Environmental Health Perspectives
Vol. 82, pp. 337–344, 1989

Dual Controls, *p*-Value Plots, and the Multiple Testing Issue in Carcinogenicity Studies

by Murray R. Selwyn*

The interpretation of statistically significant findings in a carcinogenicity study is difficult, in part because of the large number of statistical tests conducted. Some scientists who believe that the false positive rates in these experiments are unreasonably large often suggest that the use of multiple control groups will provide important insight into the operational false positive rates.

The purpose of this paper is 2-fold: to present results from two carcinogenicity studies with dual control groups, and to present and illustrate a new graphical technique potentially useful in the analysis and interpretation of tumor data from carcinogenicity studies. The experimental data analyzed show that statistically significant differences between identically treated groups will occur with regular frequency. Such data, however, do not provide strong evidence of extrabinomial variation in tumor rates.

The *p*-value plot is advocated as a graphical method that can be used to assess visually the ensemble of *p* values for neoplasm data from an entire study. This technique is then illustrated using several examples. Through computer simulation, we present *p*-value plots generated with and without treatment effects present. On average, the plots look substantially different depending on the presence or absence of an effect. We also evaluate decision rules motivated by the *p*-value plots. Such rules appear to have good power to detect treatment effects (i.e., have low false negative rates) while still controlling false positive rates.

Introduction

One of the most difficult issues associated with the interpretation of results from a carcinogenicity study in experimental animals is the question of biological versus statistical significance. Suppose, for example, that a statistically significant increasing trend in tumor rates is detected for one or two sites among a large number of such sites examined microscopically. Can we conclude that these are real effects or are the findings simply a chance event?

This issue relates directly to the question of false positive rates in these studies, an issue that has been discussed and debated in the statistical and toxicological literature for decades. The basic argument that the overall or experiment-wise error rate may be unreasonably high (*1*) follows from a simple probability calculation. If M independent statistical tests are conducted, each at the $p = 0.05$ level, then the probability of at least one significant finding is $1 - 0.95^M$. For $M = 10$ this overall probability is 0.40 and for $M = 40$, the probability is 0.87. In a carcinogenicity study, where 30 to 40 tissue or organ sites may be examined for both males and females, there is the potential for a high false positive rate (the multiplicity problem).

The actual false positive rate in a carcinogenicity study may not be nearly as high as indicated by the above probability calculation, however. As several authors (*2–4*) have pointed out, most scientific decisions are not based on a single statistically significant result at the 0.05 level, and further, as noted by Haseman (*4*), many statistical tests, in fact, operate below their nominal levels. By studying historical tumor rates from 25 studies conducted by the National Toxicology Program (NTP), Haseman (*4*) concluded that a statistical decision rule that approximated the NTP biological decision process could be formulated as follows: Declare a positive finding if the p value comparing the high dose to controls is less than 0.01 for common tumors (greater than 1% historical spontaneous rate) or if the p value is less than 0.05 for uncommon tumors. (Haseman's calculations of false positive rates used Fisher's exact test comparing the high dose to controls. Sample sizes were 50 per group.)

The applicability of Haseman's results to a wider scope of situations depends on how typical the NTP data are. In addition, one of the critical assumptions (which is generally made when analyzing unadjusted or lifetime tumor incidence data) is that the tumor counts are binomi-

*Statistics Unlimited, Inc., 2000 Commonwealth Ave., Auburndale, MA 02166.

ally distributed. In a typical carcinogenicity bioassay design with a single control group and treated groups at several exposure levels of a chemical [e.g., IARC (*5*)], the binomial assumption cannot be verified. In contrast, studies with replicate groups can be used to assess the binomial assumption. Haseman et al. (*6*) present results from a set of 18 color additive studies, each employing a dual control group design. In brief, they found no evidence of extrabinomial, within-study variability in these studies. They again reaffirmed, however, the idea that more stringent evidence than a single $p < 0.05$ for common tumors should be required for biological significance; otherwise, the experiment-wise false positive rate could be unacceptably high.

Haseman et al. are understandably cautious in generalizing their results. Therefore, there is continued interest in the results of studies with dual controls. The purpose of the present paper is 2-fold: to present findings from two studies designed with dual control groups [although the data from these studies are on a much smaller scale than the data analyzed by Haseman et al. (*4*)], and to present and illustrate a new graphical technique potentially useful in the analysis and interpretation of tumor data from rodent carcinogenicity studies. Our concern throughout relates to the false positive rate in these studies, applicable statistical procedures, and the implications of decision rules on statistical power (or equivalently the false negative rate). Our goal is to consider statistical approaches that will be helpful to other scientists in their biological interpretation of tumor data from these studies.

