A Case Study of Statistics in the Regulatory Process: The FD&C Red No. 40 Experiments [1, 2]

Stephen Lagakos [3, 4, 5] *and* **Frederick Mosteller** [3, 6, 7]

ABSTRACT—Unexpected findings in a mouse study in which the safety of FD&C Red No. 40 (Red 40) was examined led to additional experimentation and to new statistical analyses and models. The possibility of acceleration of tumors raised questions about an operational definition of acceleration and of appropriate statistical methods for assessing acceleration, especially in the face of data dredging. The evaluation of Red 40 was further complicated by cage and litter effects and the multigenerational design. In this report the investigations of these studies are reviewed and are used to illustrate how new scientific work can emerge through the regulatory process. A number of issues in animal experimentation that need to be examined are indicated.—JNCI 1981; 66:197-212.

This paper describes and discusses recent scientific investigations of Red 40, a monoazoaryl disodium disulfonate that has been approved in the United States since 1971 as a color additive in foods and drugs.

In 1974 and 1975, the patent holder for Red 40 initiated two chronic feeding experiments—one in rats and one in mice. The motivation for the experiments was to obtain the approval of Red 40 by the Governments of Canada and Great Britain for use in their countries.

In the mouse experiment, questions arose about the presence of a possible "acceleration" effect—a decreased latency period without an accompanying elevation of overall tumor incidence. These questions led to a second mouse experiment, to numerous reports by in-house and outside committees of FDA, and to public meetings.

We present a case study of these investigations. It is not our intent or our province to judge the safety of Red 40 or the regulatory action taken. Rather, we use these instructive investigations to illustrate a number of complex and important issues in the design, conduct, analysis, and interpretation of animal experiments that test for carcinogenicity.

HISTORY

Red 40 was approved by FDA in 1971 and soon became widely used in foods and drugs. With the termination of the provisional listing of FD&C Red No. 2 in 1976, the use of Red 40 further increased. Among color additives, it soon ranked, in pounds consumed, second only to FD&C Yellow No. 5 (*1*).

Although no evidence questioned its safety, the experimental testing of Red 40 as of 1974 was insufficient for approval in Canada and Great Britain: A lifetime carcinogenesis experiment had been undertaken in rats, but no similar tests had been performed in mice. To satisfy these testing requirements, Allied Chemical Company (Morristown, N.J.), the patent holder for Red 40, initiated two chronic feeding experiments, one in rats in late 1974 and one in mice in 1975. The experimental requirements were suggested by the Canadian Government after discussions with Allied Chemical Company and were sent to FDA for comment. The mouse experiment consisted of a control group and 3 dose-level groups exposed to Red 40 and was done by Hazleton Laboratories (Vienna, Va.).

In early 1976, FDA met with Allied Chemical Company and Hazleton Laboratories to review the preliminary experimental data. Allied Chemical Company reported unremarkable findings from the rat study, but unexpected occurrences in the mouse experiment. In the exposed mice, 6 tumors of the RE system occurred early in the experiment, 1 in each of the low- and middle-dose groups and 4 in the high-dose group. In contrast, none appeared in the control group. With the experiment still in its early stages, these deaths might have been due to statistical fluctuation, though they might have been a signal that Red 40 had some effect on RE tumors.

The FDA made two suggestions. The first was to kill

ABBREVIATIONS USED: FDA = Food and Drug Administration; FD&C = Food, Drug, and Cosmetic; RE = reticuloendothelial; Red 40 = FD&C Red No. 40.

---

[1] Received August 1, 1980; accepted August 1, 1980.

[2] Supported in part by National Science Foundation grant SOC75-15702 and Public Health Service grant CA00505 from the National Cancer Institute.

[3] Department of Biostatistics, School of Public Health, Harvard University, Boston, Mass. 02115.

[4] Biostatistics and Epidemiology Division, Sidney Farber Cancer Institute, 44 Binney St., Boston, Mass. 02115.

[5] *Address reprint requests to* Dr. Lagakos *at:* Department of Biostatistics, School of Public Health, Harvard University, 677 Huntington Ave., Boston, Mass. 02115.

[6] Department of Statistics, Harvard University.

[7] We thank Albert C. Kolbye, Jr.; Robert L. Elder; James Winbush; the staff of the Food and Drug Administration; members of the Working Group; John J. Gart; M. A. Gross; Jerome Cornfield (deceased); John J. Crowley; Thomas R. Fleming; and Bernard G. Greenberg for advice, critical discussions, and various courtesies. We are also grateful to John C. Bailar, III; John Emerson; and the Buffalo Color Corporation and its representative A. J. ten Braak for useful comments and criticisms. We appreciate Agnes Herzberg's design suggestions, the technical support of Gregg Dinse, and the editorial assistance of Marian McGrath. The cooperation and courtesies described above do not necessarily imply support, either in part or in whole, of the analyses and interpretations given in this paper.

and examine what amounted to 36% of the mice in the 4 groups. The purpose seems to have been to determine whether Red 40 accelerated the growth of preexisting RE tumors or whether it affected the development of RE tumors. Killing was done at week 42 of the scheduled 2-year study, and no RE tumors were detected in either the control or exposed mice.

The FDA's second suggestion was to initiate a second, much larger, mouse experiment consisting of 2 control and 3 exposed groups. The 3 exposed groups had the same dose levels of Red 40 as did the mice in the first experiment. We do not know why 2 control groups were created or why they were not regarded as a single large control group. However, the fact that 2 control groups were used led to important issues in evaluating the second experiment.

In late 1976, FDA created a Working Group of scientists from the FDA, National Cancer Institute, and National Center for Toxicological Research to monitor the rat and two mouse experiments and to analyze the data. In January 1977, the Working Group completed the first of two interim reports (*1*) on the experiments. The report considered questions of experimental design and pathology and included a detailed statistical analysis. The report concluded that the data "suggest but are too preliminary to demonstrate an association between the incidence or decreased latency of lymphomas or leukemias and exposure to FD&C Red No. 40 in CD-1 mice."

Approximately 1 year later, the Working Group completed a second interim report (*2*) based on complete data from the first experiment and on interim data from the second. The report concluded that "these experiments provide no evidence at this time that FD&C Red No. 40 is carcinogenic but we recommend that a final assessment be made when the second study is completed."

At about the same time, Dr. M. Adrian Gross from the FDA's Bureau of Drugs issued two memoranda (*3*, *4*) giving his own analysis of the mouse experiments. Gross reported that he learned about the preliminary data from the first mouse experiment from Dr. Michael F. Jacobson, a member of the Center for Science in the Public Interest, a consumer lobbyist organization. The memoranda concluded that the first mouse experiment gave clear evidence of an "acceleration" effect due to Red 40, i.e., of a decreased latency period without a corresponding increase in overall tumor incidence. Moreover, they criticized the statistical methods used by the Working Group, particularly those used to analyze time to death and to adjust statistical significance levels for the multiple-hypotheses aspects of their analyses. The memoranda also asserted that the second experiment was of little value because its 2 control groups displayed a substantial difference in RE tumor incidence rates. The "Appendix" of (*4*) comments on several internal memoranda prepared by members of the Working Group in response to (*3*). Dr. John Gart of the Working Group has made available to us two memoranda (*5*, *6*) that he prepared.

The FDA was thus faced with opposing assessments of the carcinogenicity of Red 40. As a step in resolving these differences, the Commissioner of Food and Drugs, Dr. Donald Kennedy, appointed three outside statistical consultants, provided them with copies of (*1–4*) but no raw data or internal memoranda, and asked them (*7*) to respond independently to the following seven questions:

1) Were the statistical tests used by the Working Group appropriate for detecting acceleration?
2) Would the procedures used by the Working Group detect the acceleration discussed in Dr. Gross's hypothetical example?
3) Would alternate methods (including, but not restricted to those mentioned by Dr. Gross) be more appropriate?
4) Does the pooling of all exposed animals invalidate Dr. Gross's basic contentions about the appropriateness of the Working Group's tests?
5) Is the problem of "unfairly" skewing results by examining the differences at the time of maximal difference (or at ½ time, or at random intermediate time) as small as Dr. Gross suggests?
6) Is it indeed likely that the test animal genetics were such that almost all "susceptible" animals in both the control and exposed groups would eventually die with tumors (constant incidence) although there could be significant differences in rates of tumorigenesis (differential acceleration)? If you are unable to answer this question, to what extent is the truth of that assumption essential to Dr. Gross's argument?
7) Can the second animal study by Hazleton Laboratories throw additional light on the first?

The Commissioner also invited the consultants to make any other suggestions they felt were relevant. The consultants originally invited were Drs. Jerome Cornfield, Bernard Greenberg, and Frederick Mosteller. Dr. Mosteller's request for the addition of Dr. Stephen Lagakos as a coconsultant was granted.

The consultants completed their reports (*8–10*) by October 1978. Their responses to the seven questions were qualitatively similar in many respects. The consultants agreed that the methods used by the Working Group were not particularly oriented to detecting an acceleration effect. Two of the consultants' reports (*8*, *9*) offered, in addition, their own conclusions on the carcinogenicity of Red 40. All three raised new questions concerning the design, analysis, and interpretation of the experiments.

