verified under a more rigorous protocol. If this procedure reveals evidence of bias on the part of the original pathologist, this raises serious questions regarding the objectivity of all his pathology diagnoses, including those which showed no apparent treatment-related effects.

Methods of Statistical Analysis

If a study has been properly designed and executed, the statistical analysis of the data is obviously facilitated. The methods of analyses traditionally used by the NCI in the evaluation of tumor incidence data were Fisher's exact test for pairwise comparisons and the Cochran-Armitage test for dose-response trends (*20*). These procedures compare directly the proportion of tumor-bearing animals in dosed and control groups.

In recent years there has been an increased awareness of the need to take intercurrent mortality into account in the analysis of tumor incidence data. Consequently, a number of different methods have been proposed that deal with time-to-death-with-tumor (*21–28*).

One simple approach to this problem (used, e.g., in many of the early NCI studies) is to exclude all animals that died prior to the appearance of the first tumor, and then carry out the analyses described above (Fisher's exact test; Cochran-Armitage test). In some cases this approach will be satisfactory, but in other instances more rigorous methods of analysis may be required.

Perhaps the most comprehensive discussion of this issue is given by Peto et al. (*25*). These authors emphasize the need to determine the "context of observation" of each tumor, i.e., to determine whether the tumor contributed directly or indirectly to the cause of death or whether alternatively the tumor was merely an incidental finding at autopsy in an animal dying of an unrelated cause. The distinction between "fatal" and "incidental" tumors is important, because in the evaluation of tumor incidence it is essential to distinguish between a chemical that reduces survival because of shortened tumor latency (a real carcinogenic effect) and one which also reduces survival, but in which tumors are merely being observed earlier because animals are dying of some other competing risk (a noncarcinogenic effect). Hoel and Walburg (*21*) present an example with actual experimental animal data to show how a mislead-

ing result can be obtained if a "fatal tumors" analysis is incorrectly applied to incidental tumor data. Mantel et al. (*29*) refer to this distinction as whether or not tumors are "self-evidencing," but the basic principle is the same.

The primary difficulty with survival-adjusted analyses is the need to determine the context of observation of each tumor on an individual animal basis. Peto et al. (*25*) present an example with over 4500 tumors in which 94% of all tumors could be classified as either "definitely incidental" or "definitely fatal," notwithstanding the initial reservations of pathologists regarding whether such determinations could reliably be made. Despite these results, other pathologists remain skeptical regarding the general accuracy of cause of death determinations. As part of its new modified pathology protocols, the National Toxicology Program requests that pathologists attempt to determine contexts of observation for each tumor observed in NTP studies. It is hoped that ultimately sufficient data can be generated to evaluate the feasibility of utilizing cause-of-death information on an individual animal basis in the evaluation of NTP carcinogenicity data.

Most pathologists agree that it is often possible to judge *a priori* the likelihood that a particular tumor type will be fatal or incidental. For example, one simple rule (admittedly unacceptable for general usage) is that malignant tumors are generally fatal and benign tumors are usually incidental. For those tumors determined to be virtually always "fatal" or "incidental," the context of observation for individual animals becomes less important.

For fatal tumors, life table methods (*23*,*30*) can be employed to evaluate tumor incidence data. With this approach, the proportions of tumor-bearing animals in dosed and control groups are compared at each point in time when an animal dies with the tumor of interest. The denominators of these proportions are the total number of animals at risk in each group. These results are then combined by Mantel-Haenszel methods (*31*). These methods can also be used to pool the "fatal tumor" analysis with the corresponding comparisons based on the "incidental tumors" observed at the end of the experiment to obtain an overall p value. Life table methods are sensitive both to an increased tumor incidence and to a shortened latency period.

For nonfatal tumors, the procedure described by Peto et al. (*25*), which is essentially the Hoel-Walburg (*21*) method, can be employed. According to this approach, the proportions of animals found to have tumors in dosed and control groups are compared at selected time intervals. The denominators of these proportions are the number of animals actually autopsied during the time interval. The individual time interval comparisons are then combined by Mantel-Haenszel methods to obtain a single overall result. An example illustrating the numerical computations for these various methods of analysis is given in the Appendix.