Studies with Dual Control Groups

The rationale for designing carcinogenicity studies with dual control groups is to provide a between-groups comparison that is not confounded with potential treatment effects. Any two groups that differ in terms of experimental conditions would not provide such a comparison. For such groups, differences in response could potentially be due to the different experimental conditions. As the two control groups are treated identically throughout the study, any differences in responses must simply be due to chance. If they differ more frequently than would be explainable according to the standard binomial model, then this presents some evidence that the standard model may not hold. In such cases, statistical procedures need to be modified to take into account this extrabinomial variation.

In this section, we present and analyze a subset of the results from two studies, one in rats and one in mice, with dual control groups. The first study was conducted in CD-1 mice with two vehicle control groups (control group 1 and control group 2), and three other groups fed dose levels of 25, 100, and 400 mg/kg/day of the chemical (chemical 1) in the diet for 25 months. At the start of the experiment, there were 100 mice in each of the five treatment groups for each sex. We concentrate here on the comparison between the two control groups in the experiment.

Table 1 presents the results of statistical analyses with Fisher's exact test comparing the tumor rates in the two control groups. Notice that among the 29 tumor types/sites analyzed, three comparisons are significant at the 0.05 level. When group 1 is viewed as the treated group, it has a significantly higher rate of reticulum cell sarcomas in males ($p = 0.032$). When group 2 is considered as a treated group, two comparisons are significant: lymphosarcomas in females ($p = 0.032$) and total blood vessel tumors in males ($p = 0.008$). Do these results present strong evidence of extrabinomial variation?

To address this question, we performed calculations of false positive rates in the same manner as did Haseman et al. (*6*) in their analysis of the 18 color additive studies. Basically, these authors calculated two types of false positive rates: conditional and unconditional. Conditional rates are calculated assuming that the total number of tumors in the two groups is fixed. For example, if one observes 5/100 in group 1 and 7/100 in group 2, then the total number of tumors in both groups is 12. We can then calculate the probability of all statistically significant outcomes using Fisher's exact test at the 0.05 level conditionally, given a total of 12 tumors. The unconditional method simply uses the two proportions to estimate the common tumor rate, which would be $12/200 = 0.06$ for the example. Given this as the spontaneous rate in each group, we can again calculate the probability of all statistically significant outcomes.

Table 2 presents estimated conditional and unconditional false positive rates for each of the 29 tumor types considered in Table 1. Notice that, as observed by Haseman (*4*), tumor types with low rates have negligible false positive rates, and therefore contribute minimally to the overall (experiment-wise) false positive rate. The conditional rates calculated with group 1 as the treated group and with group 2 as the treated group are not identical because of occasional differences in denominators. Sometimes these differences in false positive rates are appreciable (Table 2). Such differences are due to the discrete nature of the counts used in the test and the fact that even with denominators of almost 100, the probability distributions take large jumps. Moreover, there are sometimes considerable differences between false positive rates calculated conditionally and unconditionally (Table 2), again because of the discrete nature of the data. The unconditional rates increase smoothly with the background tumor rates observed here.

Using the unconditional false positive rates and assuming independence across the 29 sites (as did Haseman), we calculate the following probabilities:

Prob [no significant results at 0.05 level] = 0.676.
Prob [one significant result at 0.05 level] = 0.269.
Prob [two significant results at 0.05 level] = 0.050.
Prob [three or more significant results at 0.05 level] = 0.006.

Thus the overall false positive rate is $1 - 0.676 = 0.324$, and the chance of getting at least two positive results are $0.050 + 0.006 = 0.056$. Even though we have observed two significant results comparing group 2 (as treated) to group 1 as control, the probability of this occurring is

Table 1. Comparison of the dual control groups in the mouse study with chemical 1.

Tumor type	Control group 1	Control group 2	*p* values from Fisher's exact test	
			Group 1 as treated	Group 2 as treated
Females				
Hepatocellular carcinoma	0/99	0/98	1.000	1.000
Total hepatocellular tumors	5/99	6/98	0.737	0.493
Uterus: adenocarcinoma	3/98	4/98	0.778	0.500
Uterus: leiomyoma	2/98	3/98	0.816	0.500
Uterus: leiomyosarcoma	1/98	1/98	0.751	0.751
Uterus: granular cell tumor	1/98	0/98	0.500	1.000
Uterus: sarcoma	0/98	1/98	1.000	0.500
Uterus: squamous cell carcinoma	0/98	0/98	1.000	1.000
Vagina: squamous cell carcinoma	1/91	0/90	0.503	1.000
Ovary: granulosa cell tumor	1/96	0/89	0.519	1.000
Ovary: luteoma	0/96	0/89	1.000	1.000
Ovary: papillary cystadenoma	3/96	4/89	0.808	0.458
Hemangiosarcoma: all sites	6/100	5/100	0.731	0.500
Total blood vessel tumors	7/100	13/100	0.952	0.119
Lymphosarcoma	17/100	29/100	0.986	0.032
Granulocytic leukemia	1/100	0/100	0.500	1.000
Mammary gland: adenocarcinoma	1/100	2/100	0.877	0.500
Stomach: adenoma	0/96	1/96	1.000	0.500
Nose: odontoma	1/83	0/84	0.497	1.000
Males				
Hepatocellular carcinoma	10/99	10/100	0.584	0.602
Total hepatocellular tumors	19/99	19/100	0.558	0.585
Lung: carcinoma	9/100	9/100	0.597	0.597
Total lung tumors	13/100	16/100	0.789	0.344
Hemangiosarcoma: all sites	3/100	8/100	0.971	0.107
Total blood vessel tumors	3/100	13/100	0.999	0.008
Reticulum cell sarcoma	7/100	1/100	0.032	0.997
Stomach: papilloma	1/98	0/99	0.498	1.000
Nose: odontoma	3/80	4/83	0.763	0.521
Testis: gonadal stromal tumor	1/100	3/100	0.939	0.311

