Guidance for Industry

Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit comments to Dockets Management Branch (HFA-305), Food and Drug Administration, 12420 Parklawn Dr., rm. 1-23, Rockville, MD 20857. All comments should be identified with the docket number listed in the notice of availability.

For questions regarding this draft document contact (CDER) Karl K. Lin, Ph.D., 301-827-3093, e-mail link @cder.fda.gov.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)

May 2001

Pharm/Tox

G:\815dft.doc
04/23/01

Guidance for Industry

Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals

Additional copies are available from:

Office of Training and Communications
Division of Communications Management
Drug Information Branch, HFD-210
Center for Drug Evaluation and Research (CDER)
5600 Fishers Lane
Rockville, MD 20857
(Tel) 301-827-4573
(Internet) http://www.fda.gov/cder/guidance/index.htm

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)

May 2001

Pharm/Tox

TABLE OF CONTENTS

I. INTRODUCTION 1

II. BACKGROUND 1

III. VALIDITY OF THE DESIGN 2

IV. METHODS OF STATISTICAL ANALYSIS 4

A. AN OVERVIEW OF COMPLEXITIES OF STATISTICAL ANALYSIS OF TUMOR DATA 4
B. ADJUSTMENT OF TUMOR RATES FOR INTERCURRENT MORTALITY 6
C. STATISTICAL ANALYSIS OF TUMOR DATA WITH INFORMATION ABOUT CAUSE OF DEATH, TUMOR LETHALITY, BUT WITHOUT MULTIPLE SACRIFICES 8
1. Role of the Tumor in Animal's Death (Contexts of Observation of Tumor Types) 8
2. Statistical Analysis of Incidental Tumors 9
3. Statistical Analysis of Fatal Tumors 15
4. Statistical Analysis of Tumors Observed in Both Incidental and Fatal Contexts 16
5. Statistical Analysis of Mortality Independent Tumors 16
6. Exact Methods 17
D. STATISTICAL ANALYSIS OF DATA WITHOUT INFORMATION ABOUT CAUSE OF DEATH AND WITHOUT MULTIPLE SACRIFICES 21
E. STATISTICAL ANALYSIS OF DATA FROM STUDIES WITH DUAL CONTROLS 25

V. INTERPRETATION OF STUDY RESULTS 26

A. ADJUSTMENT FOR THE EFFECT OF MULTIPLE TESTS (CONTROL OVER FALSE POSITIVE ERROR) 27
B. CONTROL OVER FALSE NEGATIVE ERROR 30
C. THE USE OF HISTORICAL CONTROL DATA 30
D. EVALUATION OF VALIDITY OF DESIGNS OF NEGATIVE STUDIES 32

VI. PRESENTATION OF RESULTS AND DATA SUBMISSION 32

REFERENCES 38

Guidance for Industry[1]
Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals

This draft guidance, when finalized, will represent the Food and Drug Administration's current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. An alternative approach may be used if such approach satisfies the requirements of the applicable statutes and regulations.

I. INTRODUCTION

This document is intended to provide guidance to sponsors on the design of animal carcinogenicity experiments, methods of statistical analysis of tumor data, interpretation of study results, presentation of data and results in reports, and the submission of tumor data to FDA statistical reviewers at the Food and Drug Administration (FDA). A brief background description of the operation of statistical review of carcinogenicity studies in FDA's Center for Drug Evaluation and Research (CDER) is given in section II. A discussion of the validity of the design of the experiment is given in Section III. Section IV discusses methods of statistical analysis. Section V discusses how the results should be interpreted, and Section VI discusses data presentation and submission.

II. BACKGROUND

Assessment of the risk of drug exposure in humans includes an assessment of carcinogenicity in tests in rodents. The Division of Biometrics in the Office of Biostatistics, Center for Drug Evaluation and Research (CDER), Food and Drug Administration (FDA), is responsible for conducting statistical reviews of long-term animal (rodent) carcinogenicity studies of pharmaceuticals submitted by drug sponsors to FDA. In a carcinogenicity study of a new drug using a series of increasing dose levels, statistical tests for positive trends in tumor rates are usually of greatest interest, but as discussed in this document, in some situations, pairwise comparisons are considered to be more indicative of drug effects than trend tests.

[1] This guidance has been prepared by the Office of Biostatistics with the participation of the Office of Review Management, Center for Drug Evaluation and Research (CDER), Food and Drug Administration.

In statistical reviews of carcinogenicity studies, statisticians evaluate the validity of the designs and the appropriateness of methods of data analysis used by the sponsor. They also use raw study data in electronic form to perform additional statistical analyses.

The recommendations that follow are based on FDA's assessment of current literature, consultations with outside experts, and internal research.

III. VALIDITY OF THE DESIGN

Many factors determine the adequacy of a carcinogenicity study, including the species and strain of the animals, sample size, dose selection, method of allocation of animals, route of administration, animal care and diet, caging, drug stability, and study duration. Of particular interest to statisticians are the methods used to allocate animals to treatment groups and caging rotation, the determination of sample size, and the duration of the study.

Although not generally a statistical issue, dose selection is particularly critical. The premise of carcinogenicity testing most directly applicable to genotoxic mechanisms (when there may not be a pharmacological threshold, or identification of a threshold is difficult) is that long exposure duration and large doses in a small number of animals will be informative about the much smaller risks of lower doses and shorter durations of exposure in humans. As a result, the general goal should be to maximize rodent exposure by testing at maximum tolerated doses.

The International Conference on Harmonization (ICH) guidance entitled *S1C Dose Selection for Carcinogenicity Studies of Pharmaceuticals* (*S1C*) is an internationally accepted guidance for dose selection for carcinogenicity studies, and sponsors are advised to consult this document. The guidance allows for approaches to high dose selection based on toxicity endpoints (Sontag, Page, and Saffiotti 1976; Chu, Cueto, and Ward 1981), pharmacokinetic endpoints (multiple of maximum human exposure), pharmacodynamic endpoints, and maximal feasible dose. For further clarification, the appropriate medical review division should be consulted.[2]

Randomization should be used to allocate animals to treatment groups. Random assignment of experimental animals to different treatment groups allows the assumption that treatments will not be continually favored or handicapped by extraneous sources of variation over which the experimenter has no control (i.e., that possible bias will be minimized).

One area where bias can still be introduced, however, is in the microscopic evaluation of tissues. Currently, open or nonblinded microscopic evaluation of tissues from experimental animals is the routine practice adopted by veterinary pathologists in the generation of histopathological data in carcinogenicity studies. Veterinary pathologists do not favor blinded readings of slides of

[2] Sponsors can seek CDER's advance concurrence on carcinogenicity protocols and should consult other available guidance (e.g., ICH guidances S1A, S1B, S1C, S1C(R)). In addition, a draft guidance titled *Carcinogenicity Study Protocal Submissions* published in November 2000. Once finalized, that guidance will represent the Agency's thinking on that topic.

animal tissues/organs because they believe that blinded reading results in loss of information critical to interpretation, such as the ability to relate observations in different tissues. Furthermore, they argue that the variables constitute the baseline that defines the experimental control and that it is impractical to perform blinded slide readings because there are so many tissues from each animal. Mistakes can be easily made when assigning, opening codes, and recording results in blinded reading (Iatropoulos 1988; Prasse et al. 1986). There are others, however, who have argued that blinded evaluation should be used to prevent the bias that can be introduced by the pathologists' knowledge of the treatment groups of the tested animals (Temple, Fairweather, Glocklin, and O'Neill 1988). Certainly, blinded re-readings are common in close or disputed cases.

The number of animals remaining in a study for the full duration is an important statistical consideration. A sufficient number of animals should be used in an experiment to ensure reasonable power of statistical tests to detect true carcinogenic effects. It has been recommended that each dose and concurrent control group contain at least 50-60 animals of each sex. If interim sacrifices are planned, the initial number of animals should be increased by the number of animals scheduled for interim sacrifice. Prior assignments of treatment and designations for sacrifice of the animals should be made (Bannasch et al. 1986).

Animals are usually exposed to the test substance for essentially their entire normal life span, generally 24 months for rats and mice. The vast majority of carcinogenicity studies of pharmaceuticals using rats that are submitted to CDER for review have durations of 24 months and have reasonable survival. The duration of mouse studies ranges from 18 to 24 months, with many lasting only 18 or 21 months even though they have very low mortality at terminal sacrifice. One reason for using shorter durations in mouse studies appears to be a 1985 federal government publication stating that carcinogenicity studies should be conducted at least 18 months in mice and 24 months in rats (OFR 1985). The publication, however, goes on to say that a longer duration may be appropriate if cumulative mortality at the planned terminal sacrifice is low. CDER recommends that drug sponsors also conduct mouse studies for 24 months, unless there is excessive mortality as described below. Results of a recent study of the effect of shortened duration on the statistical power of carcinogenicity studies by Kodell, Lin, Thorn, and Chen (2000) support the CDER recommendation. The study showed that stopping at 18 months would reduce power to an unacceptable level for a variety of models of the tumorigenicity, and that the loss of power is too great to warrant an early stopping at 21 months, absent effects on survival.

However, early termination of a study for mortality, even if unavoidable, may render a study uninformative, leaving too few animals living long enough to represent adequate exposure to the chemical. This is especially important in the evaluation of the design validity of a negative study. In general, a 50 percent survival rate to weeks 80 to 90 of the 50 initial animals in any treatment group is considered adequate. The percentage can be lower or higher if the number of animals used in each treatment/sex group is larger or smaller than 50, but between 20 to 30 animals should be still alive during these weeks (Lin and Ali 1994). Whether a study could be terminated before the scheduled termination date if the survival of any treatment group goes below 50 percent or 20 to 30 surviving animals (provided that sufficient numbers of animals

were exposed through week 80 to 90) depends on the situation. For example, there is no reason to stop a study if the survival of only the low-dose group and/or the medium-dose group is altered, because the control vs. high-dose comparison will still be informative. If the survival of the high-dose group falls below 50 percent or 20-30 surviving animals after week 80, the study should be continued, either stopping dosing of animals in the high-dose group or terminating only the high-dose group, because the comparison of at least the control and low/middle doses would still be informative (the high-dose comparison would depend on the situation). A study could be terminated early if the survival of the control group (or groups) goes below 50 percent or 20-30 surviving animals after weeks 80 to 90, as the later comparisons would not be informative. Others have suggested, for example, that an experiment be terminated early when the survival of the control or the low-dose group is reduced to 20-25 percent of the original number of animals. If the mortality is increased only in the high-dose group, consideration can be given to early termination of that group (OFR 1985).[3] Because early study termination poses complex problems, it is strongly recommended that a decision to terminate a study or a study group early be made with input from the Center and the medical division responsible for the review of the associated application.