The FDA held public meetings on January 17–18, 1979. The participants included members of the Working Group, Dr. Gross, the four outside consultants

mentioned above, two additional consultants (Drs. John Crowley and Thomas Fleming), other statisticians, laboratory scientists, and members of the general public. Written comments were solicited and received from three other statisticians: Drs. Peter Armitage (*11*), Norman Breslow (*12*), and David Cox (*13*).

Each of Dr. Kennedy's seven questions was discussed at the meeting. The participants raised a number of related issues that further complicated the overall evaluation of Red 40. Some of these issues, discussed in greater detail in subsequent sections, were 1) the choice of statistical tests to compare the outcomes for the control and exposed groups, 2) the assessment of hypotheses suggested in light of the data, 3) the definition of acceleration, 4) the interim sacrifice in the first mouse experiment, and 5) the analysis of deaths without RE tumors.

The Working Group's first report (*1*) noted that the cages housing the mice had not been rotated during the two experiments. Prior to the January 1979 meeting, Drs. Lagakos and Mosteller requested cage information. These data (from the partially completed second expt only) reached all the consultants a few days before the January meeting. Neither cage nor litter effects had previously been examined, and these variables had not been included in any of the analyses of Red 40. Lagakos and Mosteller (*14*) analyzed position effects of cages and found that mice housed in the upper rows of their racks experienced a much higher incidence of RE cancers than those in the lower rows. Greenberg (*15*) analyzed litter data and suggested the possibility of a litter effect in addition to the upper row effect.

Following the meeting, the consultants sent reports to the Working Group (*16*) and the Commissioner (*17*) containing suggestions for the subsequent analysis of the Red 40 data and for research on animal experiments in general.

In April 1979, Lagakos and Mosteller (*18*) submitted to the Working Group a report that was based on the final data from the second experiment. The report analyzed time to and type of death as a function of sex, dose level, and cage information. It contained a detailed analysis of the effects of cage position and raised the possibility of an acceleration effect in the second mouse experiment. Heretofore, discussions of acceleration were confined to the first experiment. The report also pointed out a possible selection bias (*see* "A Possible Selectivity Problem") inherent in both mouse experiments which, if present, would limit their validity.

In June 1979, the Working Group held public meetings. The topics discussed included 1) the effects of cage layout on the type and time of death, 2) the difference in RE tumor incidence rates between the 2 control groups in the second mouse experiment, 3) the presence of litter effects, 4) the evidence of an acceleration effect in the second mouse experiment, and 5) the possibility of a selection bias inherent in both mouse experiments. Each of these issues is discussed in greater detail in subsequent sections.

MOUSE AND RAT EXPERIMENTS

The first mouse experiment consisted of 400 noninbred CD®-1 HaM/ICR mice, 50 of each sex from a control and 3 dose-level groups. The high-, medium-, and low-exposure levels amounted as a percentage of diet to 5.19, 1.39, and 0.37% Red 40, respectively. The high dose was believed to be approximately the maximum tolerated dose. A detailed description of the experimental design is given in (*1, 2*), and certain aspects of the design are considered later in this paper.

Upon death, the presence and type of tumors were noted in each of the 400 experimental mice. For the purposes of this paper and during most of the discussions of Red 40, the mice were classified on the presence or absence of RE tumors. (We have not explored other possible categories of tumors as a basis for the analysis of these mouse expts.) Also noted were the kind of death (natural or killed) and the age at death. With the exception of the mice killed at week 42, all mice died naturally or were killed when the long-term phase of the study had reached 104 weeks.

The second mouse experiment was similar in design to the first, except that it included 2 control groups and consisted of 100 mice/sex/group, for a total of 1,000 mice. All mice in this experiment died naturally or were killed during weeks 109–111.

The rat experiment was initiated shortly before the first mouse experiment and consisted of 50 CD albino rats per sex in each of 4 groups (1 control and 3 exposed). The dose levels of Red 40 were identical to those used in the mouse experiments. One important difference in the rat study was that the indicator tumors considered (i.e., pituitary, lymphoreticular, mammary, and uterine) were thought not to be rapidly fatal and, as a result, the Working Group's evaluation did not include time-to-death analyses.

CHOICE OF MODELS AND OF STATISTICAL TESTS

The January 1979 meeting discussed at length the choice of statistical tests for comparing a control group C and an exposed group E on the basis of time to death with RE tumors ("RE death"). Actually, the mouse experiments included three exposure levels as well as males and females. Our purposes in this section, however, are to discuss properties of statistical tests and we merely consider the situation with a single control and a single exposed group. Thus this discussion is more simplified than is required for the actual experiments.

Survivorship Functions and Competing Causes of Death

The time-to-death tests used by the Working Group, as well as those suggested by Dr. Gross, compare the survivorship function G_C for the control group with the survivorship function G_E for the exposed group.

These functions are like life tables and start at 1 (or 100%) when all animals are alive (time zero) and as time passes decrease toward 0 (or 0%) when all animals have died. They represent the distribution of latency times to RE death, which are not always observed because of deaths from other causes. We must estimate these functions from the data.

Let us define the G functions. Think of a single mouse as having two potential times to death—a potential time to RE death and a potential time to non-RE death. These times are assumed to be statistically independent and to "compete" with one another, with the smaller time determining the actual observed time and type of death. We never observe the larger time.

Then G_C is the survivorship function for the potential time to RE death in the control group, and G_E is the corresponding function for the exposed group. A statistical complication arises because in some mice, namely those with non-RE deaths, only a lower bound for the potential times to RE deaths is observed.

If no differences exist between the control and exposed groups, we would have $G_C(t)=G_E(t)$ for all t. Conversely, inequality for the G's indicates an exposure effect.

Power of Statistical Tests

The power of a statistical test is the probability that it will detect a given difference between the exposed and control populations on the basis of sample data. Some statistical tests are omnibus in the sense that they have some power against nearly all types of differences. Chi-square tests for goodness-of-fit have this feature. Some tests, however, are directional in that they have a great deal of power to detect some kinds of differences and relatively little for others. For example, the familiar t-test for a difference of means has high power to detect shifts in means between two distributions but not to detect that one is unimodal and the other bimodal. If we use an omnibus test, we get some power for many kinds of differences, but we pay for this by not concentrating the evidence on important differences that may be likely to occur.

The choice of a test statistic depends on the analyst's experiences with the kind of investigation. For example, if one kind of difference seems more plausible than others, the analyst usually chooses a test that directs the information in the data toward the anticipated kind of difference. If the analyst has no prior expectations about the kinds of differences that may occur or if differences of all types are realistically possible, a more omnibus test may be preferable.

Hazard Functions

The time-to-death tests used by the Working Group are the proportional hazards and generalized Wilcoxon tests. From many viewpoints, both are desirable tests. Variants of the former are also called the Cox, Log-rank, or Mantel-Haenszel test, and the latter is sometimes called the Breslow, Gehan, or Gilbert test. To explain the behavior and properties of these tests, let us reexpress $G_C(t)$ and $G_E(t)$ in terms of their corresponding hazard functions $h_C(t)$ and $h_E(t)$. If the G's are differentiable, the h's are given by

$$h_C(t)=\frac{g_C(t)}{G_C(t)} \text{ and } h_E(t)=\frac{g_E(t)}{G_E(t)},$$

$$\text{where } g_C(t)=\frac{-d}{dt}G_C(t) \text{ and } g_E(t)=-\frac{d}{dt}G_E(t).$$

The cumulative hazard functions $H_C(t)={}_0\int^t h_C(u)du$ and $H_E(t)={}_0\int^t h_E(u)du$ are the areas under the h functions between 0 and t and are related to the G's by $G_C(t)=-\ln H_C(t)$ and $G_E(t)=-\ln H_E(t)$. Roughly speaking, $h_C(t)$ is proportional to the chance of an RE death in a short interval after time t, given that the control animal has survived to time t. A similar interpretation holds for h_E for the exposed group. An h is also sometimes referred to as the age-specific mortality function or the cause-specific hazard function and by actuaries as the cause-specific force of mortality. Unlike the G functions, which have no physical meaning without the sometimes questionable assumption of independence of potential times to failure, the h functions can be defined in terms of observable quantities and hence are always interpretable (*19*).

Proportional Hazards and Generalized Wilcoxon Statistics

The proportional hazards test is directed to alternatives having $h_E(t)=kh_C(t)$, i.e., alternatives having proportional hazards. This multiplicative relationship simplifies to equality if $k=1$. If $k>1$, the exposed group has a larger hazard than the control group.

The proportional hazards test statistic can be written (*1*) as

$$T=\sum_i W_i[\hat{h}_E(t_i)-\hat{h}_C(t_i)], \qquad [1]$$

where the W_i are weights, the t_i are the ordered times to death, and $\hat{h}_E$ and $\hat{h}_C$ are estimates of h_E and h_C, the values of the true hazard functions. The weight W_i is related to the total number of mice surviving in both groups at time t_i. Equation [1] shows that the statistic T is merely a weighted sum of the differences between the estimated hazard functions of the exposed and control groups at the times of RE deaths. Note that if $k=1$ (no difference between control and exposed distributions), the estimates $\hat{h}_E(t)$ and $\hat{h}_C(t)$ should be close in value; hence we expect to observe small values for T. However, if $k>1$ (i.e., exposure increases the hazard), larger values of T would tend to occur; the larger the k, the larger the T, on the average.

Associated with each T is a P-value, the probability of a larger T than that observed, provided $k=1$. Thus

P-values appraise the rarity of large T's if the populations are identical.