0.056, assuming binomial variation. Hence, these data do not exhibit strong evidence of extrabinomial variation. When group 1 is viewed as treated, the probability calculation for one or more positives is 0.324. In the second study with chemical 1 in Sprague-Dawley rats, none of 14 comparisons between the dual controls resulted in a statistically significant ($p < 0.05$) difference in either direction using Fisher's exact test. Thus, evidence of extrabinomial variation is truly lacking in these two studies.

For the mouse study, when all tests are performed at the 0.05 level, the overall error rate of 32% is unacceptably high. This finding is consistent with that of Haseman (*4*), who concludes that the overall error rate is too high if all tests are conducted at the 0.05 level. Haseman considers testing at the 0.01 level for common tumors and at the 0.05 level for rare tumors. But assuming that the observed spontaneous rate for total blood vessel tumors in male mice (16/200 = 0.08) is unbiased, the observed p value of 0.008 would lead us to conclude that the group 2 effect is tumorigenic using $p < 0.01$.

Thus, even though these data sets do not provide strong evidence of extrabinomial variation, they do reinforce the idea that false positives continue to be a substantial problem in carcinogenicity studies.

p-Value Plot as a Diagnostic Tool

As noted by a number of authors (*7,8*), considerations other than p values alone bear upon the question of carcinogenicity in a particular instance. As stated in the 1980 IARC monograph (*7*):

> "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 judgment."

Haseman (*4*) argues that additional factors to be considered should include the historical control tumor rate for the tumor in question, the survival histories of the control and treated groups, dose-dependence and similarity of findings among different sexes and species, and biological plausibility in light of earlier toxicological studies, mutagenicity findings, etc.

A large part of the role of the statistician is to provide objective means for interpreting data and results to assist in making subjective judgments. In this regard, we highlight three areas in which statistical techniques may be most useful.

The first is the formal application of statistical methods incorporating historical controls into the analysis of tumor data. Several methods are currently available (*9,10*). These methods have been found to be quite informative when tumor rates are low and the potential dose-response is low enough to be uncertain.

The second area involves the evaluation of the results in one sex (e.g., males) while taking into account the results in the other sex (females). Thus, one could treat

Table 2. Estimated false positive rates when comparing the two control groups in the mouse study with chemical 1.[a]

Tumor type	Unconditional false positive rate[b]	Conditional false positive rates	
		Group 1 as treated	Group 2 as treated
Females			
Hepatocellular carcinoma	0.000	0.000	0.000
Total hepatocellular tumors	0.022	0.030	0.028
Uterus: adenocarcinoma	0.017	0.007	0.007
Uterus: leiomyoma	0.012	0.030	0.030
Uterus: leiomyosarcoma	0.001	0.000	0.000
Uterus: granular cell tumor	0.000	0.000	0.000
Uterus: sarcoma	0.000	0.000	0.000
Uterus: squamous cell carcinoma	0.000	0.000	0.000
Vagina: squamous cell carcinoma	0.000	0.000	0.000
Ovary: granulosa cell tumor	0.000	0.000	0.000
Ovary: luteoma	0.000	0.000	0.000
Ovary: papillary cystadenoma	0.017	0.009	0.048
Hemangiosarcoma: all sites	0.022	0.029	0.029
Total blood vessel tumors	0.031	0.049	0.049
Lymphosarcoma	0.036	0.032	0.032
Granulocytic leukemia	0.000	0.000	0.000
Mammary gland: adenocarcinoma	0.004	0.000	0.000
Stomach: adenoma	0.000	0.000	0.000
Nose: odontoma	0.000	0.000	0.000
Males			
Hepatocellular carcinoma	0.031	0.046	0.017
Total hepatocellular tumors	0.035	0.048	0.025
Lung: carcinoma	0.030	0.041	0.041
Total lung tumors	0.034	0.022	0.022
Hemagiosarcoma: all sites	0.022	0.029	0.029
Total blood vessel tumors	0.028	0.033	0.033
Reticulum cell sarcoma	0.018	0.032	0.032
Stomach: papilloma	0.000	0.000	0.000
Nose: odontoma	0.019	0.006	0.008
Testis: gonadal stromal tumor	0.008	0.000	0.000

[a]All tests at the 0.05 level.
[b]Common tumor rate estimated from pooled incidence. Assumes 100 animals per group.

responses of (0/50, 3/50, 6/50, 8/50) in males differently depending on whether one has observed (8/50, 6/50, 4/50, 7/50) or (1/50, 3/30, 5/50, 6/50) in females. This idea is currently being pursued in terms of research on appropriate statistical methods.