One disadvantage of the incidental tumor test is that a subjective determination of time intervals is required for purposes of grouping the data. A further drawback is that for studies in which reduced survival in the dosed group is severe, the method will have little power because only those tumors for which there is overlapping survival in dosed and control groups will be included in the statistical analysis. An alternative method of analysis for incidental tumors currently being studied by the NTP is based on logistic regression (*28*). In addition to avoiding the disadvantages cited above, the logistic regression approach allows the investigator to take into account certain covariables (e.g., cage location, litter effects) that might otherwise be confounding factors in the overall evaluation of the data.

Most statisticians (*20*,*25*) seem to believe that one-tailed tests are generally preferable to two-tailed tests in the evaluation of tumor incidence data. The reason for this preference is that the primary objective of these studies is to identify carcinogens (a one-tailed alternative). In any case, it should be clearly stated which is being used when p values are reported.

Despite the fact that past interpretations of carcinogenicity data have tended to focus on pairwise comparisons in general and high-dose effects in particular (*32*), the importance of trend tests should not be overlooked. Indeed, Peto et al. (*25*) state that trend tests combine "in a reasonably optimal way (and should therefore usually supercede)" the information obtained from pairwise comparisons. Since trend tests utilize information from all experimental groups simultaneously, they are more sensitive than pairwise comparisons to the detection of carcinogenic effects, which is an advantage that becomes more important as the number of dosed groups increases. Ironically, this very sensitivity seems to have limited the subjective value placed on these tests in the past by some investigators, namely because of the fear of making "false positive" decisions (see discussion of this issue below). Increasing emphasis will likely be given to trend tests (1) as standard designs begin utilizing additional dosed groups, (2) as more knowledge is gained regarding the likelihood of false positive results, and (3) as more information is learned about the patterns of chemically induced dose-response trends.

Interpretation of the Data

Even if a study has been carefully designed and appropriate statistical methodology employed, interpretation of results is a complex process. The following factors merit special consideration: use of historical control data, multiple comparisons issues, and interpretation of increased and decreased tumor incidence.

Historical Control Data

Over the past several years, NCI and NTP have accumulated considerable information on background tumor rates in mice and rats, particularly the $B6C3F_1$ mouse and the Fischer 344 rat. The concurrent control group is always the most important control group used

in the decision making process. However, there are some instances in which the use of historical control data can aid an investigator in the evaluation of tumor incidence data. Two examples are rare tumors and tumors that show a marginal increase relative to concurrent controls (*33*).

Before historical control data can be used in a meaningful way, however, there are a number of problems that must first be addressed. For example, nomenclature differences must be resolved. That is, nomenclature conventions and diagnostic criteria should be identical for all studies in the historical control database. Then, a decision has to be made regarding which studies should be included in the database. Next, the sources of variability in historical control tumor rates must be identified, and if possible, controlled. There are certain tumors that show considerable (extrabinomial) study-to-study variability. What factors are responsible for this variability?

NTP has identified two major sources of variability in historical control tumor rates: calendar year and laboratory. That is, tumor rates do seem to change over time, and limiting the data base to more recent studies helps control this source of variability. Secondly, laboratory-to-laboratory variability seems to be quite large for certain tumors. Thus, NTP currently gives primary emphasis to laboratory-specific tumor rates.

Finally, if historical control data are to be used in a formal testing framework, statistical procedures that adjust for extrabinomial variability should be employed. Three procedures that have been proposed for taking extrabinomial variability into account (*34–36*) are currently being studied by the NTP. A recent publication (*37*) considers the historical control issue in detail.

Multiple Comparison Considerations

Another important interpretative issue in carcinogenicity testing is how to take multiple comparisons into account. Since each NTP study consists of four separate experiments, with approximately 30 tissues examined per animal and a battery of statistical tests employed, the potential exists for finding false positives, i.e., statistically "significant" differences that are merely due to chance variation alone. When considering this issue it should be kept in mind that the false positive "problem" is limited primarily to common tumors, since for rare tumors it is virtually impossible for a sufficient number to occur in any one study by chance alone to lead to a statistically significant result.