As pointed out by the Working Group, the generalized Wilcoxon test statistic is similar to the proportional hazards test statistic except that the W_i's in equation [1] are squared. This squaring leads to a test that places more emphasis on earlier differences between $h_E(t)$ and $h_C(t)$ than does the proportional hazards test. Both tests have good power against differences in which $h_E(t) > h_C(t)$ for all t, even when the hazard functions are not proportional. They will even have good power if $h_E > h_C$ for most of the weight of the data and if the reversals are small in magnitude and have small weight.

Crossing Hazard Functions

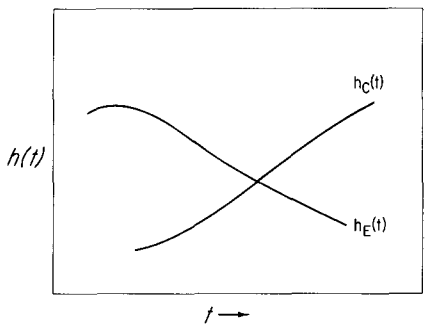
Text-figure 1 represents a situation in which the proportional hazards test can have poor power. As shown, $h_E(t)$ exceeds $h_C(t)$ at the left and is less than $h_C(t)$ on the right. For this "crossing hazards" situation, the corresponding cumulative hazard functions H_C and H_E first separate from one another, later move together again, and possibly cross. If we use the proportional hazards test statistic in this situation, the early terms in the summation in equation [1] tend to be positive and the later terms tend to be negative, which thereby causes a "cancellation" effect, with the degree of cancellation depending on the number and magnitude of negative terms. (We say "tend" because T is composed of estimates of differences rather than actual differences.) In other words, h_E and h_C could differ and still be likely to lead to small T-values that are not statistically significant at any of the usual levels. If exposure shortens the time to death only of mice destined to develop RE tumors and leaves other mice unaffected, this "acceleration effect" will produce crossing hazard functions.

A more omnibus test for comparing G_E and G_C could be obtained by a test statistic of the Kolmogorov-Smirnov type, e.g.,

$$\tilde{T} = \underset{t}{\text{maximum}} |\hat{G}_E(t) - \hat{G}_C(t)|.$$

Fleming (*20*) discussed the characteristics and use of this type of test at the January 1979 meeting. The test has some power against a broader class of alternatives to equality than the proportional hazards or generalized Wilcoxon tests. It can, however, be considerably less powerful than these tests in situations where h_E and h_C do not overlap. A simple modification of the proportional hazards or generalized Wilcoxon test that has better power against crossing hazards would be as in equation [1], but with the summation running only until a fraction (e.g., ½) of the RE deaths occur.

To summarize, statistical tests have good power against some types of alternatives and poorer power against other types of alternatives. In particular, the proportional hazards and generalized Wilcoxon tests have good power if $h_E(t) > h_C(t)$ for all t, but they could have poor power against alternatives that take the form of crossing hazard functions.



TEXT-FIGURE 1.—An illustration of crossing hazard functions.

FIRST MOUSE EXPERIMENT

Table 1, compiled from data in (*4*), gives a summary of the time-until-death and type-of-death data in the first mouse experiment. Let us consider first the overall numbers of mice with RE tumors for each sex and dosage group. Among females 8, 8, 8, and 9 had RE tumors in the control, low-, medium-, and high-dose groups, respectively. Among males 3, 3, 3, and 4 had RE tumors in the corresponding groups. Within each sex, the numbers of RE tumors in the 4 groups were remarkably similar: They varied considerably less than would be expected if no exposure effects existed. Clearly, on the basis of overall RE incidence, the data give no suggestion of an exposure effect. It is important to keep in mind that obtaining these comparisons across dose levels achieves one of the primary objectives of the experiment.

With this background we turn to the time-to-death data. The experiment has several aspects that are relevant to a detailed analysis, but here we consider the general issue of statistical tests for time to RE death as previously discussed.

Text-figure 2 gives Kaplan and Meier estimates (*21*) of the functions G_E and G_C for the female mice. Text-figure 3 gives the corresponding plots for male mice. Both graphs are drawn on semilogarithmic paper, so that they also depict the cumulative hazard functions H_E and H_C. In both text-figures, the H_E and H_C functions at first separate and later come closer together. If indicative of the true distributions, this behavior represents a crossing hazards situation, so the proportional hazards or generalized Wilcoxon tests might have little power. This crossing-hazards phenomenon is also suggested by table 1, because the RE deaths in the exposed group tend to occur sooner than those in the control group. Indeed, the first 14 RE deaths were in mice exposed to Red 40.

A vital question is whether these estimated curves reflect crossing hazards in the population, for then the proportional hazards and generalized Wilcoxon tests are not such good choices. If the underlying curves in the population are *not* crossing and what we see in

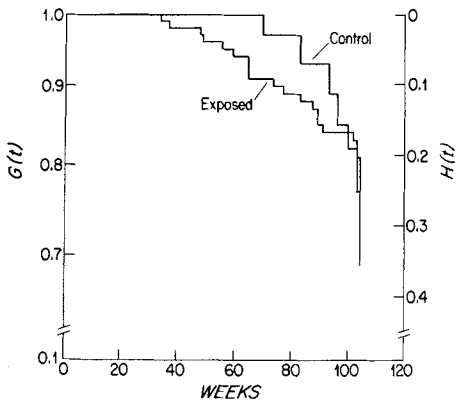
TABLE 1.—*First mouse experiment: Time to and types of death*

Category	Control		Low dose[a]		Medium dose		High dose	
	Females	Males	Females	Males	Females	Males	Females	Males
No. of mice killed at 42 wk	20	17	20	16	18	19	18	14
No. of mice killed at 104 wk	18	12	13	19	20	12	21	19
With RE tumors	2	0	3	0	2	1	3	0
Without RE tumors	16	12	10	19	18	11	18	19
Times of natural death, wk	70[b]	29	49[b]	27	30	5	34[b]	2
	77	30	60[b]	31[b]	37[b]	54[b]	36[b]	16
	83[b]	38	63	31	56[b]	67	48[b]	23
	87	48	67	35	65[b]	70	48	34
	92	53	70	55	76	79	65[b]	35[b]
	92	56	74[b]	79	83[b]	80	91[b]	35[b]
	93[b]	62	77[b]	81	87[b]	83[b]	91	67[b]
	96[b]	70	80	83[b]	90	86	98	77
	100[b]	71	80	87[b]	94	88	102	77
	102	74	89[b]	93	97	91	102	79[b]
	102	74	89	94	97	93	103[b]	84
	103[b]	76	90	99	102[b]	94		89
		85[b]	90	101		95		92
		86	97	102		97		96
		86	100			100		97
		92	102			100		98
		97[b]				100		99
		99[b]				103		
		101				103		
		102						
		103						
Total mice with RE tumors	8	3	8	3	8	3	9	4
Total mice without RE tumors	42	47	41	46	42	47	41	46

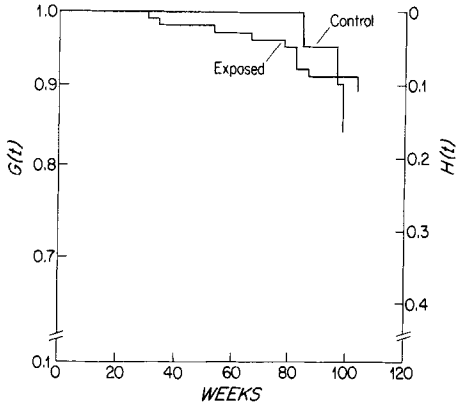
[a] Excludes 1 mouse of each sex that was missing or autolyzed.
[b] RE tumor present.

these samples is simply random fluctuation, then these tests are appropriate. The proportional hazards and generalized Wilcoxon tests, when applied to these data, produced nonsignificant *P*-values. However, when the same tests were applied earlier in chronologic time, the estimated *G*-functions had not yet begun to come together, and they produced statistically significant results. More generally, in a crossing-hazards situation, it would not be surprising for these tests to give statistically significant results midway through an experiment and nonsignificant results at the end.

The conclusions one reaches from these data depend heavily on prior expectations about the types of differences likely to occur. It may be thought that the proportional hazards and generalized Wilcoxon tests are especially appropriate for this sort of study and that analyses oriented toward strength in detecting crossing hazard functions are inappropriate and in-



TEXT-FIGURE 2.—First mouse experiment. Estimates of $G(t)$ for control (G_C) and exposed (G_E) female mice. The three exposure levels have been pooled.



TEXT-FIGURE 3.—First mouse experiment. Estimates of $G(t)$ for control (G_C) and exposed (G_E) male mice. The three exposure levels have been pooled.

effective uses of the data. For example, if someone believed that acceleration could not occur in such experiments, he or she may not want to guard against crossing hazards. In this sense, the proportional hazards and generalized Wilcoxon tests, though not oriented to detecting acceleration, are appropriate for comparing the control and exposed groups. Other investigators might think otherwise.

ALTERNATIVE FORMULATION

The time-to-death tests used by the Working Group (*1, 2*) and Gross (*3, 4*) compare the control and exposed groups on the basis of the competing-risk survivorship functions $G_C(t)$ and $G_E(t)$. The disagreement was on the choice and properties of specific statistical tests for comparing the G's.