The third idea relates to techniques to graphically display and evaluate the ensemble of p values for neoplasm data from an entire study. Our methods are similar in spirit, but less elaborate than those recently proposed by Meng and Dempster (*11*). In this section, we discuss the use of p-value plots as an informal graphical method for assessing the overall carcinogenicity of a chemical.

We assume that appropriate statistical analyses have already been conducted and that p values associated with potential treatment effects are available from a number of tumor types/sites. Given this set of p values, our approach proceeds as follows:

- Instead of working with the p values (p's), it is more convenient to work with the 1-p's. Thus a p value of 0.01 has a corresponding 1-p value of 0.99. The rationale for working with the 1-p's instead of p's is that data analysts (statisticians, scientists) may be more comfortable investigating interesting large values (e.g., outliers, right-skewed distributions) rather than small ones.
- The 1-p's are ordered from smallest to largest. Large values of 1-p, of course, correspond to small p values. Assuming that the p values are independent (not quite true because of the dependence between tumor types) and each is uniformly distributed on the interval [0,1] (also not quite true because of the discreteness of some tests), then the ordered 1-p's each follow a beta distribution. The expectations and percentiles of this approximating distribution can be obtained. The expectation of the ith largest value is $i/(n + 1)$. We obtained percentiles for each ordered 1-p using the BETAINV function in SASR. All values are plotted versus equally spaced scores. In practice, it is probably easiest to use $i/(n + 1)$. Then the expected line has unit slope.
- Each observed 1-p is plotted along with its expectation and percentiles versus equally spaced scores. Despite the fact that the assumptions above do not strictly hold, this will give an informal indication about the p values jointly.
- Of particular importance are the most extreme (highest) 1-p values. For example, are any outside the 2.5 to 97.5% envelope? Do the highest ones all lie above their expected values?

To illustrate this idea, we show in Figure 1 the p values from Fisher's exact test of Table 1 using control group 2 as treated. Notice that there are a large number of p values identically equal to 1 corresponding to the many 1-p values of 0 displayed in the lower left portion of the graph. The observed 1-p's generally lie below their expectations, indicating the conservative nature of Fisher's exact test. Also, one can see the step-function form of the 1-p curve illustrating the discreteness of the data.

To adjust for these factors, we eliminate p values equal to 1. (In practice, one would also eliminates sites with only one tumor because of the discreteness of p-value distribution.) In Figure 2, we display the resulting 1-p values with expectations and percentiles recomputed accordingly for the reduced set. Note that the test does not appear as conservative as in Figure 1, as many of the 1-p values are now above their expectations. Of particular interest are the low p values. We can compare the three lowest values to their expectations and to the 2.5% point (envelope), as shown in Table 3.

As can also be seen from Figure 2, the two lowest p values are below their expected values, but are considerably above the 2.5% values from their respective distributions. (Recall that we plot 1-p's rather than p's, so that the relevant part of Figure 2 is in the upper-right corner).

Figure 3 presents p values from trend tests [either the Cochran-Armitage test (*12,13*) or an exact trend test of Bickis and Krewski (*14*)] comparing the treated groups in the mouse study with chemical 1 to the pooled control group (pool of control group 1 and control group 2). p values of 1 have been eliminated. Observe that the 1-p values from these tests fluctuate around the expected line and lie totally within the (2.5 to 97.5%) envelope, thus indicating general conformance with the assumptions of the plotting procedure described. As the smallest p value from these trend tests is 0.09, there is no evidence of increasing tumor rates as a result of feeding the chemical to mice.

Our final example is from a chronic feeding study in mice with chemical 2. Figure 4 displays the observed 1-p

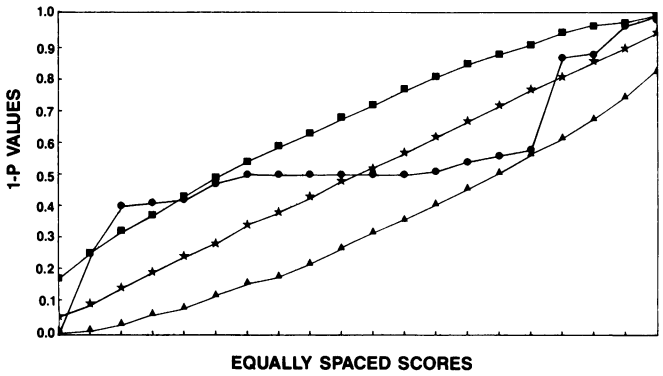

FIGURE 2. Chemical 1 mouse study: p values (Fisher's exact test, group 2 as treated, p values of 1 eliminated). Observed 1-p (●), expected values (*), 2.5% envelope (▲), and 97.5% envelope (■) versus equally spaced scores [$i/(n + 1)$].