If in discussions with CDER, the Center approves the early termination of a study under this recommendation, the study's sponsor can be assured that the study will be considered by the Center as valid in terms of adequate duration of drug exposure.

IV. METHODS OF STATISTICAL ANALYSIS

A. An Overview of Complexities of Statistical Analysis of Tumor Data

The primary purpose of a long-term rodent carcinogenicity study of a new drug is to evaluate the oncogenic potential of the drug when it is administered to animals for most of their normal life span. The drug, however, may effect the mortality of different treatment groups. Test animals living longer are more likely to develop tumors than those dying early, as demonstrated by examples in the next section, and comparisons of tumor incidence rates among treatment groups based solely on the crude proportions of animals with tumors and failure to consider the rates at which animals develop tumors can cause serious bias in the analysis (Petro et al. 1980; McKight and Crowley 1984; Gart et al. 1986). Therefore, it is essential to make adjustment for the differences in mortality among treatment groups in the analysis of tumor data.

Tumor incidence (i.e., the rate of tumor onset among the previously tumor-free population) is the most appropriate measure of tumorigenesis for two reasons (Dinse 1994; McKight and Crowley 1984; and Malani and Van Ryzin 1988): (1) the tumor incidence rate reduces biases in the crude incidence proportion of animals with tumors that could arise from differences in mortality by adjusting for time differences and by conditioning the rate at each time point on the likelihood that an animal is still alive, and

[3] This article also appeared in Gart, J.J., D. Krewski, P. N. Lee, R. E. Tarone, and J. Wahrendorf, 1986, *U.S. Interagency Staff Group on Carcinogens.*

(2) unlike the death rate with or from tumors, the tumor incidence rate does not confound information about the course of tumors with information about the onset of tumors. Most tumors except those such as skin and mammary tumors, which can be detected by palpation and visual inspection, are occult and discovered only at the time of the animal's death. The exact tumor onset times are unknown.

The analysis of tumor data is complicated when adjustments are made for differences in mortality among treatment groups because of the lack of the observable onset time of occult tumors discussed above. A huge number of statistical procedures aiming to deal with these complexities have been proposed in the literature. They followed, in general, the strategy that "without direct observations of the tumor onset times, the desired survival adjustment usually is accomplished by making assumptions concerning tumor lethality, cause of death, multiple sacrifices, or parametric models" (Dinse 1994).

The prevalence method (Hoel and Walburg 1972; Peto et al. 1980), the death-rate method (Tarone 1975; Peto et al. 1980), and the onset-rate (Peto et al. 1980), discussed in Section C below, for analyzing nonlethal, lethal, and observable tumors, respectively, are based on an assumption, or information about tumor lethality. The Peto test (Peto et al. 1980), also discussed in Section C, for analyzing data of a tumor that is considered nonlethal to a subset of animals and lethal to the rest of the animals is also based on an assumption or information as to whether the tumor caused an animal's death. The analyses will become biased if the assumption or information on tumor lethality and cause of death is not valid or accurate (Dinse 1994).

Data from carcinogenicity studies do not always contain information of tumor lethality and cause of death. Even when such information is provided, the difficulty and subjectivity in the determination of cause of death and lethality of a tumor may render the information too inaccurate and unobjective to allow valid analysis using the above statistical methods. Another way of analyzing the tumor incidence rates without relying on the tumor lethality and cause-of-death information is to use a design with multiple sacrifices at different time points. Without cause-of-death information or simplifying assumptions, multiple sacrifices of groups of animals are necessary to identify tumor incidence rates of occult tumors from the bioassay data (McKight and Crowley 1984; Kodell and Ahn 1997; Dinse 1994). Statistical methods have been proposed for analyzing tumor incidence rates on the information from multiple sacrifices rather than on the information on cause of death and tumor lethality.[4] In reality, however, very few studies are conducted with multiple sacrifices because of the cost and complexity involved. Since it is rarely used in practice, no recommendations on analysis of data with multiple sacrifices are given in this guidance.

[4] See, for example, Berlin, Brodsky, and Clifford 1979; Dewanji and Kalbfleisch 1986; Portier and Dinse 1987; Dinse 1988; Malani and Van Ryzin 1988; Willams and Portier 1992; Malani and Lu 1993; Ahn and Kodell 1995; Kodell and Ahn 1996 and 1997; and Ahn, Kodell, and Moon 2000)

Finally, for data from bioassays with no information (or assumptions) regarding tumor lethality or cause-of-death and no interval sacrifices, Dinse (1991) and Lindsey and Ryan (1993 and 1994) have proposed survival-adjusted statistical tests that focus on tumor incidence for dose-related trends by making some parametric assumptions. Dinse's test is based on the assumption of a constant difference between the death rates of animals with and without a tumor while Lindsey and Ryan's test assumes a constant ratio for those death rates. Recently, other statistical procedures of this type have been proposed in the literature for dealing with the complexities of analysis of tumor data. Those procedures do not require data on tumor lethality and cause of death, or the use of multiple sacrifices. Among those procedures, the poly-3 (in general poly-k) tests (Bieler and Portier 1988; Dinse 1994), and the ratio trend test (a modified poly-k test) (Bieler and Williams 1993; Dinse 1994) have been most extensively studied and shown to perform well under actual study conditions. Detailed discussions of the poly-k tests and the ratio trend test are given in Section D.

Some of the recently proposed statistical procedures, such as those described by Kodell, Pearce, Turturro, and Ahn (1997), and Moon, Ahn, and Kodell (2000), deal with the complexities of the tumor data analysis from a somewhat different direction. These procedures use a constrained nonparametric maximum likelihood estimation method to impute (estimate) incidence rates of fatal tumors and nonfatal tumors for time intervals preceding the final time interval of terminal sacrifice. These procedures do not require tumor lethality and cause-of-death information and are applicable to studies with only a single sacrifice. The imputed tumor incidence rates can then be used in the death-rate method, prevalence method, or the Peto test. The properties of these procedures have not yet been widely studied, and they involve extensive computations.

B. Adjustment of Tumor Rates for Intercurrent Mortality

Intercurrent mortality refers to all deaths other than those resulting from a tumor being analyzed for evidence of carcinogenicity. Like human beings, older rodents have a many fold higher probability of developing or dying of tumors than those of a younger age. Therefore, in the analysis of tumor data, it is essential to identify and adjust for possible differences in intercurrent mortality among treatment groups to eliminate or reduce biases caused by these differences. It has been pointed out that "the effects of differences in longevity on numbers of tumor-bearing animals can be very substantial, and so, whether or not they (the effects) appear to be, they should routinely be corrected when presenting experimental results" (Peto et al. 1980). The following examples demonstrate this point.

Example 1 (Peto et al. 1980). Consider a mouse study consisting of one control group and one treated group of 100 animals each. A very toxic but not carcinogenic new drug is administered to the animals in the diet for 2 years. Assume that the spontaneous incidental tumor rates for both groups are 30 percent at 15 months and 80 percent at 18 months and that the mortality rates at 15 months for the control and the treated groups are 20 percent and 60 percent, respectively, due to the toxicity of the drug. The results of this experiment are summarized in Table 1.

Table 1: Effects of Differences in Mortality on Tumor Incidence Rates, Example 1

	Control			Treated		
	T	D	%	T	D	%
15 Months	6	20	30	18	60	30
18 Months	64	80	80	32	40	80
Totals	70	100	70	50	100	50

Note: T = Incidental Tumors Found at Necropsy. D = Deaths

If one looks only at the overall tumor incidence rates of the control and the treated groups (70 percent and 50 percent, respectively) without considering the significantly higher early deaths in the treated group caused by the toxicity of this new drug, one can misinterpret the apparent significance ($p = 0.002$, 1-tailed) as showing a decrease in the treated group in this tumor type. The one-tailed p-value is 0.5, however, showing no effect of treatment when the survival-adjusted prevalence method is used.

Example 2 (Gart, Krewski, Lee, Tarone, and Wahrendorf 1986). Assume that the design used in this experiment is the same as the one used in the experiment in Example 1. Also, assume that the tested new drug in this example induces an incidental tumor that does not directly or indirectly cause animal deaths, in addition to having severe toxicity as in the previous example. Assume further that the incidental tumor prevalence rates for the control and treated groups are 5 percent and 20 percent, respectively, before 15 months of age, and 30 percent and 70 percent, respectively, after 15 months of age; and that the mortality rates at 15 months are 20 percent and 90 percent for the control and the treated groups, respectively. The results of this experiment are summarized in Table 2.

Table 2: Effects of Differences in Mortality on Tumor Incidence Rates, Example 2

	Control			Treated		
	T	D	%	T	D	%
Before 15 Months	1	20	5	18	90	20
After 15 Months	24	80	30	7	10	70
Totals	25	100	25	25	100	25

Note: T = Incidental Tumors Found at Necropsy. D = Deaths

The age-specific tumor incidence rates are significantly higher in the treated group than those in the control group. The survival-adjusted prevalence method yielded a one-tailed p-value of 0.003, revealing a clear tumorigenic effect of the new drug. The overall tumor incidence rates, however, are 25 percent for the two groups. Without adjusting the significantly higher early mortality in the treated group, the positive finding would be missed.