Due to when and how they entered the Red 40 discussions, Cornfield (*8*), Greenberg (*9*), and Lagakos and Mosteller (*10*) analyzed time-to-death differently than did the Working Group and Gross. Cornfield (*8*) and Lagakos and Mosteller (*10*) focused on the survivorship function $F(t)$ of the conditional distribution of time to death, given the presence of an RE tumor. When defined for control and exposed groups, this leads to $F_C(t)$ and $F_E(t)$, respectively. Consider a population of N mice among which n mice have RE deaths and the remaining $N-n$ mice have non-RE deaths. Then $F(t)$ is the proportion of the n RE deaths that occur after age t. To interpret $G(t)$ analogously, the $N-n$ unobserved "potential" RE death times, as descrubed in "Choice of Models and of Statistical Tests," must be considered for mice with non-RE deaths. Then $G(t)$ is the proportion of times that exceed t in the pool of the n observed and $N-n$ potential RE death times.

A detailed discussion and comparison of the F- and G-functions is given in (*10*). The point we want to emphasize here is that F and G describe somewhat overlapping but different aspects of the joint distribution of time to and type of (i.e., RE vs. non-RE) death. As a result, a comparison of F_C and F_E is not the same as a comparison of G_C and G_E.

In controlled experiments in which exposure has no effect on outcome, $F_C = F_E$ and $G_C = G_E$. Hence comparisons based on either F's or G's are appropriate. When exposure is related to outcome, however, the effect will manifest itself differently in the F's than in the G's. This means that in some situations it may be preferable to base comparisons on G_C and G_E, but in others it may be better to use F_C and F_E. When overall RE mortality rates for the control and exposed groups are equal, comparisons based on the F-functions are particularly oriented toward acceleration. Thus it is not surprising that the tests applied by Cornfield (*8*) and Lagakos and Mosteller (*18*) gave smaller P-values than those obtained by approaches based on the G's. (We delay to the next section treatment of the possibility that acceleration was suggested by the data.) In other circumstances, e.g., when the amount and pattern of intercurrent deaths or sacrificing differ between the control and exposed mice, comparisons based on overall RE mortality and the F-functions can be misleading, and use of the G-functions would be preferable.

In choosing a specific test statistic for comparing F_C and F_E, the analyst is faced with the same kinds of considerations of power as in comparisons of G_C and G_E. Some tests are well suited for certain types of differences between F_C and F_E, whereas other tests will be better in other situations. When there is no interim censoring, a proportional hazards test, similar in form to equation [1] but based on F_C and F_E, can be used (*18*). This test will have good power when the hazard functions for F_C and F_E do not cross. This illustrates the fact that for detecting acceleration, the proportional hazards test may have relatively poor power when applied to G_C and G_E, but good power when applied to F_C and F_E. More generally, in comparisons of control and exposed groups, the choice of formulations (e.g., G vs. F) as well as particular tests (e.g., proportional hazards, generalized Wilcoxon) can be important. In many experiments, of course, the exposure effect is so strong that a simple comparison of tumor incidence rates is sufficient.

ASSESSING HYPOTHESES SUGGESTED IN LIGHT OF THE DATA

It is important to recognize that the hypothesis of acceleration was not specified in advance of the first mouse experiment, and it does not seem to be a primary end point in many other carcinogenesis experiments. Nevertheless, acceleration of tumors has been recognized as a form of carcinogenesis by both FDA (*22*) and WHO (*23*). It became an issue here only after the data from the first mouse experiment suggested it as a possibility.

Nearly all statisticians agree that for a hypothesis suggested in the light of the data, significance levels (P-values) computed in the usual ways do not carry the same meaning as they would if the hypothesis were prespecified. Beyond this, we have no generally accepted way of interpreting these significance levels. The following passage, from (*16*), provides what we believe to be a useful statement of the present understanding. It applies to other problems as well as that of carcinogenesis testing:

> In analyzing experiments on carcinogenesis, good statistical practice requires examining the data from many points of view, not all of which can be specified in advance. Such examination may lead to the development of hypotheses, which, ideally, should then be evaluated by use of a data set independent of the set which suggested them. In carcinogenesis experiments, which normally take 2 to 3 years, this is a counsel of perfection, and in practice the hypotheses must often be statistically evaluated using the same data set which generated them. There is near universal agreement among statisticians that low P-values obtained when the hypothesis is suggested by the data may provide less evidence against the null hypothesis than would the same low P-value obtained for a pre-specified

alternative. But there is no consensus why this is so and on how to proceed with the necessary adjustment.

A number of techniques are available which assist one in thinking about the problem but which should not be viewed as providing exact, unequivocal evaluations. A class of procedures, often called multiple comparison procedures, can provide exact evaluations in some cases, but only after specifying a set of statements for which one can calculate the *simultaneous* probability that all are simultaneously correct. But the choice of this set, which can have an important influence on the calculated *P*-value, is a matter of judgement and cannot be settled mathematically. Furthermore, even after settling on a set, it is not always clear that the calculated probability provides an appropriate characterization of the uncertainties involved.

A common multiple comparison procedure uses the first term of the Bonferroni inequalities.[8] This always provides an upper limit to the correct *P*-value for the set selected. When the statements are nearly independent, the upper limit will be very close to the correct value, but when they are highly correlated, it may be very far from it. Thus, even aside from the problems mentioned above, non-quantifiable judgment cannot always be avoided in the statistical evaluation.

A less traditional view of the evaluation of hypotheses suggested by data, but one which commands support in some parts of the statistical community, is that the plausibility of hypotheses suggested by examining the data must enter into the statistical evaluation. But this is inevitably a matter of judgment and biological rather than statistical judgment. We would not recommend that quantification of such judgments be undertaken by anyone uncomfortable with this task. But it is also true that qualitative plausibility judgments are often necessary even though there may be no objective, universally accepted way of arriving at them.

In their evaluations of the acceleration hypotheses, the Working Group (*1, 2*) used the Bonferroni inequality and corrections for repeated significance testing. Gross (*3, 4*) did not attempt adjustment. Cornfield (*8*) approached the problem by analytically combining the evidence in the data for acceleration with prior beliefs of its plausibility to produce posterior odds for acceleration.

It was our personal impression that at the close of the January 1979 meetings, most present thought that the evidence for acceleration was not conclusive, although their feelings about its plausibility varied. Some said that they knew of no biologic mechanism that would lead to decreased latency without increased RE incidence, even though FDA and WHO had considered this possibility earlier.

[8] Consider n independent events, each with probability α. The probability of at least one occurring is $1-(1-\alpha)^n$. If α is small and n is not too big, this quantity is approximated by $n\alpha$. For example, if each of two independent events has a 0.05 chance of occurring, the probability that at least one occurs is about 0.1; more precisely, it is 0.0975. The fact that 0.1 is larger than 0.0975 explains the use of the expression "inequality." When events are not independent, the approximation also offers an upper bound on the correct probability. These ideas can be generalized and applied to the problem of significance testing. For example, if n hypotheses are tested and it is desired to obtain an overall type I error probability of $P<0.05$, then each hypothesis can be assessed at the $P=0.05/n$ level. This will ensure an overall type I error rate of $P<0.05$.

DEFINING ACCELERATION

Another point discussed at the January 1979 meeting was the meaning of acceleration. Lagakos and Mosteller (*10*) pointed out that the concept of acceleration was not clearly defined in operational terms. Both FDA (*22*) and WHO (*23*), in reports on carcinogenesis testing, state that an earlier occurrence of tumors in treated animals than in controls indicates a carcinogenic effect even if the overall tumor incidences are the same. However, their reports leave open the question of what the measure of shortening should be.

In the simplest situations, any reasonable measure of shortening of time to tumor (or time to death), e.g., the mean or median, might suffice. In other situations in which, for example, trade-offs between a few more long lives or many more short lives may be at issue, the choice of a measure is important and can affect the conclusions reached. For example, Lagakos and Mosteller (*18*) illustrated the arbitrariness of the definition of acceleration by applying eight different metrics to the data in the first mouse experiment. We got *P*-values ranging from 0.003 to 0.316. This clearly indicates how the choice of a metric can affect resulting *P*-values. Adjusting for the fact that the hypothesis was suggested by the data would increase these *P*-values and thus weaken the statistical significance of the acceleration hypothesis. (Because eight measures are under discussion, consideration should also be given to the problem of multiplicity. The 0.003 is probably too small and the 0.316 too large, just because they are the smallest and largest of eight numbers measuring similar things.)

More generally, suppose that a control and an exposed group have identical lifetime incidence rates. If the survivorship function of time to RE death for the exposed group is always smaller than that for the control group, it is reasonable to speak of an acceleration in the time to RE death in the exposed group. However, if these distributions cross, we need to define which group, if either, has accelerated or whether the word gives a poor description of the comparison. Comparing two functions by a single word, metric, or number imposes a weight function or payoff function on the curves. Thus it is not so easy or automatic to compare the shortening of time as the word "acceleration" might suggest. The problem of defining carcinogenic effects is even more complicated than this, because incidence rates as well as times to death may differ between control and exposed animals.

INTERIM SACRIFICES

In the early stages of the first mouse experiment, Allied Chemical Company notified FDA that some unexpectedly early RE deaths occurred among the exposed mice as compared with no RE deaths in the control mice. The FDA suggested that a number of mice be killed immediately. This sacrifice was done in such a way that 30 animals were left alive in each

sex-dose group. The 2 animals with the largest serial numbers in each cage were killed unless an animal had already died, and then only 1 was killed. In all, 142 mice (36% of the original sample) were killed at week 42, and among these no RE tumors were found.