One possible strategy to deal with this problem (employed, e.g., in the early NCI bioassays) is to use a Bonferonni-type multiple comparisons adjustment. A discussion of Bonferonni adjustments has been given by Mantel (*16*).

Because interpretation of carcinogenicity data is a complex process, other investigators believe that it is not necessary to employ any formal rigid multiple-comparisons adjustment. As noted by Peto et al. (*25*), "*P*-values are objective facts, but unless a *p*-value is very extreme, the proper use of it in the light of other information to decide whether or not the test agent really is carcinogenic involves subjective judgement."

Thus, no rigid "decision rule" should be employed in the interpretation of carcinogenicity data. Although the statistical significance of an observed tumor increase is perhaps the single most important piece of evidence used in the evaluation process, a number of other factors must be taken into account as well: (1) whether the effect was dose-related, (2) whether the effect was supported by related nonneoplastic changes and/or by similar evidence in other sex–species groups, (3) whether the effect occurred in a target organ, (4) relative survival of dosed and control animals, (5) the historical rate of the particular tumor, and (6) the biological "meaningfulness" of the effect.

Are false positives a major problem in NCI/NTP carcinogenicity studies? Certain investigations of this issue (which assumed that any $p < 0.05$ effect was automatically regarded as biologically meaningful) estimated these rates to be quite high: 20 to 50% for a single sex-species group (*38*). Other investigators questioned the validity of this particular decision rule and emphasized (as discussed above) that the interpretation of carcinogenicity results incorporates biological knowledge and corroborative evidence such as the presence of a dose–response relationship or experimentally consistent results in different species or sexes (*39*). These investigators and other researchers (*20,40*) concluded that the overall false-positive rate generally does not greatly exceed the nominal level.

In a recent examination of 25 NTP feeding studies, a simple statistical rule was derived which appears to mimic closely the scientific judgment process used in these experiments. This "rule" was as follows: regard as carcinogenic any chemical that produces a high-dose increase in a common tumor that is statistically significant at the 0.01 level or a high-dose increase in an uncommon tumor that is statistically significant at the 0.05 level. It should be re-emphasized that no rigid decision procedure was (or should be) used in these studies. However, the overall false-positive rate associated with this particular decision rule (which appears to closely approximate the overall evaluation process) was estimated and found to be no more than 7 to 8% (*41*). Thus, false positives do not appear to be a major problem in NCI/NTP studies.

Increased and Decreased Tumor Incidences

One noteworthy characteristic of NCI/NTP carcinogenicity studies is the tendency of these experiments to show significantly decreased as well as increased tumor incidences. In fact, an examination of 25 recent NTP feeding studies revealed that the frequencies of statistically significant ($p < 0.01$) increases and decreases in tumor incidence in these studies were approximately

the same (*42*). Some investigators have responded to these types of findings by taking the position that such studies are uninterpretable, should be abandoned altogether, and a "drastically different design" employed (*43*).

Other authors feel that a more prudent and scientific course of action is to investigate the patterns of tumor increases and decreases and attempt to identify possible causal factors, as well as to better characterize the biologic process or processes responsible for these opposing responses (*44*). Survival differences can account for many of the decreased tumor incidences observed in dosed groups, since if the chemical's toxic effect results in early mortality, these animals may not have survived sufficiently long to be considered at risk for developing tumors.

Another explanatory factor is decreased weight gain. Frequently, the chemical under test reduces body weight gain, and decreased incidences of certain endocrine and reproductive system tumors (notably mammary gland fibroadenoma in female F344 rats) are clearly related to decreased weight gain in the dosed groups (*42*). This result agrees with similar observations made in earlier studies (*45–49*).