Peto et al. (1980) recommend that, whether or not survival among treatment groups is significantly different, tumor rates should routinely be adjusted for survival when presenting experimental results. The Cox test (Cox 1972; Thomas, Breslow, and Gart 1977; Gart et al. 1986); the generalized Wilcoxon or Kruskal-Wallis test (Breslow 1970; Gehan 1965; Thomas, Breslow, and Gart 1977); and the Tarone trend tests (Cox 1959; Peto et al. 1980; Tarone 1975) are routinely used to test for heterogeneity in survival distributions and significant dose-response relationships (trends) in survival.

C. Statistical Analysis of Tumor Data With Information About Cause of Death, Tumor Lethality, but Without Multiple Sacrifices

1. Role of the Tumor in Animal's Death (Contexts of Observation of Tumor Types)

One way to choose the appropriate survival-adjusted methods in the analysis of tumor data is to base analysis on the role that a tumor plays in causing the animal's death. Tumors can be classified as *incidental*, *fatal*, and *mortality independent* or (observable) according to the contexts of observation described in Peto et al. (1980). Tumors that are not directly or are indirectly responsible for an animal's death, but are merely observed at the autopsy of the animal after it has died of an unrelated cause, are said to have been observed in an *incidental context*. Tumors that kill the animal, either directly or indirectly, are said to have been observed in a *fatal context*. Tumors, such as skin tumors, for which detection occurs at times other than when the animal dies are said to have been observed in a *mortality independent* (or observable) *context*. To apply a survival-adjusted method correctly based on such information, it is essential that the role of a tumor in an animal's death (or the context of observation of a tumor) be determined as accurately as possible.

Different statistical techniques have been proposed for analyzing data of tumors when information about the role of a tumor in causing death is available. For example, the prevalence method, the death rate method, and the onset rate method are recommended for analyzing data on tumors observed in incidental, fatal, and mortality independent contexts of observation, respectively (Peto et al. (1980)). In that paper, Peto et al. demonstrate the possible biases resulting from misclassification of incidental tumors as fatal tumors, or of fatal tumors as incidental tumors.

The determination of whether a tumor is incidental, fatal or mortality independent is often difficult, especially for the first two classifications, as it is often hard to tell whether a tumor caused an animal's death. According to Haseman (1999), in practice, a continuum exists between these two extremes: many tumors contribute ultimately to an animal's death, but are not instantly (or even rapidly lethal). Such tumors technically are neither incidental nor fatal, and it remains unclear how such tumors should be regarded. Even if the information on the circumstances of individual animals and tumors is reliable and available, it is overly simplistic to assume that all tumors of a given type are 100 percent fatal or 100 percent incidental. It is likely that there will be a mixture of incidental and fatal tumors.

As noted above, alternative survival-adjusted statistical procedures that do not need such information have been developed and used for tumor data analysis. Some of the procedures are discussed briefly in the Section IV.C.1 and in detail in Section IV.D. The alternative procedures should be used to replace the procedures proposed by Peto et al. (1980) in the analysis of tumor data when there is no information available or the information is not accurate enough to perform a meaningful statistical analysis.

2. Statistical Analysis of Incidental Tumors

The prevalence method described in the paper by Peto et al. (1980) should be used in testing for positive trends in prevalence rates of incidental tumors. The method is described briefly here.

The method focuses on the age-specific tumor prevalence rates to correct for intercurrent mortality differences among treatment groups in the test for positive trends or differences in incidental tumors. The experiment period is partitioned into a set of intervals plus interim (if any) and terminal sacrifices. The incidental tumors are then stratified by those intervals of survival times. The selection of the partitions of the experiment period does not matter very much as long as the intervals are "not so short that the prevalence of incidental tumors in the autopsies they contain is unstable, nor yet so large that the real prevalence in the first half of one interval could differ markedly from the real prevalence in the second half" (Peto, et al. 1980).

In each time interval, for each group, the observed and the expected numbers of animals with a particular tumor type found in necropsies are compared. The expected number is calculated under the null hypothesis that there is no dose-related trend. Finally, the differences between the observed and the expected numbers of animals found with the tumor type after their deaths are combined across all time intervals to yield an overall test statistic using the method described in a paper by Mantel and Haenszel (1959).

The following derivation of the Peto prevalence test statistic uses the notations in Table 3. Let the experiment period be partitioned into the following m intervals $I_1, I_2, \ldots, I_m$. As mentioned before, interim (if any) and terminal sacrifices should be treated as separate intervals.

Let R_k be the number of animals that have not died of the tumor type of interest but come to autopsy in time interval k, P_{ik} be the proportion of R_k in group i, and O_{ik} be the observed number of autopsied animals in group i and interval k found to have the incidental tumor type.

Define $O_{.k} = \sum_i O_{ik}$.

The number of autopsied animals expected to have the particular incidental tumor in group i and interval k, under the null hypothesis that there is no treatment effect, is:

$$E_{ik} = O_{.k}\, P_{ik}.$$

The variance-covariance of $(O_{ik} - E_{ik})$ and $(O_{jk} - E_{jk})$ is:

$$V_{ijk} = \alpha_k P_{ik}(\delta_{ij} - P_{jk})$$

where

$$\alpha_k = O_{.k}(R_k - O_{.k}) / (R_k - 1)$$

and

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases}$$

Define

$$O_i = \sum_k O_{ik}$$

$$E_i = \sum_k V_{ijk}.$$

and $V_{ij} = \sum_k V_{ijk}$.

The test statistic T for the positive trend in the incidental tumor is defined as:

$$T = \sum_i D_i (O_i - E_i)$$

with estimated variance

$$V(T) = \sum_i \sum_j D_i D_j V_{ij}$$

where D_i is the dose level of the ith group.

Under the null hypothesis of equal prevalence rates among the treatment groups, the statistic

$$Z = T / [V(T)]^{1/2}$$

is approximately distributed as a standard normal.

Table 3: Notations Used in the Derivation of Peto Prevalence Test Statistics

Interval		Group	0	1	. . .	i	. . .	r	Sum
		Dose	D_0	D_1	. . .	D_i	. . .	D_r	
I_1	R_1		O_{01}	O_{11}	. . .	O_{i1}	. . .	O_{r1}	$O_{.1}$
			P_{01}	P_{11}	. . .	P_{i1}	. . .	P_{r1}	$P_{.1}$
I_2	R_2		O_{02}	O_{12}	. . .	O_{i2}	. . .	O_{r2}	$O_{.2}$
			P_{02}	P_{12}	. . .	P_{i2}	. . .	P_{r2}	$P_{.2}$
.	.		.	.	. . .	.	. . .	.	.
.	.		.	.	. . .	.	. . .	.	.
.	.		.	.	. . .	.	. . .	.	.
I_k	R_k		O_{0k}	O_{1k}	. . .	O_{ik}	. . .	O_{rk}	$O_{.k}$
			P_{0k}	P_{1k}	. . .	P_{ik}	. . .	P_{rk}	$P_{.k}$
.	.		.	.	. . .	.	. . .	.	.
.	.		.	.	. . .	.	. . .	.	.
.	.		.	.	. . .	.	. . .	.	.
I_m	R_m		O_{0m}	O_{1m}	. . .	O_{im}	. . .	O_{rm}	$O_{.m}$
			P_{0m}	P_{1m}	. . .	P_{im}	. . .	P_{rm}	$P_{.m}$

Notes:

R_k: Number of animals that have not died of the tumor type of interest, but come to autopsy in time interval k.

P_{ik}: Proportion of R_k in group i.

O_{ik}: Observed number of autopsied animals in group i and interval k found to have the incidental tumor type.

$O_{.k}$: $\sum_i O_{ik}$.

As noted above, to use the prevalence method, the experimental period should be partitioned into a set of intervals plus interim (if any) and terminal sacrifices. The following partitions (in weeks) are used most often by statisticians in CDER in 2-year studies: (1) 0 – 50, 51 – 80, 81 – 104, interim sacrifice (if any), and terminal sacrifice; (2) 0 – 52, 53 – 78, 79 – 92, 93 – 104, interim sacrifice (if any), and terminal sacrifice (proposed by National Toxicology Program); and (3) partition determined by the *ad hoc runs* procedure described in Peto et al. (1980).

The data for liver hepatocellular adenoma in male mice from a carcinogenicity study are used as an example to explain the prevalence method for testing the positive trend in tumor rates of an incidental tumor. There were four treatment groups. The control group had 100 animals, the three treated had 50 animals each. The dose levels used were 0, 10, 20, and 40 mg/kg/day for the control, low-, medium-, and high-dose groups, respectively. The study lasted for 106 weeks. In this example, the study period was partitioned into four intervals, 0 – 50, 51 – 80, 81 – 106, and terminal sacrifice. The numbers of animals died and necropsied, and the numbers of necropsied animals with liver hepatocellular adenoma by treatment group in each interval are included in Table 4.

Table 4: Data of Liver Hepatocellular Adenoma of Male Mice

Time Intervals (Weeks)	Groups											
	Control			Low			Medium			High		
	T	N	%	T	N	%	T	N	%	T	N	%
0 – 50	0	6	0	0	2	0	0	2	0	0	4	0
51 – 80	1	26	4	1	18	6	3	17	18	1	13	8
81 – 106	4	37	11	2	14	14	2	14	14	7	19	37
Terminal Sacrifice	2	31	6	5	16	31	3	17	18	4	14	29
Total	7	100	7	8	50	16	8	50	16	12	50	24

Notes: T = Number of necropsies with the above tumor.
N = Number of necropsies during a time interval.
% = Percent of necropsies with the above tumor.

The observed incidences and the expected incidences of the tumor type calculated under the null hypothesis that there is no trend (or drug induced increase) are shown in Table 5. The expected tumor rates in each interval were calculated in the following way. First, the tumor rate for the interval using data of all treatment groups in the interval was estimated. For example, the estimated tumor rate for

the interval 51 – 80 weeks was 6/74 = 0.0811. Second, the expected incidences for individual groups in the interval were calculated by multiplying the numbers of necropsies by the estimated tumor rate. For the interval 51 – 80 weeks, the expected tumor rates for the control, low-, medium-, and high-dose groups were 26x(6/74)=2.11, 18x(6/74)=1.46, 17x(6/74)=1.38, and 13x(6/74)=1.05, respectively.