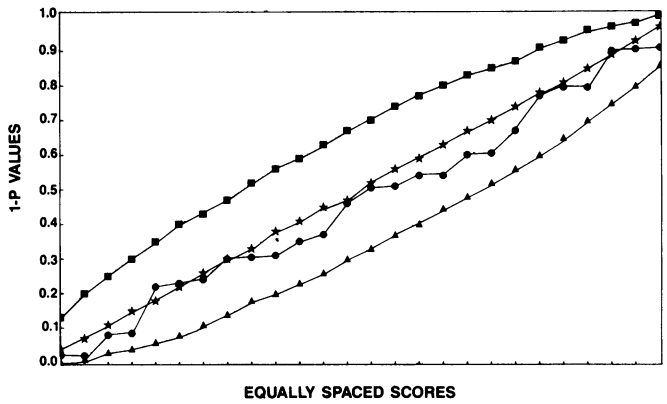

FIGURE 3. Chemical 1 mouse study: p values from trend tests (p values of 1 eliminated). Observed 1-p (●), expected values (*), 2.5% envelope (▲), and 97.5% envelope (■) versus equally spaced scores [$i/(n + 1)$].

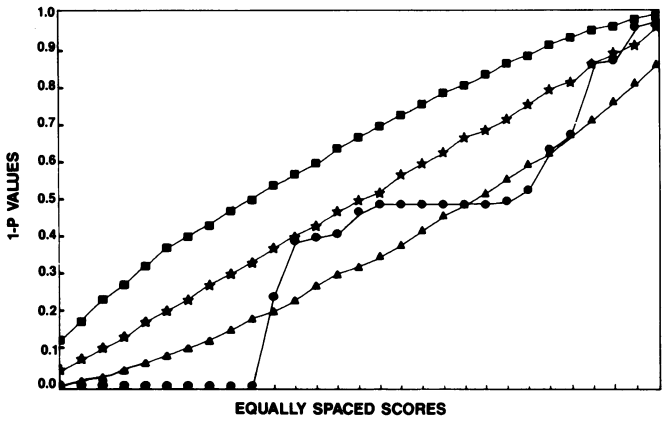

FIGURE 1. Chemical 1 mouse study: p values (Fisher's exact test, group 2 as treated). Observed 1-p (●), expected values (*), 2.5% envelope (▲), and 97.5% envelope (■) versus equally spaced scores [$i/(n + 1)$].

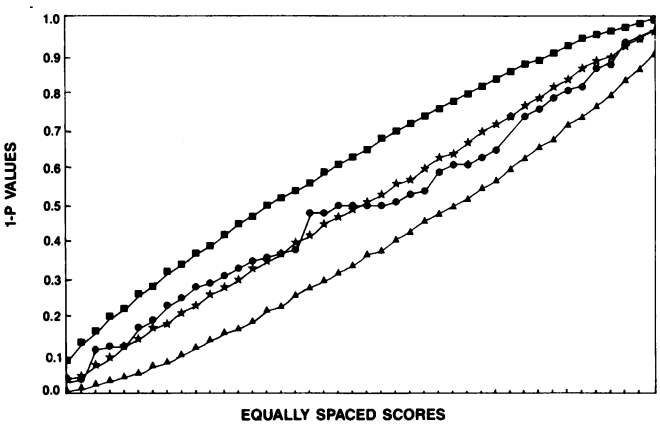

FIGURE 4. Chemical 2 mouse study: p values from trend tests (p values of 1 eliminated). Observed 1-p (●), expected values (*), 2.5% envelope (▲), and 97.5% envelope (■) versus equally spaced scores [$i/(n + 1)$].

Table 3.

Observed p value	Expected p value	Lower 2.5 percentile
0.008	0.048	0.001
0.032	0.095	0.012
0.107	0.143	0.032

values, their expectations, and the (2.5 to 97.5%) envelope. Again, p values of 1 have been eliminated prior to this analysis. Including all results for both males and females, a total of 42 individual trend tests have been conducted. The observed 1-p's fluctuate about their expected values. The three lowest p values are shown in Table 4.