Haseman (*42*) also found an inverse correlation between the incidences of liver tumors and leukemia in F344 rats. A similar negative association between the incidences of lymphomas and liver tumors in CF-1 mice exposed to DDT was reported by Breslow et al. (*50*). This result was confirmed in a later study (*51*) that involved extensive serial sacrifice and thus overcame doubts about whether the original finding may have been an artifact of the rapid lethality of lymphomas. The point is that while one should not ignore negative trends in tumor incidence, a closer examination of the data may in many cases reveal possible associations that may serve as explanatory factors. Also, one should not rule out the possibility that in some instances certain chemicals may in fact exhibit "true" anti-carcinogenic effects.

Appendix

In this section we give a detailed comparison of the three methods of analysis used most often in the analysis of tumor incidence data: the fatal tumors (life-table) analysis, Peto's incidental tumors analysis and the Cochran-Armitage linear trend test.

The discussion that follows deals with only the trend tests but similar comments apply for the pairwise comparisons. All three test procedures follow the same basic approach: (1) break down the data by time interval; (2) calculate expected number of tumors in each time interval; (3) calculate the corresponding variance term V; (4) combine the data over all time intervals, compare observed and expected tumor incidence, and calculate the final test statistic.

The primary differences between the procedures are as follows.

The Cochran-Armitage test in essence uses a single time interval (the entire study period) and hence does not adjust for intercurrent mortality. In situations in which there is a marked treatment effect on survival, this can be a serious shortcoming of this approach.

The fatal tumors analysis assumes that all tumors of a given type in animals dying prior to terminal sacrifice (TS) are "fatal," (i.e., either directly or indirectly caused the death of the animal). The incidental tumors analysis assumes that all such tumors are "incidental," (i.e., were merely observed at autopsy in animals dying of an unknown or unrelated cause). At present NTP protocols do not require specification of the cause of death for each individual animal, so it is not possible to determine precisely which tumors are fatal and which are incidental. The Cochran-Armitage test does not require an assumption regarding tumor "lethality."

For a particular time interval the fatal tumors analysis is based on all animals alive at the beginning of the interval. In contrast, the incidental tumors analysis is based only on the animals that die and are autopsied during the interval.

For the fatal tumors analysis, each week at which a tumor is observed is regarded as a separate "time interval." For the incidental tumors analysis, there is a certain flexibility with regard to choice of time intervals, and NTP generally uses (for a 2-year study) five time intervals (expressed as weeks): 0-52, 53-78, 79-92, 93-TS and TS.

An example is now presented which illustrates the three methods. The data are taken from the NTP 11-Aminoundecanoic Acid Technical Report (*52*), and the lesions are malignant lymphoma in male mice.

Let O denote the number of observed tumors; E, the number of expected tumors; N the number of animals at risk; and D, dose levels [without loss of generality (for equally spaced doses) taken as 0,1,2]. Σ denotes summation, and R are the ΣO for a given time interval and M are ΣN for a given time interval.

Then, for each time interval, for the three dose groups i, we have

$$E_i = N_i R/M \qquad i = 1,2,3$$

and

$$V = \frac{(M - R)R[M\Sigma ND^2 - (\Sigma ND)^2]}{M^2 (M - 1)}$$

[The Cochran-Armitage test uses M^3 rather than $M^2(M - 1)$ in the denominator of V].

The individual values of O, E and V are then summed over all time periods, and the final test statistic is based on the total and can be calculated as

$$Z = \Sigma D(O - E)/(V)^{0.5}$$

The p-value that corresponds to this test statistic can be found from tables of the Standard Normal

Table A-1.

Time interval, weeks	Controls			Low dose			High dose			
	O	*E*	*N*	*O*	*E*	*N*	*O*	*E*	*N*	*V*
0–109	2	5	50	9	5	50	4	5	50	9

Table A-2.

Time interval, weeks	Controls			Low dose			High dose			
	O	*E*	*N*	*O*	*E*	*N*	*O*	*E*	*N*	*V*
73	0	0.405	47	0	0.371	43	1	0.224	26	0.5965
85	1	0.415	44	0	0.387	41	0	0.198	21	0.5661
88	0	0.404	42	0	0.394	41	1	0.202	21	0.5650
93	0	0.414	41	1	0.384	38	0	0.202	20	0.5712
100	0	0.427	41	1	0.375	36	0	0.198	19	0.5725
106	0	0.407	37	0	0.385	35	1	0.209	19	0.5763
107	0	0.411	37	1	0.389	35	0	0.200	18	0.5665
109	1	3.326	37	6	3.056	34	1	1.618	18	4.2150
Total	2	6.209		9	5.740		4	3.051		8.2291

Table A-3.