Although post-hoc considerations of the interim sacrifice decision have no bearing on the evaluation of Red 40, they may be of use in future experiments in which there are early indications of possible exposure effects.

Two potentially useful things could have been learned from such an interim sacrifice. The first would have been to discover whether Red 40 accelerated the growth of preexisting RE tumors. If acceleration occurred, FDA could have taken rapid measures to curb the human consumption of Red 40. Second, if a sufficient number of RE tumors had been found in the exposed mice as compared with none (or few) in the controls, a statistically significant carcinogenic effect might have been observed without over a year of waiting until the close of the experiment. As pointed out by the Working Group (*1, 2*), a limitation of the interim sacrifice is that is could not distinguish whether Red 40, if carcinogenic, increases overall RE incidence or only shortens the latency period of RE tumors.

The main disadvantage of the interim sacrifice was that it substantially reduced the available information for analyzing incidence and time to death. Though not a major consideration, it also complicated these analyses from a technical point of view.

The decision to undertake an interim sacrifice was criticized by Gross (*3, 4*). He argued that RE tumors are well known to be uniformly fatal and that their course (the period from tumor onset until death) is very brief; hence finding any (or many) RE tumors in advance in either the control or exposed mice was unlikely. The second report of the Working Group (*2*) mentioned that the course of RE tumors was short. The Working Group, formed only after the interim sacrifice, took no position on its merit. The decision to sacrifice was later defended by Greenberg (*9*).

If the course of RE tumors is very rapid near the time of sacrifice, there is little to gain by interim sacrifice. The results of the terminal sacrifice suggest, however, that the course of these tumors, at least in older mice, may be more than a matter of several days. The course may vary with age, being more rapid for younger mice. If so, the use of time to death as a surrogate for time to tumor onset needs to be reexamined. In "Appendix 1," we present a method for judging the lethality of a tumor and apply this to the mouse data.

Ad hoc arguments could be made on how and when an interim sacrifice should be done. As far as we know, this question has not been quantitatively or theoretically addressed. It involves modeling in a problem for which the appropriate model is exactly the issue. The solution involves the trade-off between the information gained from the sacrifice and the information lost by preventing the mice from dying naturally. It seems to us that this is an area in which challenging theoretical research of practical importance can be profitably pursued. The magnitude of the task, as judged from other decision problems with sequential features, appears to us to be substantial.

NON-RE DEATHS

It is common in analyzing animal carcinogenesis experiments to focus attention on one type or group of diseases and give little direct attention to others. For example, the Red 40 analyses focused on tumors of the RE system. This approach was partly due to the stance that the substance being tested, if a carcinogen, will likely affect certain indicator tumors. Other diseases may be of interest but occur so rarely that their formal analysis is not feasible. This stance is also reinforced by the guidelines for carcinogenesis testing, which focus on the formation of tumors and, by so doing, might be interpreted as suggesting that diseases of other types are not relevant.

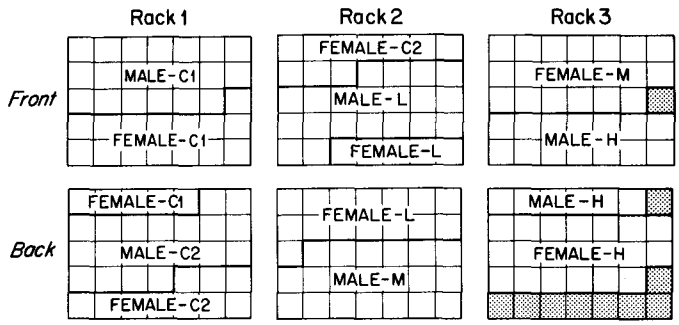
Sometimes, differences between control and exposed mice appear in terms of other diseases or their prevention, and then their interpretation is not always clear. For example, in the first mouse experiment, the male controls tended to have earlier times to non-RE deaths than the exposed males (*10*), and this did not appear to be due to a "competing risk" phenomenon or, as far as we know, an infectious disease or any other easily explainable cause. Thus Red 40 may have a favorable effect on time to non-RE deaths. Of course, the observed difference may have been due entirely to statistical variation.

Although the law is primarily oriented to carcinogenicity, it is good practice to review systematically the effect of exposure on all times to death without regard to type. A mistake sometimes made in human experimentation is to note and act on the basis of an increase in a death rate from a specific disease associated with a treatment, without reviewing total performance.

CAGE AND LITTER EFFECTS

In his letter to the consultants, Commissioner Kennedy indicated that their opinions and analyses need not be limited to his seven specific questions. Lagakos and Mosteller requested information about the placement of mice in cages and about cage layout. Shortly before the January 1979 meeting, we received this cage information for the second mouse experiment but not for the first.

The mice in the second experiment were housed in three large racks, each having a front and back section and each section consisting of five rows of cages. There were seven cages per row and 5 mice per cage. The mice were assigned to the cages systematically, beginning with the first control group males in the top of the front of rack 1 and ending with the high dose females in the bottom of the back of rack 3 (text-fig. 4).



TEXT-FIGURE 4.—Second mouse experiment. Assignment of mice to cages. C1 = 1st control group; C2 = 2d control group; L = low dose; M = medium dose; H = high dose. *Shaded boxes* represent empty cages.

These cage positions were maintained throughout the experiment.

At the January meeting, we presented a preliminary analysis (*14*) that disclosed a strong correlation between cage row and RE death rates, which varied from 17% (bottom row) to 32% (top row) (table 2). We could not explain this strong association by sex, dosage group, or rack column or position. A subsequent analysis (*18*) also indicated that cage position (front vs. back) might be correlated with non-RE mortality and that position was correlated with time to non-RE death.

These findings further complicated the analysis of the second experiment. Because of the systematic way of assigning animals to cages, the cage effects, if not accounted for, could mask a real difference between the control and exposed groups or could appear to yield a difference when none existed. Furthermore, the findings affected the interpretation of the first experiment because assignments of mice to cages had also been systematic and no cage analyses had been done.[9]

At the January meeting, Greenberg (*15*) reported evidence for a litter effect in the second experiment. This effect would matter to the analysis because, by design, all mice from a given litter received the same treatment and littermates of the same sex were housed in the same or contiguous cages.

If a litter effect exists, variability among animals within a litter may be less than that among animals from different litters. In extreme situations, animals in the same litter always have the same outcome, and then two such animals provide no more information than one. More generally, the effective sample size lies somewhere between the number of litters and the total number of mice.

The result is that in analyses that make no distinction among animals on the basis of litters, the *P*-values for exposure effects tend to be smaller than they should be, i.e., they give overly significant results. To put it another way, these analyses respond in a way that suggests that the experiment was more precise than it actually was. As far as we know, statistical methods for litter effects in time-to-death analyses in which all littermates receive the same treatment have not yet been developed.

SECOND MOUSE EXPERIMENT

Table 3 summarizes the numbers of RE and non-RE tumors for the second mouse experiment. Note the large variation in RE tumor rates relative to the first mouse experiment. Among the exposed groups no evidence was found of a typical dose-response relationship. Another peculiar aspect was the difference in RE death rates between the 2 control groups. Among 100 male mice in the first control group, 25 had RE tumors as compared with only 10 among 100 male mice in the second control group. For female mice, 33 of 100 in the first control group had RE tumors as compared with 25 of 99 in the second control group. (With an exact test for two 2×2 tables (*24*), this difference is statistically significant at the $P = 0.008$ level.) These differences are surprising because as far as we know, the 2 control groups were generated and handled in the same way.

One explanation for this control group difference in RE death rates could be a variation in cage positioning. The 2 control groups were well balanced across rows, but as the Working Group pointed out (*25*), the groups were highly confounded with cage position: front versus back (table 4). Overall, ⅞ of the mice in the first control group compared to only ¼ of the mice in the second control group occupied the fronts of racks. The situation was even more severe among male mice, because all those in the first control group were in the front of the first rack and all those in the second control group were in the back of this rack. Thus the imbalances between the control groups with respect to cage position could have led to the observed difference in RE death rates. The confounding of these two factors makes this explanation practically impossible to verify.

In our own analysis of these data (*26*), we were unable to explain fully the observed difference in control groups in terms of sex, cage row, and cage

[9] Mr. Gregg Dinse, who assisted us in our analyses, obtained the cage data from the first mouse expt from FDA in March 1980. His preliminary analyses indicate a strong row effect on RE incidence, but in the opposite direction to that we indicated in table 2 for the second mouse expt.

TABLE 2.—*Second mouse experiment: RE tumor incidence by cage row*

Row	No. of mice	No. of RE deaths	Percent RE deaths
1, top	201[a]	64	32
2	209[b]	50	24
3	205	37	18
4	205	37	18
5, bottom	175	30	17

[a] Excludes 4 mice with unknown RE status.
[b] Excludes 1 mouse with unknown RE status.