Table 5: Observed and Expected Tumor Incidences Liver Hepatocellular Adenoma of Male Mice

Time Intervals (Weeks)	Observed & Expected Incidences	Groups			
		Control	Low	Medium	High
0 – 50	Observed	0	0	0	0
	Expected	0	0	0	0
51 – 80	Observed	1	1	3	1
	Expected	2.11	1.46	1.38	1.05
	Observed	4	2	2	7
	Expected	6.61	2.50	2.50	3.39
Terminal Sacrifice	Observed	2	5	3	4
	Expected	5.56	2.87	3.05	2.51
Total	Observed	7	8	8	12
	Expected	14.28	6.83	6.93	6.95

Note: The expected tumor incidences were calculated under the null hypothesis that there is no trend.

The test statistics T's and their variances V(T)'s for the data of the 5 intervals calculated by the formulas listed above are included in Table 6. It is noted that the first interval, 0 - 50 weeks, did not contribute anything to the overall test result since none of the 14 animals that died during the first time interval developed liver hepatocellular adenoma. The overall result shows a statistically significant positive trend in tumor rates of this tumor (with one-sided p-value 0.002).

Table 6: Test Statistics, Their Variances, z-values, and P-value of Peto Prevalence Analysis of Incidental Tumors

Liver Hepatocellular Adenoma of Male Mice

Time Intervals (Weeks)	T-Stat T	Variance of T-Stat V(T)	$z = \frac{T}{[V(T)]^{0.5}}$	P-Value
0 - 50	-	-	-	-
51 - 80	25.6756	1116.583	0.7683	0.2211
81 - 106	129.2857	3091.314	2.3253	0.0100
Term. Sacr.	79.7435	2445.855	1.6124	0.0534
Overall Total	234.7048	6653.752	2.8773	0.0020

Note: The z and p-value columns do not add up to the totals. The z and p-value of overall total row were calculated based on the T and V(T) of the row.

Also as noted above, this method used normal approximation in the test for positive trend or difference in tumor prevalence rates. The accuracy of the normal approximation depends on the number of tumor occurrences in each group in each interval, the number of intervals used in the partitioning, and the mortality patterns. The approximation may not be stable and reliable when the numbers of tumor occurrences across treatment groups are small. In this situation, an exact permutation trend test based on an extension of the hypergeometric distribution (to be discussed in Subsection III.C.6) should be used to test for the positive trend in tumor prevalence rates.

3. Statistical Analysis of Fatal Tumors

It is recommended that the death rate method described in Peto et al. (1980) be routinely used to test for the positive trend or difference in incidence of tumors observed in a fatal context.

The notations of Subsection III.C.2 with some modifications will be used in this section to derive the test statistic of the death rate method. Now let $t_1 < t_2 < \ldots < t_m$ be the time points when one or more animals died of the fatal tumor of interest. These time points are used to replace the intervals used in the prevalence method. The notations in Table 3 are redefined as follows:

R_k: The number of animals at risk of all groups just before t_k.

P_{ik}: (The same as in the prevalence method) Proportion of R_k in Group i.

O_{ik}: Observed number of animals in Group i dying of the fatal tumor of interest at time t_k.

$$O_{.k} = \sum_i O_{ik}.$$

As in the prevalence method, the test statistic T for the positive trend in the fatal tumor is defined as:

$$T = \sum_i D_i(O_i - E_i)$$

with estimated variance

$$V(T) = \sum_i \sum_j D_i D_j V_{ij}.$$

where D_i, O_i, E_i, and V_{ij} are defined similarly as in Subsection III. C.2.

Under the null hypothesis of equal tumor rates across the treatment groups, the statistic

$$Z = T / [V(T)]^{1/2}$$

is distributed approximately as standard normal.

4. Statistical Analysis of Tumors Observed in Both Incidental and Fatal Contexts

When a tumor is fatal for some animals and is incidental for other animals in the experiment, data for the incidental and fatal tumors should be analyzed separately by the prevalence and the death rate methods. Results from the different methods can then be combined to yield an overall result. The combined overall result can be obtained simply by adding together either the separate observed frequencies, the expected frequencies, and the variances, or the separate T statistics and their variances (Peto et al. 1980).

5. Statistical Analysis of Mortality Independent Tumors

Tumors that are mortality independent, such as skin tumors and mammary gland tumors, which are visible and/or can be detected by palpation in living animals, are analyzed by CDER statistical reviewers using the onset rate method. The onset rate method for mortality independent tumors and the death rate method for fatal tumors are essentially the same in principle except that the endpoint in the onset rate method is the occurrence of such a tumor (e.g., skin tumor reaching some prescpecified size) rather than the time or cause of the animal's death.

In the onset rate method, all those animals that, although still alive, have developed the particular mortality independent tumor and hence are no longer at risk for such a tumor are excluded from the calculation of the numbers of animals at risk. The R_k, P_{ik}, and O_{ik} described in Section III.C.3 are now redefined as follows for the onset rate method:

R_k: The number of animals alive and free of the mortality independent tumor of interest in all groups just before t_k.

P_{ik}: (The same as in the death rate method) Proportion of R_k in Group i.

O_{ik}: Observed number of animals in Group i found to have developed the mortality independent tumor of interest at time t_k.

The test statistic T and its estimated variance V(T) are the same as those defined in the death rate method.

6. Exact Methods

As noted in previous sections, the prevalence method, the death rate method, and the onset rate method used normal approximation in the test for the positive trend in tumor incidence rates. Mortality patterns, the number of intervals used in the partitioning of the study period, and the numbers and patterns of tumor occurrence in each individual interval have effects on the accuracy of the normal approximation. It is also well known that the approximation results may not be stable and reliable, and tend to underestimate the exact p-values when the total numbers of tumor occurrence across treatment groups are small (Ali 1990). In this situation, the exact permutation trend test should be used to test for the positive trend (Gart et al. 1986; Goldberg 1985). The exact permutation trend test is a generalization of the Fisher's exact test to a sequence of 2x(r+1) tables. The exact permutation trend test procedure described below is for tumors observed in an incidental context. However, the positive trends in incidence rates of tumors observed in a fatal or in a mortality independent context can be tested in a similar way. In those cases, the number of 2x(r+1) tables will be equal to the number of time points when one or more animals died of a particular fatal tumor, or when one or more animals developed a particular mortality independent tumor. Fairweather et al. (1998) contains a discussion on the limitations of applying exact methods to fatal tumors.

The exact method is derived by conditioning on the row and column marginal totals of each of the 2x(r+1) tables formed from the partitioned data set of Table 3. Consider the k-th interval I_k (in Table 3) and rewrite it as in Table 7. Let the column totals C_{0k}, C_{1k}, …, C_{rk} and the row totals $O_{.k}$ and $A_{.k}$ be fixed. Define $P_{ik} = C_{ik}/R_k$. Then the quantities $E_{ik} = O_{.k}P_{ik}$, $V_{ijk} = \alpha_k P_{ik}(\delta_{ij}-P_{jk})$, E_i, and V(T) (defined in Subsection III.C.2) are all known constants.

Table 7: The Data in the k-th Time Interval I_k Is Written as a 2 x (r + 1) Table

Group	0	1 . . .	i	. . .	r	
Dose	D_0	D_l . . .	D_i	. . .	D_r	Total
# w tumor	O_{0k}	O_{lk} . . .	O_{ik}	. . .	O_{rk}	$O_{.k}$
# w/o tumor	A_{0k}	A_{lk} . . .	A_{ik}	. . .	A_{rk}	$A_{.k}$
Total	C_{0k}	C_{lk} . . .	C_{ik}	. . .	C_{rk}	R_k

Now let y be the observed value of $Y = \sum D_i O_i$, where $O_i = \sum_k O_{ik}$, the total number of tumor bearing animals of the tumor of interest in treatment group i. Then (under conditioning on the column and row marginal totals in each table) the observed significance level or

$$\text{p-value} = P[\sum D_i O_i >= y] = P(\sum_i D_i \sum_k O_{ik} >= y) = P(\sum_k \sum_i D_i O_{ik} >= y)$$

$$= P(\sum_k Y_k >= \sum y_k) = P(Y >= y),$$

where $Y = \sum Y_k = \sum_k \sum_i D_i O_{ik}$ and $y = \sum y_k$, the observed value of Y.

This p-value (P(Y>=y) is computed from the exact permutational distribution of Y. Given the observed row and column marginal totals in a 2x(r+1) table, all possible tables having the same marginal totals can be generated. Let S_k (k=1,2,…,K) be the set of all such tables generated from the k-th observed table. From a set of K tables taking one from each S_k and assuming independence between the K tables, the above expression for the p-value can now be written as

$$\text{p-value} = \sum [P(Y_1 = y_i) \ldots P(Y_k = y_k)]$$

where $y_k = \sum_i D_i O_{ik}$ (k=1,2,…,K), the sum is over all sets of K tables such that $y_i + y_2 + \ldots + y_k >= y$, the observed value of Y, and $P(Y_k = y_k)$ is the conditional probability given the marginal totals in the k-th table, i.e.,

$$P(Y_k=y_k) = \left[\binom{C_{0k}}{O_{0k}} \binom{C_{1k}}{O_{1k}} \cdots \binom{C_{rk}}{O_{rk}} \right] \Big/ \binom{R_k}{O_k}$$

Example (Lin and Ali 1994). Consider an experiment with 3 treatment groups (control, low-, and high-dose) with dose levels $D_0=0$, $D_1=1$, and $D_2=2$, respectively. Suppose the study period is partitioned into the intervals 0-50, 51-80, 81-104 weeks, and the terminal sacrifice week. Consider a tumor type (classified as incidental) with data in Table 8.