Time interval, weeks	Controls			Low dose			High dose			
	O	*E*	*N*	*O*	*E*	*N*	*O*	*E*	*N*	*V*
0–52	0	0	2	0	0	3	0	0	22	0.0000
53–78	0	0.154	2	0	0.462	6	1	0.385	5	0.4852
79–92	1	0.909	5	0	0.545	3	1	0.545	3	1.2496
93–108	0	1.600	4	3	1.600	4	1	0.800	2	1.4933
109	1	3.326	37	6	3.056	34	1	1.618	18	4.2150
Total	2	5.989		9	5.663		4	3.348		7.4431

Distribution. For the Cochran-Armitage linear trend test, we have, for the values shown in Table A-1,

$$Z = [(9 - 5) + 2(4 - 5)]/(9)^{0.5} = 0.67,\ p = 0.252$$

The Cochran-Armitage test currently employed by the NTP uses a continuity correction, i.e.,

$$Z = [|(9 - 5) + 2(4 - 5)| - 0.5]/(9)^{0.5} = 0.50,\ p = 0.309$$

For the fatal tumor (life table) analysis we have, for the data of Table A-2,

$$Z = [(9 - 5.740) + 2(4 - 3.051)]/(8.2291)^{0.5} = 1.798,\ p = 0.036$$

or, if a continuity correction is employed,

$$Z = [|(9 - 5.74) + 2(4 - 3.051)| - 0.5]/(8.2291)^{0.5} = 1.624,\ p = 0.052$$

For the Peto incidental tumor analysis we have, for the data of Table A-3,

$$Z = [(9 - 5.663) + 2(4 - 3.348)]/(7.4431)^{0.5} = 1.701,\ p = 0.044$$

or, if a continuity correction is employed,

$$Z = [|(9 - 5.663) + 2(4 - 3.348)| - 0.5]/(7.4431)^{0.5} = 1.518,\ p = 0.065$$

For these data, adjusting for intercurrent mortality revealed some evidence of a dose-related trend that would have been missed had the usual Cochran-Armitage test been carried out. This latter test does not take into account the fact that at the high dose 44% of the animals died during the first year (compared with 4% of the controls) and the first malignant lymphoma was not observed until week 73.

I would like to thank Drs. Michael Hogan and James Huff for their helpful suggestions and Ms. Kay Moore for typing this manuscript.

REFERENCES

1. National Academy of Sciences. Drinking Water and Health. National Research Council, National Academy of Sciences Washington, DC, 1977.
2. International Agency for Research on Cancer. Long-term and Short-term Screening Assays for Carcinogens: A Critical Appraisal. IARC Monographs, Supplement 2, Lyon, 1980.
3. Occupational Safety and Health Administration. Identification, classification and regulation of potential occupational carcinogens. Fed. Reg. 45(15): 5001–5296 (1980).
4. Food Safety Council. Proposed System for Food Safety Assessment. Food Safety Council, Washington, DC, 1980.
5. Office of Technology Assessment. Assessment of Technologies for Determining Cancer Risks from the Environment. Congress of the United States, Office of Technology Assessment, Washington, DC 1981.
6. Sontag, J. M., Page, N. P., and Safiotti, U. Guidelines for