Despite the fact that the smallest p value among the 42 is 0.01, this value is not very far from its expectation and considerably higher than the 2.5% value. The other small p values are not unusual at all.

p-Value Plots: Computer-Simulated Data

The examples in the previous section illustrate the concept that small p values can easily arise due to chance. Their presence is consistent with an overall null hypothesis of no treatment-related effects, as shown by the p value plots. But is the converse true? When there are real treatment effects, do p-value plots appear different than they do without such effects present?

In an effort to shed some light on these questions, we conducted several computer simulation studies where we generated binomially distributed data (using the SAS function RANBIN), statistically analyzed them, and produced corresponding p value plots. We provide the details behind the simulations and our results in this section.

The statistical procedure used throughout was the Cochran-Armitage (CA) trend test (*12,13*). For the first set of simulations, we generated data consistent with the null hypothesis of no treatment effect. Initially 5000 independent sets were generated, each set consisting of a control and three nonzero dose groups of size 60 (doses 1, 2, 3). Of these, 1000 were generated with a spontaneous tumor rate of 2%, 2000 were generated with a rate of 5%, and the remaining 2000 were generated with a spontaneous rate of 10%. With these 5000 cases as input data to the CA test, we calculated rejection rates of the test for both a continuity-corrected and noncontinuity-corrected version. The results are shown in Table 5.

Thus, with spontaneous rates of 10% or less, the noncontinuity corrected version of the CA test operates close to its nominal level. When tumor rates are low (i.e., 2%), the test is somewhat conservative, but not nearly as conservative as Fisher's exact test would be in the same situation. In what follows we continue to work with the CA test without continuity correction.

The 5000 p values from this null case were randomly grouped into 200 sets with 25 tumor types/sites per set, such that, within each set, 5 came from the 2% background rate, 10 came from the 5% background rate, and the other 10 came from the 10% background rate. This gave us a collection of 200 studies, each with 25 tumor types/sites per study. Within each study, the p values were ordered from smallest to largest. We then investigated the distributions of the order statistics obtained in this way.

In particular, we examined the 25th percentile (over the 200 studies), median (over the 200 studies), and 75th percentile (over the 200 studies) of the ordered p values from the 200 studies. These were plotted along with expected values and the (2.5 to 97.5%) envelope of the p-value plot as defined earlier. Figure 5 displays the results. Notice that the median of the 200 studies matches closely with the expected values and that the 25th and 75th percentile curves both lie well within the envelope.

A second set of independent simulations was generated to represent the situation where treatment effects were present. The method of generating the data was the same as for the null case, except at two sites per study. Instead of counts generated at a constant 5% rate, counts for these two sites were generated as independent binomials with rates 0.005, 0.025, 0.075, and 0.100, respectively, in the four groups (doses 0, 1, 2, 3). Cochran-Armitage test statistics and corresponding p values were calculated as described.

Table 4.

Observed p	Expected p	Lower 2.5 percentile
0.010	0.023	0.001
0.048	0.047	0.006
0.056	0.070	0.015

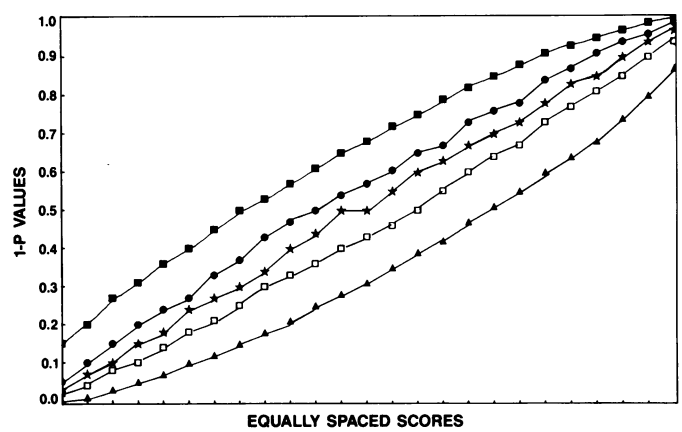

FIGURE 5. Simulation results, case 1 (no treatment effects). Twenty-fifth percentile of simulation runs (□), median of simulation runs (*), 75th percentile of simulation runs (●), 2.5% envelope (▲), and 97.5% envelope (■) versus equally spaced scores [$i/(n + 1)$].

Table 5.

Background rate	Number of cases	No continuity correction: Proportion rejected (α = 0.05)	No continuity correction: Proportion rejected (α = 0.01)	With continuity correction: Proportion rejected (α = 0.05)	With continuity correction: Proportion rejected (α = 0.01)
0.02	1000	0.035	0.005	0.014	0.004
0.05	2000	0.052	0.006	0.035	0.005
0.10	2000	0.053	0.012	0.039	0.008

We again examined the 25th percentile, median, and 75th percentiles from the ordered p values for the 200 studies. Figure 6 shows a very different pattern from the one presented in Figure 5. Although the lower left portion of Figure 6 looks similar to Figure 5, all three curves from the simulated data lie above the expected values in the right portion of the figure (corresponding to small p values or large 1-p values). For the largest 1-p values, all the curves are close to the upper envelope with the median curve being almost coincident at the two largest 1-p values (smallest p values).