Table 8: Hypothetical Tumor Data for Exact Permutation Trend Test

Time intv.		Dose levels 0	1	2	Total
0 - 50	O	0	0	0	0
	C	1	3	3	7
51 - 80	O	0	0	0	0
	C	4	5	7	16
81 - 104	O	0	0	2	2
	C	10	12	15	37
Term. Sacr.	O	0	1	0	1
	C	35	30	25	90

O = observed tumor count, C = number of animals necropsied

Since all the observed tumor counts (i.e., O's) in the first two time intervals are zeros, the data for these intervals will not contribute anything to the test statistic, and these intervals may be ignored. The observed subtables formed from the last two intervals are given in Table 9.

Now, generate all possible tables from observed subtable 1. Since the marginal totals are fixed, these tables may be generated by distributing the total tumor frequency $O_{.1}(=2)$ among the three treatment groups. Thus, each table will correspond to a configuration of this distribution of $O_{.1}$. The configurations, the values of Y_1, and the $P(Y_1=y_1)$ are shown in Table 10.

Table 9: Observed Subtables From the Above Hypothetical Tumor Data

	Observed subtable 1					Observed subtable 2			
Dose	0	1	2	Total	Dose	0	1	2	Total
O	0	0	2	$2=O_{.1}$	O	0	1	0	$1=O_{.2}$
A	10	12	13	$35=A_{.1}$	A	35	29	25	$89=A_{.2}$
C	10	12	15	$37=R_1$	C	35	30	25	$90=R_2$

Table 10: All Possible Configurations of $o_{.1}$ and the Corresponding Hypergeometric Probabilities

Configurations	y_1	$P(Y_1=y_1)$
0, 0, 2	4	.15766
0, 2, 0	2	.09910
2, 0, 0	0	.06757
0, 1, 1	3	.27027
1, 0, 1	2	.22523
1, 1, 0	1	.18018

To illustrate the computation of y_1 and $P(Y_1=y_1)$ consider the last row. Here $y_1= D_0x1 + D_1x1 + D_2x0 = 0x1 + 1x1 + 2x0 = 1$, and

$$P(Y_1=1) = \left[\binom{10}{1}\binom{12}{1}\ldots\binom{15}{0}\right] \Big/ \binom{37}{2} = 0.18018$$

The configurations and probabilities obtained from observed subtable 2 are given in Table 11.

Note that the first configuration (0,0,2) in Table 8 corresponds to the observed subtable 1 with a value of $y_1= (0x0)+(2x2)=4$ and a probability of .15766, and the second configuration (0,1,0) in Table 8 corresponds to the observed subtable 2 with a value of $y_2= (0x0)+(1x1)+(0x0)=1$ and a probability of .33333. Thus, the observed value of $y = y_1+y_2 = 4+1=5$. Now the exact p-value (right-tailed) is calculated as follows:

$$P(Y = Y_1+Y_2 >=5) = P(Y_1=4,Y_2=1)+P(Y_1=4, Y_2=2)+P(Y_1=3, Y_2=2)$$

$$= .15766 \times .33333 + .15766 \times .27778 + .27027 \times .27778$$

= .17142

Table 11: All Possible Configurations of $O_{.2}$ and the Corresponding Hypergeometric Probabilities

Configurations	y_1	$P(Y_1=y_1)$
0, 0, 1	2	.27778
0, 1, 0	1	.33333
1, 0, 0	0	.38889

For the purpose of comparison, it should be noted that the normal approximated p-value for the data set in the above example is .0927.

D. Statistical Analysis of Data Without Information About Cause of Death and Without Multiple Sacrifices

As noted previously, in the analysis of tumor data, it is essential to identify and adjust for possible differences in intercurrent mortality among treatment groups to eliminate or reduce biases caused by these differences. It is also necessary for the analysis to appropriately account for tumor lethality. The widely used prevalence method, the death rate method, and the onset rate methods for analyzing incidental, fatal, and mortality independent tumors, respectively, described in previous sections rely on good information on tumor lethality and cause of death. There are situations in which sponsors have not included tumor lethality and cause of death information in their statistical analyses and electronic data sets. Under those situations, statistical reviewers in CDER either treated all tumors as incidental or relied on cause of death assessments by the reviewing pharmacologists and toxicologists in the Center. There are consequences in misclassifying tumors as lethal or not in survival adjusted statistical tests. The prevalence method will reject the null hypothesis of no positive trend less frequently than it should as the lethality of a tumor increases (Peto et al. 1980; Dinse 1994). This will increase the probability of failing to detect true carcinogens.

The Bailer-Portier poly-3, and poly-6 (in general poly-k) tests (Bailer and Portier 1988; Dinse 1994) have been proposed for testing linear trends in tumor rates. These tests are basically modifications of the survival unadjusted Cochran-Armitage test (Cochran 1954; Armitage 1955, 1971) for linear trend in tumor rate. If the entire study period is considered as one interval, the data for a particular tumor type will be in the form of Table 12. The notations in Table 12 to be used to explain these tests are the same as those in Table 7 except that the k-th interval now is the entire study period. The second subscript, k, for the k-th interval was dropped from the notations.

Table 12: The Data Using the Entire Study Period as an Interval

Group	0	1	. . .	i	. . .	r	
Dose	D_0	D_1	. . .	D_i	. . .	D_r	Total
# w. tumor	O_0	O_1	. . .	O_i	. . .	O_r	O
# w/o tumor	A_0	A_1	. . .	A_i	. . .	A_r	A
Total	C_0	C_1	. . .	C_i	. . .	C_r	R

The Cochran-Armitage test statistic for linear tend in tumor rate is defined as (Armitage 1955):

$$\chi_{CA}^2 = \frac{R\{R\sum O_i D_i - O\sum C_i D_i\}^2}{O(R-O)\{R\sum C_i D_i^2 - \{\sum C_i D_i\}^2\}} \quad \text{or}$$

$$= \frac{\{\sum D_i (O_i - E_i)\}}{\sum E_i D_i^2 - (\sum E_i D_i)^2 / O}$$

Where $O = \sum O_i$, $A = \sum A_i$, $R = \sum C_i$, $E_i = O C_i / R$.

The test statistic χ_{CA}^2 is distributed approximately as χ^2 on one degree of freedom.

The Cochran-Armitage linear trend test is based on a binomial assumption that all animals in the same treatment group have the same risk of developing the tumor over the duration of the study. However, as noted previously, the animal's risk of developing the tumor increases as study time increases. The assumption is thus no longer valid if some animals die earlier than others. It has been shown that as long as the mortality patterns are similar across treatment groups, the Cochran-Armitage test is still valid, although it may be slightly less efficient than a survival adjusted test (Dinse 1994). However, if the mortality patterns are different across treatment groups, the Cochran-Armitage test can give very misleading results.

The Bailer-Portier poly-3 test adjusts for differences in mortality among treatment groups by modifying the number of animals at risk in the denominators in the calculations of overall tumor rates in the Cochran-Armitage test to reflect "less-than-whole-animal contributions for decreased survival" (Bailer and Portier 1988). The modification is

made by defining a new number of animals at risk for each treatment group. The number of animals at risk for the i-th treatment group C^*_i is defined as

$$C^*_i = \sum W_{ij}$$

where w_{ij} the weight for the j-th animal in the i-th treatment group, and the sum is over all animals in the group.

Bailer and Portier (1988) proposed the weight w_{ij} as follows:

$w_{ij} = 1$ to animals dying with the tumor, and

$w_{ij} = (t_{ij} / t_{sacr})^3$ to animals dying without the tumor

where t_{ij} is the time of death of the j-th animal in the i-th treatment group, and t_{sacr} is the time of terminal sacrifice.

The power of 3 used in the weighting is from the observation that tumor incidence can be modeled as a polynomial of order 3. Similarly the poly-6 test (or the general poly-k test) assigns the weight $w_{ij} = (t_{ij} / t_{sacr})^6$ (or $w_{ij} = (t_{ij} / t_{sacr})^k$) to animals dying without the tumor when the tumor incidence is close to a polynomial of order 6 (or order k).

The class of Bailer-Portier poly-k tests are carried out by replacing the C_i's by the new numbers of animals at risk C^*_i's in the calculation of the above Cochran-Armitage test statistic.

The class of Bailer-Portier poly-k tests adjust differences in survival, do not need the information about cause of death, and call for only a (the terminal) sacrifice. Results of simulation studies by Bailer and Portier (1988), and Dinse (1994) show that the tests performed very well under many conditions simulated. They are also relatively robust to (not affected greatly by) tumor lethality.

Bieler and Williams (1993) pointed out that, since animal survival time is generally not a fixed quantity, the numerators and denominators of the adjusted quantal response estimates

$$p^*_i = O_i / C^*_i$$

are both subject to random variation.

Bieler and Williams (1993) proposed a test called the ratio trend test (also called Bieler-Williams poly-3 test), which is another modification to the Cochran-Armitage linear trend test. The ratio trend test employs the adjusted quantal response rates calculated in Bailer and Portier (1988) and the delta method (Woodruff 1971) in the estimation of the variance of the adjusted quantal response rates $p^*_i = O_i / C^*_i$.

The computational formula for Bieler-Williams ratio trend (modified C-A) test statistic is given as follows:

$$\chi^2_{BW} = \frac{\sum m_i p^*_i D_i - (\sum m_i D_i)(\sum m_i p^*_i) / \sum m_i}{\{c\,[\sum m_i D_i^2 - (\sum m_i D_i)^2 / \sum m_i]\}^{1/2}}$$

where

$c = \sum\sum (r_{ij} - r_{i.})2 / [R - (r + 1)]$

$m_i = (C_I^*)^2 / C_i$

$r_{ij} = y_{ij} - p^* w_{ij}$

$r_{i.} = \sum r_{ij} / Ci$

y_{ij} = tumor response indicator (0 = absent at death, 1 = present at death) for the jth animal in the ith group.