TABLE 3.—*Second mouse experiment: Types of death by sex and dose group*

Sex	Category	No. of mice in various groups: First control	Second control	Low dose	Medium dose	High dose
Male	No. of mice	100	100	99	100	99
	Natural deaths	63	59	61	66	71
	RE tumors present	19	10	14	8	13
	No RE tumors	44	49	47	58	58
	Terminal sacrifice	37	41	38	34	28
	RE tumors present	6	0	6	1	4
	No RE tumors	31	41	32	33	24
	Total with RE tumors	25	10	20	9	17
Female	No. of mice	100	99	99	99	100
	Natural deaths	59	48	65	56	56
	RE tumors present	24	17	28	24	19
	No RE tumors	35	31	37	32	37
	Terminal sacrifice	41	51	34	43	44
	RE tumors present	9	8	4	2	3
	No RE tumors	32	43	30	41	41
	Total with RE tumors	33	25	32	26	22
	Total with RE tumors, both sexes	58	35	52	35	39

TABLE 4.—*Second mouse experiment: Distribution of control mice by sex, control group, and rack position*

Control group	No. of mice in front of rack			No. of mice in back of rack		
	Males	Females	Total	Males	Females	Total
1	100	75	175	0	25	25
2	0	50	50	100	50	150

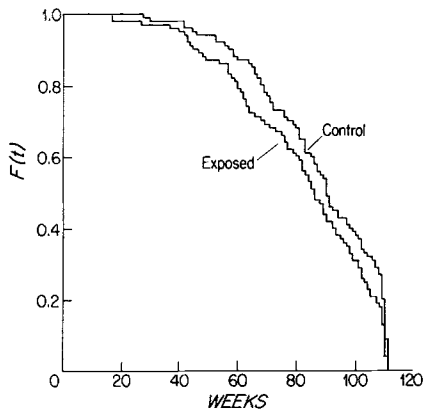
position. Because we have no other explanation for the difference in control groups, the conclusiveness of the second experiment is diminished.

ASSESSING ACCELERATION IN THE SECOND MOUSE EXPERIMENT

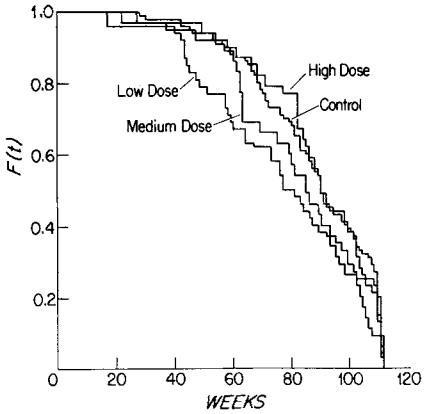
At the January 1979 meeting, discussions of acceleration focused on the first mouse experiment, for which we had no detailed cage information. The final results from the second experiment were not yet available, and preliminary analyses by the Working Group did not suggest any indications of acceleration.

In April 1979, we (*18*) reported to the Working Group our analyses of the final data from the second experiment and indicated the suggestion of an acceleration effect (text-fig. 5). We noted that the magnitude of the effect was smaller than that observed in the first experiment and was statistically significant only when the three exposure levels were pooled. For these exposed groups, time to RE death was observed to be directly related to dose, i.e., to have an "inverse" dose-response relationship (*26*) (text-fig. 6). This inverse effect lessened belief in the acceleration hypothesis. Text-figure 6 also shows the similarity between the pooled controls and the high-dose group.

By the June 1979 meeting, the Working Group had also analyzed the final data from the second experiment and reviewed the question of acceleration, including the inverse dose-response relationship (*25*). Our



TEXT-FIGURE 5.—Second mouse experiment. Conditional distributions, F_C and F_E, of time to death for mice that died with RE tumors by wk 111.



TEXT-FIGURE 6.—Second mouse experiment. Conditional distributions, F_C and F_E, of time to death for each group of mice that died with RE tumors by wk 111.

interpretation of the discussion is that most scientists present regarded an inverse dose-response effect as unlikely and hence believed that the observed result was a statistical fluctuation, a confounding effect with some other explanatory variable, or an indication of a bias in the experimental design.

A POSSIBLE SELECTIVITY PROBLEM

A basic premise in both mouse experiments is that the control and exposed groups are, apart from any treatment and cage effects, comparable in all respects. The multigenerational aspect of the experimental design may have compromised this premise (*18*).

To generate the mice used in the experiments, parental males and females were paired and assigned to the control, low-, medium-, or high-dose group. These parents were then fed their assigned doses of Red 40 before and during mating, and the mothers were continued on the same doses throughout gestation and weaning. To be eligible for the second experiment, a litter had to have at least 3 pups of each sex. From each such litter, 3 mice of each sex were randomly selected for inclusion. Thus all mice within a litter received the same exposure level in utero, and those selected for inclusion in the experiment continued throughout their lifetimes on the same exposure level as the parents. The rationale behind this two-generational design was to provide every possible opportunity for the test chemical to manifest its carcinogenic effect. This design introduces the substance in utero, when the organism may be especially vulnerable.

The possible selectivity bias arises because only the mice surviving until birth had a chance to be in the experiment. The bias might be further aggravated because only litters with at least 3 pups of each sex were used. If exposure to Red 40 affected the birth process, the litters qualifying for inclusion in different treatment groups might not be comparable. For example, suppose that Red 40 let only the hardier mice survive until birth. Then the exposed mice included in the experiment could be stronger than the controls; hence the 2 groups could not be legitimately compared.

A selectivity bias is difficult to detect and, if present, cannot usually be accounted for in the analysis. One possible indication of its presence would be a difference in the distributions of the litter sizes of the exposed and control mice. However, these data did not seem to be available. Even if a difference in distributions of litter size occurred, we still would not know whether the exposed mice were more or less resistant to RE tumors or other diseases than were the untreated mice.

AMALGAMATING INFORMATION AND MAKING DECISIONS

What started as a seemingly simple question—"Is Red 40 a carcinogen?"—unfolded into a number of complex issues, the overall message of which was not at all obvious even after the second mouse experiment. After conducting several experiments, gathering expert opinions from a working group and from outside sources, holding public meetings, and collecting numerous additional opinions, FDA was faced with making a decision about Red 40.

The first step in this process is an overall assessment of the experiments, including any further analyses that seem warranted. As we have seen, the assessment is complicated by questions of hypotheses suggested by the data, choice of test statistics, prior beliefs about and the definition of an acceleration effect, cage and litter effects, confounding, inverse dose-response effects, and a possible selectivity problem. The task is to surmise what the data are saying and how strongly they seem to be saying it. One must also keep in mind the closely related question of power: Were the experiments of sufficient size and sufficiently well designed that important differences had a high chance of being detected and were observed differences large enough to be considered important?

The next step is to decide if there is enough evidence to reach a conclusive decision about the safety of Red 40, and if not, whether additional experimentation is warranted. This step usually involves consideration of human risks and benefits, costs, timing, substitute products, and priorities, and in a way that is difficult to quantify.

DISCUSSION AND RECOMMENDATIONS

The purpose of this paper is to present a case study of the events and issues surrounding recent experiments on Red 40. It is not our task or intention to evaluate Red 40 itself but to describe how scientific investigations can develop from regulatory situations and how problems can and do arise. Because of these complications, the Red 40 story is a useful example to persons concerned with scientific experimentation. The efforts of the FDA in recognizing the need for a Working Group, suggesting a second experiment, soliciting outside opinions, financing analyses, holding public meetings, and maintaining a continuing interest in developments have been constructive, and they add up to a considerable contribution in helping to solve some of the questions raised by the complications that have appeared. Also, the concerns expressed by Dr. Gross have contributed to these valuable consequences.

For students of statistics, the development of the acceleration hypothesis provides a good example of the difficulty of evaluating hypotheses suggested in light of the data. The paper illustrates, in a practical situation, the difficulties in analyzing competing risk data, namely, the choice of formulations, the choice of statistical tests, and the grouping of causes of death into broader categories (e.g., RE vs. non-RE). The complications arising from cage and litter effects and from confounding cage position and control group underscore the importance of experimental design. Comparing control and exposed groups raises ques-

tions when duplicate control groups exhibit substantial differences.

For students of health policy and administration, this discussion shows how complex the results of a scientific study can be. Although we all like simple and clear-cut conclusions based on specific methods and unequivocal data, real investigations are rarely so straightforward. Outside committees and public meetings invariably create time delays, extra costs, and more interpretations to collate. Nevertheless, this route allows the opinions of persons outside the system to be heard, thus broadening the base of the discussion.

For the consumer, there seem to be two messages. First, in a one-shot experiment, surprises often lurk and many things can and often do go amiss; second, experimental results are often not clear and hence their ultimate interpretation depends on one's own prior beliefs and values. Thus two rational people with differing value systems could reach different decisions from the same data.

For the FDA, National Cancer Institute, National Center for Toxicological Research, and other agencies and scientific groups, a number of points arise regarding animal experiments. One is the role of multigenerational designs for testing carcinogenicity. Their advantage over designs using within-litter randomizations is the increased sensitivity from in utero exposure to the test substance. As we have seen, however, multigenerational designs raise questions about litter effects and selectivity. One way to avoid the former in a multigenerational design is to use only a single mouse from each litter (*27*). The possible selectivity bias is more difficult to avoid, measure, or account for. We think that these issues should be studied to determine which type of design is better for various types of potential carcinogens.