Thus, we see that, at least in the long run, p-value plots generated with treatment effects present do look substantially different from those produced under the null case. Theoretically then, they can be used informally to help distinguish true positive results from false positives.

Decision Rules Based on *p*-Value Plots

We can further use the simulated data presented in the previous section to evaluate various decision rules in terms of both level and power. One such decision rule (call it CA01) would be analogous to the one studied by Haseman (*4*), i.e., declare a positive finding if the smallest p value is less than 0.01. (We assume here that all tumors are common tumors.)

In addition, we can consider several decision rules based on whether one or more of the largest 1-p values lies above a suitably chosen upper envelope. We have studied both the 97.5%(A) and 95%(B) envelopes and based decision rules on the 1, 2, or 3 largest 1-p values. Thus, the decision rule B3 will declare a positive finding if any one of the three largest 1-p values exceeds the 95% envelope value.

Results from the first simulation case with no treatment-related effects give us information on level. Results from the second simulation case give us information on power. These are summarized in Table 6.

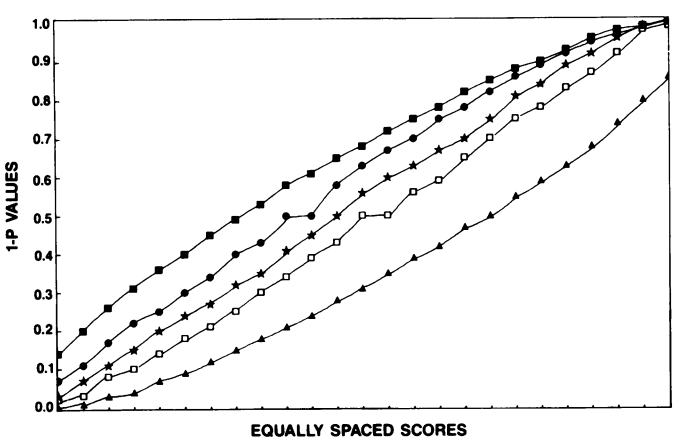

FIGURE 6. Simulation results, case 2 (treatment effects at two sites). Twenty-fifth percentile of simulation runs (□), median of simulation runs (*), 75th percentile of simulation runs (●), 2.5% envelope (▲), and 97.5% envelope (■) versus equally spaced scores [$i/(n + 1)$].

Table 6. Rejection rates (percent).

Decision rule	Simulation 1 (No treatment effects)	Simulation 2 (Treatment effects present)
CA01	19	91
A1	0.5	51.5
A2	2.5	68.5
A3	4.5	71
B1	2	68
B2	6	82.5
B3	8	85.5

In Table 6, one can see the traditional trade-off between level and power. Increased power comes only at the expense of increasing the experiment-wise false positive rate. For this example, it appears that rule B2 provides a good compromise between level and power. With 25 p values in the plot, rule B2 declares a positive finding if the smallest p value is less than 0.002 or if the next smallest p value is less than 0.014. Simes (*15*) studies a procedure with decision rule rejecting if the ith smallest p value is less than $i\alpha/N$, where N is the number of tests. He proves that this procedure has an overall type I error rate of α when the tests are independent.

Conclusions and Discussion

From our analysis of the mouse study data with chemical 1, we see that statistically significant results can arise due to chance. These data do not, however, provide overwhelming evidence of extrabinomial variation, although such an hypothesis cannot be rejected, either. As Haseman and colleagues point out (*6*), such data reinforce the idea of cautiously interpreting statistically significant results. Even the use of the more conservative decision rule based on $p < 0.01$ may not be conservative enough for some data sets.

We have developed and advocated the p-value plot as an aid in decision-making for carcinogenicity studies. These plots are useful diagnostic tools to informally assess the overall treatment-related effects of the experimental compound. When many statistical tests have been made, this fact is automatically incorporated. The plots easily exhibit the conservative nature of some tests (contrast, for example, Fig. 1 with Figs. 3 and 4). By eliminating results from tests with p values equal to 1 (or for which the total number of tumors in all groups is very low), the observed 1-p values will often fluctuate about their expected values, thus eliminating potential bias from the procedure. The plots can distinguish cases in which treatment effects are present from the null case (compare Figs. 5 and 6).

Through the use of a limited set of computer simulations, we have evaluated several decision rules. With a distribution of spontaneous tumor rates considerably higher than those from the 25 NTP studies reviewed by Haseman, and with the Cochran-Armitage trend test instead of the more conservative Fisher's test, we saw the rule using $p < 0.01$ had a false positive rate of 19%. In contrast, using rules constructed from p-value plots, we were able to reduce this rate to acceptable levels. Of course, this

resulted in a slight loss in power to detect treatment-related effects.