Bieler and Williams (1993) showed that the Bailer-Portier poly-3 trend test is anti-conservative when tumor incidence rates are low and treatment toxicity is high. Their study also showed that for tumors with low background rates, the ratio trend test (Bieler-Williams poly-3 test) yielded actual Type I errors close to the nominal levels used and was observed to be less sensitive than the Bailer-Portier poly-3 trend test to misspecification of the shape of tumor incidence function and the magnitude of treatment toxicity.

A more recent simulation study by Chen, Lin, Juque, and Arani (2000) showed the following additional results about the characteristics of the Bailer-Portier poly-3 and the ratio trend tests (Bieler-William poly-3 test). For individual tumor types, the two tests for trend yield attained Type I errors around the nominal levels (5 percent and 1 percent) for tumors with spontaneous rates in the range between 2 percent to 20 percent. When spontaneous rates are below the range, the two tests become conservative (i.e., less likely to show statistically significant results). For tumors with spontaneous rates above 20 percent to 60 percent (the upper rate used in the simulation), the ratio trend test still maintains the attained Type I error rates close to the nominal levels, but the Bailer-Portier poly-3 test becomes more and more conservative as the rates go up. The introduction of the compound symmetric correlation structure (although a not very realistic structure) among tumors corrects the problem of conservativeness somewhat in the Bailer-Portier poly-3 test, but the patterns of conservativeness continue to exist.

The ratio trend test (Bieler-William poly-3 test), like the Bailer-Portier poly-3 test, adjusts differences in survival, does not need the information about cause of death, and results only in a (the terminal) sacrifice. Results of simulation studies (Bieler and Williams 1993; Chen, Lin, Huque, and Arani 2000) show that the tests performed very well under many simulated conditions. It is also shown to be relatively robust to (not affected greatly by) tumor lethality, misspecification of the shape of tumor incidence function, and the magnitude of treatment toxicity. The ratio trend test (Bieler-William poly-3 test) should be used to replace the asymptotic tests that depend on the information of tumor lethality and cause of death when the information is unavailable.

Theoretically, exact versions of the class of tests can be developed for testing data of studies with small numbers of tumor bearing animals by applying the test procedures to all possible permuted configurations of the outcome. However, because these tests use risk sets based on all animals in each treatment group, the computations involved in the exact tests will be extensive. Therefore, for studies with small numbers of tumor bearing animals, the current practice of treating them as incidental tumors and applying the exact permutation trend test should continue.

E. Statistical Analysis of Data From Studies With Dual Controls

There are two categories of studies with dual control groups. The first category usually consists of studies using an untreated control group and a vehicle control group (Category A). Other variations of nonidentically treated, nondrug treatment controls groups are also occasionally used and are included in Category A for statistical purposes. The second category (Category B) includes studies that use two identical control groups (Society for Toxicology 1982; Haseman, Winbush, and O'Donnell 1986).

The main reasons for using two differently treated controls, generally an untreated control and a vehicle control in a study in Category A are to determine whether the vehicle has effects on tumor incidence and pattern, body weight, and food consumption (in dietary studies) on the test animals, and to make sure that the control animals are subjected to the same influences (e.g., gavage or injection) as the drug treated animals, so that all animals will be subject to equal physiological response and stress (i.e., to isolate the treatment effect from other possible effects) (Gart et al. 1986; Dayan 1988).

There are arguments for and against using two identical control groups in a study (Category B). The arguments for this design are that the results from the two identical controls can be used as a mechanism for identifying the extent of control variability (Gart et al. 1986) and the results can be used to help evaluate the biological significance of increases in tumor incidence in the treated groups (i.e., true increases versus noise). From the biological perspective, the dual control data can be viewed as equivalent to having contemporary historical data. In this case, consideration of other appropriate historical control data is essential if the results with the two contemporary controls are different. As described below, however, there may be difficulties in statistical analysis of data from a study using this design.

Statisticians and pharmacologists/toxicologists should decide collaboratively which of the two control groups is appropriate for the analysis of data from a study in Category A. Ordinarily, analyses of data of the vehicle control and the treated groups are the most meaningful assessment of drug effect. Even in this case, however, the untreated control can give information about spontaneous variability. There are other situations in which three analyses – control 1 versus treated groups, control 2 versus treated groups, and control 1 plus control 2 versus treated groups – are performed. Because concerns about the possible effects of the vehicle substance on the test animals are the reason for using the vehicle control in addition to the untreated control, it is also of interest to compare the mortality, tumor rates, body weight, and food consumption (in dietary studies) between the two control groups.

Data from dual identical control groups may or may not be combined for statistical analysis of data. If comparisons of the controls for Category B studies show no large differences in mortality and tumor rate, the data from the two control groups are usually combined to form a single control group in subsequent analyses (Haseman et al. 1990). If the data show evidence of differences in mortality or tumor incidence between the identical controls, three tests – control 1 versus treated groups, control 2 versus treated groups, and control 1 plus control 2 versus treated groups – for each tumor/organ combination should be carried out.

In the second case, the question of how to interpret the results of a study in Category B can be approached from two perspectives. First, a trend or a difference in tumor rate could be considered significant only if it is significant for both of the controls. The basis for this conclusion would be that a real finding should be reproducible. Alternatively, the trend or the difference in tumor rate between groups could be considered significant as long as any one of the three tests (i.e., drug vs. control 1, drug vs. control 2, and drug vs. pooled control) either control and pooled control shows a significant result, assuming that most carcinogenicity studies are relatively under powered. The first approach is conservative in the sense that the null hypothesis will be rejected less often. The second approach, on the other hand, will result in an increased false positive rate.

Currently, no good information exists about how to appropriately adjust the significance levels for the above two approaches to maintain the 10 percent overall false positive rate used by the Center. In general, the test result could be regarded as providing only equivocal evidence of a positive finding unless all the three tests yield consistent results (i.e., all statistically significant or all not statistically significant) (Haseman et al. 1990). In such instances, from a biological prospective it is particularly important to evaluate the control response relative to a historical control.

V. INTERPRETATION OF STUDY RESULTS

Interpreting results of carcinogenicity experiments is a complex process, and there are risks of both false negative and false positive results. The relatively small number of animals used and

low tumor incidence rates can result in the failure to detect the carcinogenicity of a drug (i.e., a false negative). Because of the large number of comparisons involved (usually 2 species, 2 sexes, and 30 or more tissues examined), a great potential exists for finding statistically significant positive trends or treatment-placebo differences due to chance alone (i.e., a false positive). Therefore, it is important that an overall evaluation of the carcinogenic potential of a drug take into account the multiplicity of statistical tests of significance for both trends and pairwise comparisons. The evaluation should also make use of historical information and other information related to biological relevance (e.g., positive findings at the same site in the other sex and/or in the other species, and evidence of increased preneoplastic lesions at the target organs/tissues).

A. Adjustment for the Effect of Multiple Tests (Control Over False Positive Error)

It is well known that, for a multi-group study (e.g., 3 doses and placebo), trend tests are more powerful (i.e., more likely to detect an effect) than pairwise comparisons. Tests for trend instead of pairwise comparison tests between control and high-dose groups are therefore the primary tests in the evaluation of drug related increases in tumor rate.

Statistical and nonstatistical procedures have been proposed for controlling the overall false positive rate. Surveys of some of those procedures can be found in Lin and Ali (1994), and Fairweather et al. (1998). In this guidance document, only the statistical decision rules for controlling the overall false positive rates associated with trend tests and pairwise comparisons used by the Center in interpreting the final results of carcinogenicity studies are discussed. The decision rules were developed based on historical control data of CD rats and CD mice (strains that are most widely used in studies of pharmaceuticals) to achieve an overall false positive rate of around 10 percent for the standard two-species, two-sex in-vivo studies and the alternative ICH one-species, two-sex in-vivo studies.

In the past, statisticians in CDER used the statistical decision rule described in Haseman (1983) in tests for significance of trends in tumor incidence. The decision rule was originally developed for pairwise comparison tests in tumor incidence between the control and the high-dose groups and was derived from results of carcinogenicity studies conducted at National Toxicology Program (NTP). Strains of Fischer 344 rats and B6C3F1 mice were used in the NTP studies. Like most studies of pharmaceuticals, four treatment/sex groups with 50 animals in each group were used in the NTP studies. All of the NTP studies lasted for 2 years. The decision rule tests the significant differences in tumor incidence between the control and the dose groups at 0.05 level for rare tumors and at 0.01 level for common tumors. A tumor type with a background rate of 1 percent or less is classified as rare by Haseman; more frequent tumors are classified as common. Haseman's original study and a second study using more recent data with higher tumor rates show that the use of this decision rule in the control-high pairwise comparison tests would results in an overall false positive rate between 7 to 8 percent and between 10 to 11 percent, respectively (Haseman 1983, 1984a, 1991).

Concerns have been raised that applying the rule described by Haseman (1983) to analyses of trend tests would lead to an excessive overall false positive error rate as data from all treatment groups are used in the tests and considerably lower tumor rates can yield a wrongly significant result. Results from recent studies within and outside FDA show that this concern is valid. Based on studies conducted by CDER and NTP, the overall false positive error resulting from interpreting trend tests by use of the above decision rule is about twice as large as that associated with control-high pairwise comparison tests.

Based on recent studies using real historical control data of CD mice and CD rats from Charles River Laboratory and simulation studies conducted internally and in collaboration with NTP, a new statistical decision rule for tests for a positive trend in tumor incidence has been developed. This new decision rule tests the positive trend in incidence rates in rare and common tumors at 0.025 and 0.005 levels of significance, respectively. The new decision rule achieves an overall false positive rate of around 10 percent in a standard two-species and two-sex study (Lin 1995, 1997; Lin and Rahman 1998a, 1998b). The 10 percent overall false positive rate is seen by CDER statisticians as appropriate in a new drug regulatory setting.