A second issue is whether acceleration is sufficiently likely that it should be checked for routinely. The consultants recommended (*17*) that past investigations be reviewed for acceleration effects to see whether they were rare. The point was made that if they had not been looked for in the past analyses, their frequency was hard to guess. As indicated earlier (*see* "Choice of Models and of Statistical Tests"), the usual statistical tests for time to death are not especially oriented toward an acceleration effect and hence not particularly appropriate for testing it. We recommend that a decision be made whether routine checking for acceleration is of value and, if so, how acceleration should be defined in operational terms.

Another analytical issue is the interpretation of differences between control and exposed groups with respect to time to non-RE death, and the role this should play in the overall evaluation of the data. We believe that this end point plus time to death from all causes should be routinely examined for possible benefits or damages from exposure.

Findings about non-RE deaths and findings for nonprimary hypotheses related to the investigation raise the question of the place of exploratory data analyses in the regulatory context. In the consultants' view (*17*), in addition to the standard analyses and related tests, the regulatory reports should have a place for exploratory analyses that do not have the same status as the more standard analyses. The exploratory analyses serve to alert the scientific and regulatory community and others to matters that might need to be studied or attended to in later work or in other studies of a related character. They do not necessarily bear strongly on the decisions in the particular study. These studies are, of course, being done in the light of the data, so their results will need confirming in future data. That is the spirit of the enterprise. Exploratory analyses do represent a way of retrieving and retaining findings that might otherwise be lost from the specific studies. It is especially helpful when such studies are done by the original investigators because they may know much more about the actual experimental circumstances and decisions than other analysts can discover.

The associations between rack row or position and outcome indicate the need to consider cage layout in the analysis of laboratory experiments. We believe that cage layout variables should always be included in the analysis because this increases the sensitivity of treatment comparisons and can guide the investigator to better control of environmental factors in the laboratory. Cage layout should also be considered in the design of experiments. One approach is to arrange cages in a way that "balances" treatment groups with respect to rows, columns, positions, and racks. This arrangement would free estimates of exposure effects from biases due to cage layout. "Appendix 2" gives two specific designs that could have been used in studies similar to the two mouse experiments. These designs were constructed by making variations on suggestions by Dr. Agnes Herzberg (personal communication). Other designs with particular merits could also be constructed. Alternatively, cages can be assigned to the dosage groups in a totally random way. We prefer a balanced design to random allocation because it ensures that factors of interest will not be confounded, it leads to slightly more sensitive tests, and it is easier to implement. For either approach, the results should be analyzed in a way (*28, 29*) that accounts for possible sex and cage layout effects.

A different approach is the rotation of cages during the experiment so that, for example, each treatment group occupies each row of a rack for a comparable period of time. This approach is not as desirable as balanced or randomized cage allocations. Unexpected exposure to a toxic substance on a given day or other environmental factors whose effects may vary with age or duration can still lead to biases that cannot easily be accounted for in the analysis. If there were some other special merit to rotation, both it and balanced allocation could probably be used.

Finally, the question of whether, when, and how to interrupt an ongoing experiment for the unscheduled sacrifice of some of the animals should be studied.

After such studies are made, it would be useful if FDA and other agencies issued guidelines on how to proceed when unexpected early results appear.

APPENDIX 1: ASSESSING TUMOR LETHALITY

As we stated earlier (*see* "Interim Sacrifices"), assessment of tumor lethality might have a bearing on the decision to make an unscheduled interim sacrifice. More generally, the lethality or nonlethality of a tumor is important in determining the type of statistical analysis that should be used. For example, time-to-death analyses are appropriate for tumors with a short course because then time to death is a surrogate for time to tumor. It is generally accepted, however, that such analyses are not appropriate for nonlethal tumors.

Suppose that at a given time t either all or a random sample of animals are killed and examined for the presence of a particular irreversible tumor. For example, in a lifetime feeding experiment, t might be the time of the terminal sacrifice. Let N denote the number of animals killed, and suppose that n_1 of these are found to have tumors and that $n_2=N-n_1$ do not have the particular tumor. The estimated tumor prevalence rate at time t is therefore n_1/N.

Now suppose (t_1,t_2) denotes a time interval near t for which the prevalence at t_1 is about the same as at t. The interval should be short enough so that an animal without a tumor at time t_1 is not likely to develop a tumor and die before t_2. Let M be the number of animals alive at time t_1, and suppose that m_1 (m_2) of these die naturally by time t_2 with (without) tumors.

The estimated prevalences of animals with or without tumors at time t_1 are $(n_1/N)M$ and $(n_2/N)M$, respectively. The estimated conditional probability that an animal alive at time t_1 will die before time t_2 is therefore $l_1=m_1/[(n_1/N)M]$ for animals with tumors and $l_2=m_2/[(n_2/N)M]$ for those without. These lethality statistics, l_1 and l_2, and the corresponding parameters, λ_1 and λ_2, that they estimate are similar in spirit to those considered by Turnbull and Mitchell (*30*) and others. When $\lambda_1=\lambda_2$, the presence of a tumor at time t_1 is not associated with a higher probability of death by time t_2 than in the absence of the tumor; hence the tumor can be called "nonlethal." When $\lambda_1>\lambda_2$, the tumor is "comparatively lethal" in the sense that its presence at time t_1 increases the risk of death by time t_2. In the extreme situation $\gamma_1=1$, all animals alive with tumors at time t_1 are certain to die by time t_2. For ease of exposition, we call such tumors "rapidly lethal."

Suppose that we want to assess the hypothesis H_{RL}:$\lambda_1=1$, that a tumor is rapidly lethal. Then because $l_1=1$ if and only if $n_1/N=m_1/M$, H_{RL} can be assessed by testing for homogeneity in the 2×2 contingency with entries:

m_1	$M-m_1$	M
n_1	n_2	N

When $\lambda_1<1$, m_1/M will tend to be less than n_1/N.

Alternatively, suppose we wanted to assess the hypothesis H_{NL}:$\lambda_1=\lambda_2$, that the tumor is nonlethal. Then because $l_1=l_2$ if and only if $m_1/(m_1+m_2)=n_1/N$, H_{NL} can be evaluated by testing for homogeneity in the 2×2 contingency table with entries:

m_1	m_2	$m=m_1+m_2$
n_1	n_2	N

If $m_1/m>n_1/N$, we have evidence that the tumor is comparatively lethal. Expressed another way, the tumor is nonlethal if the proportion of animals with tumors is approximately equal among killed animals and among those that died naturally.

In many practical situations, we deal with animal categories or strata (e.g., sex) with different tumor rates. In these circumstances, H_{RL} and H_{NL} can be assessed by forming a separate 2×2 contingency table for each of the K-strata. The K-tables can then be analyzed with methods for K 2×2 tables [*see*, for example, (*24*)]. When several covariates are associated with tumor rate, the techniques described by Breslow (*31*) could be applied.

To illustrate the methods, we consider the data near and at the end of the two mouse experiments. Recall that RE tumors were generally thought to be rapidly lethal, so an assessment of H_{RL} is of particular interest. Table 5 gives the results from the second mouse experiment, based on natural mortality during weeks 105-108 and the terminal sacrifice beginning in week 109. For assessing H_{RL} in females, the 2×2 tables obtained by considering the control and exposed mice separately are:

Controls			Exposed		
4 (4%)	96	100	4 (3%)	131	135
17 (18%)	75	92	9 (7%)	112	121

The Fisher-Irwin exact test gives $P=0.002$ (controls) and $P=0.15$ (exposed), and a combined test (*23*) gives $P=0.0005$.

For males, the tables are:

Controls			Exposed		
1 (1%)	88	89	2 (2%)	119	121
6 (7%)	72	78	11 (11%)	89	100

with significance levels $P=0.05$ (controls), $P=0.004$ (exposed), and $P=0.0003$ (combined). Thus for both sexes the data suggest that in the second experiment, RE tumors present at week 105 will not necessarily cause death within 4 weeks.

Let us now consider the results from the first mouse experiment, based on natural mortality during weeks

TABLE 5.—*Numbers and types of deaths near and at the end of the second mouse experiment*

Mouse group	No. of mice alive by wk 105	No. of mice that died in wk 105–108			No. of mice killed in wk 109–111			Estimated lethality	
		Total	With RE tumors	Without RE tumors	Total	With RE tumors	Without RE tumors	With RE tumors, l_1	Without RE tumors, l_2
Controls									
Males	89	11	1	10	78	6	72	0.15	0.12
Females	100	8	4	4	92	17	75	0.22	0.05
Exposed									
Males	121	21	2	19	100	11	89	0.15	0.18
Females	135	14	4	10	121	9	112	0.40	0.08

TABLE 6.—*Numbers and types of deaths near and at the end of the first mouse experiment*

Mouse group	No. of mice alive by wk 100	No. of mice that died in wk 100–103			No. of mice killed at wk 104			Estimated lethality	
		Total	With RE tumors	Without RE tumors	Total	With RE tumors	Without RE tumors	With RE tumors, l_1	Without RE tumors, l_2
Controls									
Males	15	3	0	3	12	0	12	NE[a]	0.20
Females	22	4	2	2	18	2	16	0.82	0.10
Exposed									
Males	57	7	0	7	50	1	49	0	0.12
Females	60	6	2	4	54	8	46	0.22	0.06

[a] Not estimable.

100-103 and the terminal sacrifice in week 104 (table 6). For female mice, the 2×2 tables for assessing H_{RL} are:

Controls

2 (9%)	20	22
2 (11%)	16	18

Exposed

2 (3%)	58	60
8 (15%)	46	54

with corresponding significance levels $P=1.0$ (controls), $P=0.04$ (exposed), and $P=0.09$ (combined).