Although it may appear that our findings are inconsistent with those of Haseman, this is not so. The historical control tumor rates reported in Haseman (*4*) contained many low rates. Nine of the 27 type/sites had rates less than 2% in rats and seventeen of the 27 had rates less than 2% in mice. Had these rates been higher, the overall false positive rates reported by Haseman would have been greater. Second, Haseman's analysis was based on Fisher's exact test, which is very conservative, regardless of the background rate. In contrast, we showed that the Cochran-Armitage trend test operates close to its nominal level when the spontaneous tumor rate is in the 5 to 10% range. Thus, our studies are complementary rather than contrary to those of Haseman.

Because hypothesis testing seems to be the most common statistical approach to the analysis of data from carcinogenicity studies, we believe that it is therefore most appropriate (in the spirit of the Neyman-Pearson approach to hypothesis testing) to first control the Type I error rate in these studies. By employing statistical techniques that control the overall false positive rate, such as those based on p-value plots or other methods of adjustment (*11,16,17*), this goal can be achieved. Several such decision rules have been proposed in this paper and evaluated in the limited set of simulation studies conducted. Other such rules can easily be contemplated. Their properties can further be investigated under a variety of conditions, again with computer simulation.

The main advantage of p-value plots is that they provide the statistician and toxicologist with a simple way to visually summarize the results of numerous statistical tests and compare the p values obtained to those that would be expected in the absence of compound-related effects. By examining the shape of the observed 1-p curve as well as its largest values, information about the overall effects of treatment may be deduced. Thus, we believe that such graphical evaluations of data may serve as a useful tool in interpreting results from carcinogenicity studies. The final decision process, however, needs to be an interdisciplinary effort with input from pathologists, toxicologists, and statisticians (*8*).

Research support was provided in part by SmithKline Beckman Animal Health Products. The author thanks Vernon Chinchilli, Christine Kopral, and the two referees for their constructive comments and suggestions.

REFERENCES

1. Salsburg, D. Use of statistics when examining lifetime studies in rodents to detect carcinogenicity. J. Toxicol. Environ. Health 3: 611-628 (1977).
2. 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).
3. Fears, T. R., and Tarone, R. E. Response to: Use of statistics when examining lifetime studies in rodents to detect carcinogenicity. J. Toxicol. Environ. Health 3: 629-632 (1977).
4. Haseman, J. K. A re-examination of false-positive rates for carcinogenesis studies. Fundam. Appl. Toxicol. 3: 334-339 (1983).
5. IARC. Long-Term and Short-Term Assays for Carcinogens: A Critical Appraisal. IARC Scientific Publications, No. 83, Lyon, 1986.
6. Haseman, J. K., Winbush, J. S., and O'Donnell, M. W., Jr. Use of dual control groups to estimate false positive rates in laboratory animal carcinogenicity studies. Fundam. Appl. Toxicol. 7: 573-584 (1986).
7. IARC. Guidelines for simple sensitive, significance tests for carcinogenic effects in long-term animal experiments. Annex to Long-Term and Short-Term Screening Assays for Carcinogens: A Critical Appraisal. IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, Supplement 2. International Agency for Research on Cancer, Lyon, 1980, pp. 311-426.
8. Gart, J. J., Krewski, D., Lee, P. N., Tarone, R. E., and Wahrendorf, J. The Design and Analysis of Long-Term Animal Experiments, IARC Scientific Publications, No. 79, Lyon, 1986.
9. Tarone, R. E. The use of historical control information in testing for a trend in proportions. Biometrics 38: 215-220 (1982).
10. Dempster, A. P., Selwyn, M. R., and Weeks, B. J. Combining historical and randomized controls for assessing trends in proportions. 78: 221-227 (1983).
11. Meng, C. Y. K., and Dempster, A. P. A Bayesian approach to the multiplicity problem for significance testing with binomial data. Biometrics 43: 301-312 (1987).
12. Cochran, W. G. Some methods for strengthening the common χ^2 test. Biometrics 10: 417-451 (1954).
13. Armitage, P. Tests for linear trends in proportions and frequencies. Biometrics 11: 375-386 (1955).
14. Bickis, M., and Krewski, D. Statistical design and analysis of the long-term carcinogenicity bioassay. In: Toxicological Risk Assessment, Vol. 1 (D. B. Clayson, D. Krewski, and I. Munro, Eds.), CRC Press, Boca Baton, FL, 1985, pp. 125-147.
15. Simes, R. J. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 751-754 (1986).
16. Heyse, J. F., and Rom, D. Adjusting for Multiplicity of statistical tests in the analysis of carcinogenicity studies. Biometrical J., 30: 1-15 (1988).
17. Westfall, P. H., and Young, S. S. P-value adjustments for multiple tests in multivariate binomial models. Presented at the 1988 Spring Biometrics meeting, Boston, MA.