Regulatory statistical literature emphasizes methods for testing for positive trends in tumor rate (Lin 1988, 2000); Lin and Ali 1994; Chen and Gaylor 1986; Dinse and Haseman 1986; and Dinse and Lagokos 1983). There are situations, however, in which pairwise comparisons between control and individual treated groups may be more appropriate than trend tests because trend tests assume that a carcinogenic effect is related to doses or systemic exposure weights, or ranks. The assumption may be true for simple direct acting carcinogens in studies not complicated by excessive toxicity. However, there are many cases in which the response is to a drug metabolite, is mediated through a receptor (or enzyme) that may be saturated even at the low dose, is compounded by dose-related toxicity, or is complicated by other nonlinear effects. Under those situations, pairwise comparisons may be appropriate and the decision rule described in Haseman (1983) should be used in interpreting the results of the pairwise comparison tests.

Sponsors should conduct both trend tests and pairwise comparison tests and present the results of these two types of tests in the formats used in Table 15. A recent complication to the use of the trend test is the choice by a sponsor not to do histopathologic evaluation of all treatment groups. Although studies conducted using this design have been evaluated by CDER, such an approach is not usually recommended.

The high cost (between 1 and 2 million dollars) and long time (a minimum of 3 years) it takes to conduct a standard long-term, in-vivo carcinogenicity study and the increased insight into the mechanisms of carcinogenicity provided by advances in molecular biology have led to alternative in-vivo approaches to the assessment of carcinogenicity. The International Conference on Harmonization (ICH) has developed guidance for use in

the United States and in other regions entitled *S1B Testing for Carcinogenicity of Pharmaceuticals* (1998). This guidance outlines experimental approaches to the evaluation of carcinogenic potential that may obviate the need for the routine use of two long-term rodent carcinogenicity studies and allows for the alternative approach of conducting one long-term rodent carcinogenicity study together with a short- or medium-term rodent test. The short- or medium-term rodent test systems include such studies as initiation-promotion in rodents, transgenic rodents, or newborn rodents, which provide rapid observation of carcinogenic endpoints in-vivo. In general, these studies do not produce false positive results because tumor background rates are very low. False positives therefore arise primarily from the 2-year rodent study. Results from an agency study using historical control data of CD rats and CD mice (Lin 1997; Lin and Rahman 1998b) showed that the use of significance levels of 0.05 and 0.01 in tests for positive trend in incidence rates of rare tumors and common tumors, respectively, will result in an overall false positive rate around 10 percent in a study in which only one 2-year rodent bioassay (plus the shorter rodent study) is conducted.

The decision rules for testing positive trend or differences between control and individual treated groups in incidence rates for standard studies using two species and two sexes as well as studies following ICH guidance and using only one 2-year rodent bioassay are summarized in Table 13.

The developed decision rules for tests for positive trend and for difference in pairwise comparisons are based on the proposition that the carcinogenic effect of a drug is considered positive if one or more tumor types tested in any of the four experiments (or two experiments under an alternative ICH study) of species/sex combination show a significant positive trend in tumor incidence rates (or one or more tumor types show a significant difference in tumor incidence rates when the results of the control-high pairwise comparisons are used in the final interpretation). The decision rules were developed assuming the use of the two-species-and-two-sex (or one-species-and-two-sex) standard design of a two-year study with 50 animals in each of the four treatment/sex groups.

Table 13: Statistical Decision Rules for Controlling the Overall False Positive Rates Associated With Tests for Positive Trend or With Control-High Pairwise Comparisons in Tumor Incidences to Around 10 Percent in Carcinogenicity Studies of Pharmaceuticals

	Tests for Positive Trend	Control-High Pairwise Comparisons
Standard 2-Year Studies with 2 Species and 2 Sexes	Common and rare tumors are tested at 0.005 and 0.025 significance levels, respectively.	Common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively.
Alternative ICH Studies (One Two-Year Study in One Species and One Short- or Medium-Term Study, Two Sexes)	Common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively.	Under development and not yet available.

B. Control Over False Negative Error

To make sure that the false negative rate is not excessive, reviewing pharmacologists, pathologists, and medical officers evaluate the adequacy of the gross and histological examination of both control and treated groups, the adequacy of the dose selection, the duration of the experiment in relation to the normal life span of the tested animals, and the survival of animals in the study.

C. The Use of Historical Control Data

The concurrent control group is always the most appropriate and important in testing drug related increases in tumor rates in a carcinogenicity experiment. However, if used appropriately, historical control data can be very valuable in the final interpretation of the study results. Large differences between studies can result from differences in nomenclature, pathologists reading slides, the specific animal strain used and laboratory conditions. It is therefore extremely important that the historical control data chosen be from studies comparable to the current study, generally recent studies from the same laboratory using the same strain of rodent.

Historical control data are particularly useful in classifying tumors as rare or common. A statistically significant increase in a rare tumor is unlikely as a chance occurrence so that it is critical to decide whether a tumor is rare or not. Rare tumors are generally tested with less stringent statistical decision rules (see Table 13). Historical control data can also be used as a quality control mechanism for a carcinogenicity experiment by assessing the reasonableness of the spontaneous tumor rates in the concurrent control

group (Haseman 1984b; Haseman, Huff, and Boorman 1984), and for evaluation of disparate findings in dual concurrent controls.

For common tumors, in cases of marginally significant trends or differences, historical control data can help investigators determine whether the findings are real or false positives. Historical control data can also help investigators determine whether nonsignificant findings in rare tumors are true negative or false negative results due to the lack of power in the statistical tests used. A widely used informal method is to determine whether the tumor rates in the treated groups in the experiment are within the range of reliable historical control data. If they are, a marginally significant finding for a common tumor may be discounted as resulting from a random occurrence of a low concurrent control rate. Similarly, a nonsignificant increase for a rare tumor can be considered truly negative if the treated tumor rates are within the historical range.

The above informal method of using historical control data in the interpretation of statistical test results is not very satisfactory because the range of historical control rates is usually too wide. This is especially true in situations in which the historical tumor rates of most studies used are clustered together, but a few other studies give rates far away from the cluster. When the range of historical control data is simply calculated as the difference between the maximum and the minimum of the historical control rates. The range does not consider the shape of the distribution of the rates. The upper confidence intervals for binomial proportions constructed by the methods described in Louis (1981), Blyth (1986), Vollset (1993), Jovanovic and Viana (1996), and Jovanovic and Levy (1997) should probably replace the historical range in the above informal method.

In addition to the informal use of historical control data in the interpretation of the statistical testing results mentioned above, more formal statistical procedures have been proposed that allow the incorporation of appropriate historical control data in tests for trend in tumor rate. For example, Tarone (1982), Hoel (1983), Hoel and Yanagawa (1986), Tamura and Young (1986, 1987), and Prentice et al. (1992) proposed some empirical procedures using the beta-binomial distribution to model historical control tumor rates and to derive approximate and exact tests for trend. The results from those studies show that the incorporation of the historical control data improves the power of the tests. The greatest improvement of power is shown in the tests of rare tumors. Dempster et al (1983) proposed a Bayesian procedure to incorporate historical control data into statistical analysis. The procedure uses the assumption that the logits of the historical control tumor rates were normally distributed.

These formal statistical procedures work well in situations in which historical data from a large number of studies with relatively large control groups are available to provide reliable estimations of the parameters of the prior distributions. However, the maximum likelihood estimators (MLEs) of the prior parameters were shown to be unstable, and the distributions of the MLEs were skewed to the right (i.e., with bunching above the mean and a long tail below the mean). The skewness was severe in cases in which only

historical data of a few small control groups were available. The skewness of the MLEs inflated the Type I error of the tests. Also, these procedures were developed to incorporate historical control data into the Cochran-Armitage test for linear trend in tumor incidence. Since the Cochran-Armitage test is a survival-unadjusted procedure, these procedures cannot be applied to studies with significant differences in survival among treatment groups. Recently, Ibrahim and Ryan (1996) developed a method for incorporating historical control information into survival-adjusted tests for trend in tumor incidence. When using this method, the study period should be partitioned into intervals. In each interval, the multinomial distribution should be used to model the observed numbers of animals dying with the tumor, and Dirichlet distribution should be used as the prior distribution for the historical control tumor rates. This method applies only to fatal tumors.

D. Evaluation of Validity of Designs of Negative Studies

In negative or equivocal studies, that is, studies for which either the sponsor's or FDA's statisticians detected no statistically significant positive trend or difference in tumor rate, the statistical reviewers will perform a further evaluation of the validity of the design of the experiment to see if there were sufficient numbers of animals living long enough to provide adequate exposure to the chemical and to be at risk of forming late-developing tumors. The reviewers also want to see if the doses used were adequate to present a reasonable tumor challenge to the tested animals (Haseman 1985).

As a rule of thumb, a 50 percent survival rate of the 50 initial animals in any treatment group between weeks 80 to 90 of a 2-year study would be considered to yield a sufficient number of animals with adequate exposure. The percentage can be lower or higher if the number of animals used in each treatment/sex group is larger or smaller than 50, but between 20 to 30 animals should be still alive during these weeks.

The adequacy of doses selected and of the animal tumor challenge in long-term carcinogenicity experiments is evaluated by pharmacologists and the CDER Carcinogenicity Assessment Committee (CAC) based on the previously described ICH approaches as well as on the results of the long-term carcinogenicity experiments. To assist the evaluation, CDER statistical reviewers are often asked to provide analyses of body weight and mortality differences and, occasionally, other differences between treated and control groups.

VI. PRESENTATION OF RESULTS AND DATA SUBMISSION

To facilitate the statistical reviews, sponsors should present study results and data in such a way that FDA statistical reviewers are able to verify the sponsors' calculations, to validate their statistical methods, and to trace back the sponsors' conclusions through their summaries and analyses of the raw data (FDA 1987).

In the sponsor's report, in addition to the volumes containing study data of individual animals, a statistical analysis section should be included containing summary statistics of the study data, results of statistical analyses of the data, results and findings, and main conclusions of the study. In the statistical analysis section, the sponsor should include descriptions of the statistical procedures used and pertinent literature references. The descriptions of statistical methodology and references are particularly important if the sponsor decides to use designs and methods of analysis and interpretation other than those recommended in this guidance document.