For male mice the corresponding tables are:

Controls

0 (0%)	15	15
0 (0%)	12	12

Exposed

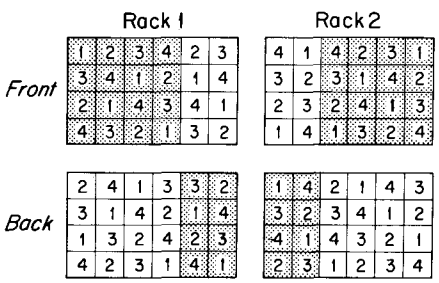
0 (0%)	57	57
1 (2%)	49	50

Note that the estimated RE prevalence rates are extremely small. This means that there is virtually no power to assess the probability of death by week 103, given an RE tumor at week 100, because the data indicate that there are almost no mice with RE tumors at week 100. Expressed statistically, the RE percentage in the top row cannot be significantly less than that in the bottom row when the latter is zero or nearly zero. Thus for the first experiment, there is some evidence against H_{RL} in the exposed females, but virtually no information to assess H_{RL} in males.

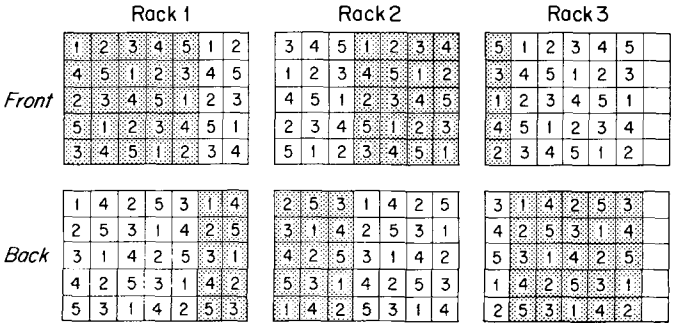
APPENDIX 2: CAGE LAYOUT DESIGN

Text-figure 7 gives an example of a cage layout design that could be used in an experiment with about the same number of mice and the same dosage and control plan as the first mouse experiment. (We assume that there are 2 side-by-side racks having front and back sections and that each section can accommodate 4 rows of cages and 6 cages per row.) The numbers 1, 2, 3, and 4 on the cages refer to the control and 3 exposed groups in some random order. With 4 mice per cage, the design leads to 48 mice per sex-dose group.

The front and back sections each consist of three 4×4 Latin squares placed end to end. This means that each dose level appears once in each column and each row of each 4×4 Latin square. Such a design should



TEXT-FIGURE 7.—Possible cage layout design for a study similar to the first mouse experiment. *Shaded area* = males; 1, 2, 3, 4 = random assignment of control group and mice given low, medium, or high doses, respectively; 4 mice/cage.



TEXT-FIGURE 8.—Possible cage layout design for a study similar to the second mouse experiment. *Shaded area*=males; 1, 2, 3, 4, 5= random assignment of 2 control groups and mice given low, medium, or high doses, respectively; 5 mice/cage.

prevent any treatment-comparison biases due to environmental row, column, or position effects.

The cages are also balanced to avoid biases from regional effects: Each of the four 2×2 "quadrants" in each 4×4 Latin square contains all 4 dose groups. There is also front-to-back balance in the sense that for each front-back pair of Latin squares, the 16 pairs of cages contain all 16 combinations of pairs of dose groups.

Text-figure 8 gives a cage layout design that might be used in an investigation with 5 groups and sample sizes like those of the second mouse experiment. (Here we assume that there are three side-by-side racks having front and back sections and that each section holds 5 rows of cages and up to 7 cages per row.) The numbers 1, 2, 3, 4, and 5 refer to the 5 groups in some random order. With 5 mice per cage, this design leads to 100 mice per sex-dose group.

The front and back sections each consist of four 5×5 Latin squares. Within each Latin square, each group appears once in each row, column, and diagonal. The four cages surrounding a given cage (i.e., above, below, and to the sides) represent the 4 other groups. Furthermore, within each front-back pair of Latin squares, the 25 pairs of cages contain all 25 combinations of pairs of dose groups. Thus this design will tend to minimize biases due to row, column, position, and regional effects.

It should be understood that these designs are suggestions and not necessarily appropriate for all situations.

REFERENCES

An asterisk (*) before the reference number indicates that the information is available from the Freedom of Information Office, Food and Drug Administration, Rockville, Maryland 20857.

* (1) Working Group. Interim report on FD&C Red No. 40. Jan. 19, 1977.

* (2) ———. Second interim report on FD&C Red No. 40. Drafted Feb. 1978; released April 1978.

* (3) GROSS MA. Memorandum to Richard R. Bates. Carcinogenicity of Red No. 40. Feb. 8, 1978.

* (4) ———. Memorandum to Thomas P. Grumbly. Carcinogenicity testing of FD&C Red No. 40. June 13, 1978.

* (5) GART JJ. Memorandum to Albert C. Kolbye. March 2, 1978.

* (6) ———. Memorandum to Albert C. Kolbye. July 11, 1978.

* (7) KENNEDY D. Letter to Frederick Mosteller. July 25, 1978.

* (8) CORNFIELD J. Letter to Donald Kennedy. Oct. 1, 1978.

* (9) GREENBERG BG. Letter to Robert L. Elder. Sept. 25, 1978.

* (10) LAGAKOS S, MOSTELLER F. Memorandum to Donald Kennedy. Oct. 11, 1978.

* (11) ARMITAGE P. Written comments presented on Jan. 17, 1979. In: Transcript of the ad hoc meeting on statistical procedures to analyze bioassay data for tumor acceleration (FD&C Red No. 40), Jan. 17-18, 1979, Washington, D.C., pp. 97-100.

* (12) BRESLOW N. Written comments presented on Jan. 17, 1979. In: Transcript of the ad hoc meeting on statistical procedures to analyze bioassay data for tumor acceleration (FD&C Red No. 40), Jan. 17-18, 1979, Washington, D.C., pp. 100-105.

* (13) COX DR. Written comments presented on Jan. 17, 1979. In: Transcript of the ad hoc meeting on statistical procedures to analyze bioassay data for tumor acceleration (FD&C Red No. 40), Jan. 17-18, 1979, Washington, D.C., pp. 105-109.

* (14) LAGAKOS S, MOSTELLER F. Presentation on Jan. 17, 1979. In: Transcript of the ad hoc meeting on statistical procedures to analyze bioassay data for tumor acceleration (FD&C Red No. 40), Jan. 17-18, 1979, Washington, D.C., pp. 7-30.

* (15) GREENBERG BG. Presentation on Jan. 17, 1979. In: Transcript of the ad hoc meeting on statistical procedures to analyze bioassay data for tumor acceleration (FD&C Red No. 40), Jan. 17-18, 1979, Washington, D.C., pp. 38-51.

* (16) CROWLEY J, CORNFIELD J, FLEMING TR, GREENBERG BG, LAGAKOS SW, MOSTELLER F. Memorandum to Albert C. Kolbye. Feb. 23, 1979.

* (17) ———. Letter to Donald Kennedy. Feb. 15, 1979.

* (18) LAGAKOS S, MOSTELLER F. Mixed population analysis of mouse experiments on FD&C Red No. 40. April 18, 1979.

(19) PRENTICE RL, KALBFLEISCH JD, PETERSON AV JR, FLOURNOY N, FAREWELL VT, BRESLOW NE. The analysis of failure times in the presence of competing risks. Biometrics 1978; 34:541-554.

* (20) FLEMING TR. Presentation on Jan. 17, 1979. In: Transcript of the ad hoc meeting on statistical procedures to analyze bioassay data for tumor acceleration (FD&C Red No. 40), Jan. 17-18, 1979, Washington, D.C., pp. 52-68.

(21) KAPLAN EL, MEIER P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; 53:457-481.

(22) FDA Advisory Committee on Protocols for Safety Evaluation. Panel on carcinogenesis report on cancer testing in the safety evaluation of food additives and pesticides. Toxicol Appl Pharmacol 1971; 20:419-438.

(23) WHO Scientific Group. Principles for the testing and evaluation of drugs for carcinogenicity. WHO Tech Rep Ser 1969; 426:1-26.

(24) ZELEN M. The analysis of several 2×2 contingency tables. Biometrika 1971; 58:129-137.

* (25) Remarks of members of Working Group. In: Transcript of the ad hoc meeting for Red 40, June 11, 1979, Washington, D.C.

* (26) LAGAKOS S, MOSTELLER F. Opening presentation. In: Transcript of the ad hoc meeting for Red 40, June 11, 1979, Washington, D.C.

(27) FOX JG, THIBERT P, ARNOLD DL, KREWSKI DR, GRICE HC. Toxicology studies. II. The laboratory animal. Food Cosmet Toxicol 1979; 17:661-675.

(28) COX DR. Regression models and life tables (with discussion). J R Stat Soc B 1972; 34:187-220.

(29) PETO R, PIKE MC, ARMITAGE P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. Br J Cancer 1977; 35: 1-39.

(30) TURNBULL BW, MITCHELL TJ. Exploratory analysis of disease prevalence data from survival sacrifice experiments. Biometrics 1978; 34:555-570.

(31) BRESLOW N. Regression analysis of the log odds ratio. Biometrics 1976; 32:409-416.