Carcinogen Bioassay in Small Rodents. DHHS Publication (NIH) 76-801, National Cancer Institute, Bethesda, MD, 1976.
7. International Life Sciences Institute. The selection of doses in chronic toxicity/carcinogenicity studies. In: Current Issues in Toxicology (H. C. Grice, Ed.), Springer Verlag, New York, 1984, pp. 9–49
8. Haseman, J. K. Dose selection issues in carcinogenicity testing. Fund. Appl. Toxicol., in press.
9. Food Safety Council. Chronic toxicity testing. Proposed system for food safety assessment. Food. Cosmet. Toxicol. (Suppl. 2) 16: 97–108 (1978).
10. Organisation for Economic Co-operation and Development. OECD Guidelines for Testing of Chemicals, Paris, France, 1981.
11. Environmental Protection Agency. Health Effects. Test Guidelines, Office of Pesticides and Toxic Substances, Washington, DC, 1982.
12. Portier, C., and Hoel, D. G. Optimal design of the chronic animal bioassay. J. Toxicol. Environ. Health 12: 1–19 (1983).
13. Portier, C., and Hoel, D. G., Design of the chronic animal bioassay for goodness-of-fit to multistage models. Fundamental Appl. Toxicol. in press.
14. National Toxicology Program. NTP Technical Report on the Carcinogenesis Bioassay of Propyl Gallate, NTP TR 240, Dept. of Health and Human Services, 1982.
15. National Toxicology Program. NTP Technical Report on the Carcinogenesis Bioassay of Ziram, NTP TR 238, Dept. of Health and Human Services, 1983.
16. Mantel, N. Assessing laboratory evidence for neoplastic activity. Biometrics 36: 381–399 (1980).
17. Mantel, N., Bohidar, N. R., and Ciminera, J. L. Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. Cancer Res. 37: 3863–3868 (1977).
18. Mantel, N., and Ciminera, J. L. Use of logrank scores in the analysis of litter-matched data on time to tumor appearance. Cancer Res. 39: 4308–4315 (1979).
19. Weinberger, M. A. How valuable is blind evaluation in histopathologic examinations in conjunction with animal toxicity studies. Toxicol. Pathol. 7: 14–17 (1980).
20. Gart, J. J., Chu, K. C., and Tarone, R. E. Statistical issues in interpretation of chronic bioassay tests for carcinogenicity. J. Natl. Cancer Inst. 62: 957–974 (1979).
21. Hoel, D. G., and Walburg, H. E. Statistical analysis of survival experiments. J. Natl. Cancer Inst. 49: 361–372 (1972).
22. Peto, R. Guidelines on the analysis of tumor rates and death rates in experimental animals. Brit. J. Cancer 29: 101–105 (1974).
23. Tarone, R. E. Tests for trend in life table analysis. Biometrika 62: 679–682 (1975).
24. Turnbull, B. W., and Mitchell, T. J. Exploratory analysis of disease prevalence data from survival/sacrifice experiments. Biometrics 34: 555–570 (1978).
25. Peto, R., Pike, M., Day, N, Gray, R., Lee, P., Parish, S., Peto, J., Richard, S., and Wahrendorf, J. Guidelines for simple, sensitive, significant tests for carcinogenic effects in long-term animal experiments. International Agency for Research Against Cancer. Monographs: Long-Term and Short-Term Screening Assays for Carcinogens: A Critical Appraisal. World Health Organization Geneva, 1980, Supplement 2, pp. 311–426.
26. Kodell, R. L., Shaw, G. W., and Johnson, A. M. Nonparametric joint estimates for disease resistance and survival functions in survival/sacrifice experiments. Biometrics 38: 43–58 (1982).
27. Kodell, R. L., Farmer, J. H., Gaylor, D. W. and Cameron, A. M. Influence of cause-of-death assignment on time-to-tumor analyses in animal carcinogenesis studies. J. Natl. Cancer Inst. 69: 659–664 (1982).
28. Dinse, G. E., and Lagakos, S. W. Regression analysis of tumor prevalence data. J. Roy. Statist. Soc. Ser. C 32: 236–248 (1983).
29. Mantel, N., Tukey, J. W., Ciminera, J. L., and Heyse, J. F. Tumorigenicity assays, including use of the jackknife. Biomet. J. 24: 579–596 (1982).
30. Cox, D. R. Regression models and life tables (with discussion). J. Roy. Statist. Soc. B34: 187–220 (1972).
31. Mantel, N., and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22: 719–748 (1959).
32. Chu, K. C., Cueto, C., and Ward, J. M. Factors in the evaluation of 200 National Cancer Institute carcinogen bioassays. J. Toxicol. Environ. Health 8: 251–280 (1981).
33. Haseman, J. K. Statistical support of the proposed National Toxicology Program Protocol. Toxicol. Pathol. 11: 77–82 (1983).
34. Hoel, D. G. Conditional two-sample tests with historical controls. Contributions to Statistics: Essays in Honor of Norman L. Johnson (P. K. Sen, Ed.), North-Holland Publishing Co., Amsterdam, 1983, pp. 229–236.
35. Dempster, A. P., Selwyn, M. D., and Weeks, B. J. Combining historical and randomized controls for assessing trends in proportions. J. Am. Statist. Assoc. 78: 221–227 (1983).
36. Tarone, R. E. The use of historical control information in testing for a trend in proportions. Biometrics 38: 215–220 (1982).
37. Haseman, J. K., Huff, J., and Boorman, G. A. Use of historical control data in carcinogenicity studies in rodents. Toxicol. Pathol., in press.
38. Salsburg, D. S. Use of statistics when examining lifetime studies in rodents to detect carcinogenicity. J. Toxicol. Environ. Health 3: 611–628 (1977).
39. Fears, T. R., Tarone, R. E., and Chu, K. C. False-positive and false-negative rates for carcinogenicity screens. Cancer Res. 37: 1941–1945 (1977).
40. Haseman, J. K. Response to "Use of statistics when examining Lifetime studies in rodents to detect carcinogenicity." J. Toxicol. Environ. Health 3: 633–636 (1977).
41. Haseman, J. K. A re-examination of false-positive rates for carcinogenicity bioassays. Fund. Appl. Toxicol. 3: 334–339 (1983).
42. Haseman, J. K. Patterns of tumor incidence in two-year cancer bioassay feeding studies in Fisher 344 rats. Fund. Appl. Toxicol. 3: 1–9 (1983).
43. Salsburg, D. S. The lifetime feeding study in mice and rats—an examination of its validity as a bioassay for human carcinogens. Fund. Appl. Toxicol. 3: 63–67 (1983).
44. Haseman, J. K., Huff, J. E., and Moore, J. A. Response to "The lifetime feeding study in mice and rats—an examination of its validity as a bioassay for human carcinogens." Fund. Appl. Toxicol. 3: 3/a–5/a (1983).
45. Roe, F. J. C., and Tucker, M. J. Recent developments in the design of carcinogenicity tests on laboratory animals. Proc. Eur. Soc. Study Drug Toxicity 15: 171–177 (1978).
46. Ross, M. H., and Bras, G. Lasting influence of early caloric restriction on prevalence of neoplasms in the rat. J. Natl. Cancer Inst. 47: 1095–1113 (1971).
47. Rowlatt, C., Franks, L. M., and Sheriff, M. U. Mammary tumor and hematoma supression in dietary restriction in C3H A^{av} mice. Brit. J. Cancer 28: 83 (1973).
48. Tucker, M. J. The effect of long-term food restriction on tumors in rodents. Int. J. Cancer 23: 803–807 (1979).
49. Conybeare, G. The effect of quality and quantity of diet on survival and tumor incidence in outbred Swiss mice. Food Cosmet. Toxicol. 18: 65–75 (1980).
50. Breslow, N. E., Day, N. E., Tucusov, V. S., and Tomatis, L. Associations between tumor types in a large-scale carcinogenesis study of CF-1 mice. J. Natl. Cancer Inst. 52: 233–239 (1974).
51. Wahrendorf, J. Simultaneous analysis of different tumor types in a long-term carcinogenicity study with scheduled sacrifices. J. Natl. Cancer Inst. 70: 915–921 (1983).
52. National Toxicology Program. NTP Technical Report on the Carcinogenesis Bioassay of 11-Aminoundecanoic Acid NTP TR 216, Dept. of Health and Human Services, 1982.
53. Haseman, J. K. Exact sample sizes for use with the Fisher-Irwin test for 2 × 2 tables. Biometrics 34: 106–109 (1978).