Tables 14, 15, and 16 include examples of formats for presenting summaries and results of analyses of survival and tumor data. Presentations of data summaries and analyses results should be made for each species/sex combination. Descriptive statistics such as mean, standard deviation, and range, which are important in characterizing the distinctive and essential features of a study, should also be reported by species/sex combination. Graphics that are useful and informative in presenting study results should be used to display summary data, especially summary statistics over time.

Two sets of formats and specifications were previously used regularly by the Divisions of Biometrics, Office of Biostatistics. They were (a) the Divisions of Biometric Formats and Specifications for Submission of Animal Carcinogenicity Study Data, and (b) the Submitters Toxicological Uniform Data Information Exchange Standard (STUDIES). Because mistakes have often been made by sponsors in data sets using the STUDIES formats, the Office of Biostatistics now recommends that sponsors submit the data sets in the simpler divisions of biometrics formats and specifications described in Lin (1998). Discussions of the statistical analyses on which the formats were developed can be found in Lin (1998).

Data sets described in the above Divisions of Biometrics formats and specifications document are divided into two groups, Group A and Group B, depending on whether the data will be used immediately in the statistical review and evaluation of the carcinogenicity studies. Group A includes data sets that are always used by statisticians performing a statistical review and evaluation of the carcinogenicity studies. Group B includes data sets that may be used by medical officers, pharmacologists, toxicologists, and statisticians in their final interpretations of the study results. Sponsors are urged to submit the two groups of data sets together with their original, initial submissions of the hardcopy NDA or IND. However, if a sponsor under some special circumstances cannot submit the two groups of data sets together, the Group A data sets should be submitted first.

The FDA has issued a guidance document (1999) to encourage and assist drug applicants in submitting an electronic archival copy of a new drug applications (NDAs), including amendments and supplements. The Agency's effort to encourage applicants to submit applications electronically is an integrated part of the Agency's Electronic Records; Electronic Signatures regulation (Electronic Records; Electronic Signatures, Office Federal Register, March 20, 1997). The above submission of data from carcinogenicity studies to statisticians should be a part of an electronic NDA. The information in the formats and specifications, discussed above, have been incorporated into the Agency's guidance on regulatory submission of electronic applications (FDA 1999). Drug sponsors should follow the guidance and recommendations

included in the nonclinical pharmacology and toxicology section of the guidance in their preparation and submission of electronic carcinogenicity study data.

Table 14: Example Format for Showing Summary of Deaths and Sacrifices of Male Mice

	Control					Low					Medium					High				
Week	E	D	S	N	NP	E	D	S	N	NP	E	D	S	N	NP	E	D	S	N	NP
34	70	--	--	--	--	70	--	--	--	--	70	--	--	--	--	70	--	--	--	--
35	70	1	1	1	1	70	--	--	--	--	70	--	--	--	--	70	--	--	--	--
36	68	--	--	--	--	70	1	--	--	1	70	--	--	--	--	70	--	--	--	--
39	68	--	--	--	--	69	--	--	--	--	70	--	--	--	--	70	1	--	1	--
41	68	--	--	--	--	69	--	--	--	--	70	1	--	--	1	69	--	--	--	--
43	68	--	--	--	--	69	--	--	--	--	69	1	--	1	--	69	--	--	--	--
49	68	--	--	--	--	69	--	--	--	--	68	--	--	--	--	69	1	--	--	--
52*	68	--	10	10	--	69	--	10	10	--	68	--	10	10	--	68	--	10	10	--
53	58	1	--	1	--	59	--	--	--	--	58	--	--	--	--	58	--	--	--	--
58	57	--	--	--	--	59	--	--	--	--	58	3	--	3	--	58	--	--	--	--
62	57	1	--	1	--	59	--	--	--	--	55	--	--	--	--	58	1	--	1	--
65	56	--	--	--	--	59	--	--	--	--	55	--	--	--	--	57	1	--	1	--
70	56	--	--	--	--	59	--	--	--	--	55	3	--	3	--	56	1	--	1	--
71	56	--	1	1	--	59	1	1	2	--	52	1	--	1	--	55	--	--	--	--
(Continue to the end of the study)																				
Term*	41	2	39	41	--	40	--	40	40	--	36	--	36	--	--	38	1	37	36	2
Mean survival	668					680					650					632				

Notes: E = Number of animals entering the period; D = Deaths; S = Sacrificed moribund; N = At least one tissue was examined microscopically; NP = No tissues were examined microscopically; * = Scheduled and terminal sacrifices.

Table 15: Example Format for Showing Summary of Incidences and Results of Statistical Tests (P-values) of Neoplastic Lesions (in Male Mice)

Organ/Tissue And Tumor	Control	Low	Medium	High
Number of animals at the beginning	50	50	50	50
Liver	(40)	(45)	(50)	(43)
Hepatocellular adenoma	4	5	7	10
(Context of obser. of the tumor)[#]				
Unadjusted P-values[##]:				
Exact test	P=	P=	P=	P=
Asymptotic test	P=	P=	P=	P=
Hepatocellular carcinoma	2	2	5	3
(Context of obser. of the tumor)[#]				
Unadjusted P-values[##]:				
Exact test	P=	P=	P=	P=
Asymptotic test	P=	P=	P=	P=
Hemangioma	0	0	2	3
(Context of obser. of the tumor)[#]				
Unadjusted P-values[##]:				
Exact test	P=	P=	P=	P=
Asymptotic test	P=	P=	P=	P=
Hepatoma	0	1	1	2
(Context of obser. of the tumor)[#]				
Unadjusted P-values[##]:				
Exact test	P=	P=	P=	P=
Asymptotic test	P=	P=	P=	P=

See the footnotes on next page.

Table 15 (Continued): Example Format for Showing Summary of Incidences and Results of Statistical Tests (P-values) of Neoplastic Lesions (in Male Mice)

Organ/Tissue And Tumor	Control	Low	Medium	High
Lung	(45)	(47)	(49)	(45)
Bronchiolar/alveolar adenoma (Context of obser. of the tumor)[#]	2	1	4	8
Unadjusted P-values[##]:				
Exact test	P=	P=	P=	P=
Asymptotic test	P=	P=	P=	P=
Bronchiolar/alveolar carcinoma (Context of obser. of the tumor)[#]	2	2	5	4
Unadjusted P-values[##]:				
Exact test	P=	P=	P=	P=
Asymptotic test	P=	P=	P=	P=
(List the numbers of animals with the tissues examined, overall tumor incidences, and the p-values of trend tests and pairwise comparisons for all organs/tissues and tumors.)				

Notes: Numbers in parentheses are numbers of animals with the tissues examined microscopically.
The p-values under the control group are from trend tests.
The p-values under each dosed group are from pairwise comparisons between that dosed group and the control group.
[#]Contexts of observation of the tumor, if information is available, should be one of the four possibilities: fatal, incidental, mortality independent, and mixture of fatal and incidental. Use N.A. to indicate that the information is not available.
[##]Unadjusted P-values are the p-values unadjusted for effect of multiple tests.

Table 16: Example Format for Showing Historical Control Data (in Male Rats)

The historical control data are based on the carcinogenicity studies conducted at XYZ Laboratory between 1995 and 2000.

Species: Mouse, Sex: Male, Strain: Crl:CD-1 Mice

	Historical Control Incidences			
Studies	Tumor type 1	Tumor type 2	. . .	Tumor type T
Study #1 (1992)	1/49	4/49	. . .	8/50
Study #2 (1992)	1/50	3/50	. . .	4/50
.	.	.	. . .	.
.	.	.	. . .	.
.	.	.	. . .	.
Study #n (1996)	0/50	2/50	. . .	5/50
Total	2/347	23/417	. . .	34/417
Standard Deviation	1.0%	3,2%	. . .	4.0%
Range	0%-2%	0%-10%	. . .	3%-17%

REFERENCES

Ahn, H., and R.L. Kodell (1995), "Estimation and Testing of Tumor Incidence Rates in Experiments Lacking Cause-of-Death Data," Biometrical Journal, 37, 745-763.

Ahn, H., and R.L. Kodell, and H. Moon (2000), "Attribution of Tumor Lethality and Estimation of Time to Onset of Occult Tumors in the Absence of Cause-of-Death Information," Applied Statistics, 49, 157-169.

Ali, M.W. (1990) "Exact Versus Asymptotic Tests of Trend of Tumor Prevalence in Tumorigenicity Experiments: A Comparison of P-values for Small Frequency of Tumors," Drug Information Journal, 24, 727-737.

Armitage, P. (1955), "Tests for Linear Trends in Proportions and Frequencies," Biometrics, 11, 375-386.

Armitage, P. (1971), Statistical Methods in Medical Research, John Wiley, New York.

Bailer, A., and C. Portier (1988), "Effects of Treatment-Induced Mortality on Tests for Carcinogenicity in Small Samples," Biometrics, 44, 417-431.

Bannasch, P., R.A. Griesemer, F. Anders, B. Becker, J.R. Cabral, G.D. Porta, V.J. Feron, D. Henschler, N. Ito, R. Kroes, P.N. Magee, B. McKnight, U. Mohr, R. Montesano, N.P. Napalkov, S. Nesnow, A.E. Pegg, G.N. Rao, V.S. Turusov, J. Wahrdrendorf, and J. Wilbourn (1986), "Long-Term Assays for Carcinogenicity in Animals," in Long-Term and Short-Term Assays for Carcinogens: A Critical Appraisal, Editors, R. Montesano, H. Bartsch, H. Vainio, J. Wilbourn, and H. Yamasaki, IARC Scientific Publications No. 83, Lyon, France.

Berlin, B., J. Brodsky, and P. Clifford (1979), "Testing Disease Dependence in Survival Experiments with Serial Sacrifice," Journal of American Statistical Association, 74, 5-14.

Bieler, G.S., and R.L. Williams (1993), "Ratio Estimates, the Delta Method, and Quantal Response Tests for Increased Carcinogenicity," Biometrics, 49, 793-801.

Blyth, C.R. (1986), "Approximate Binomial Confidence Limits," Journal of American Statistical Association, 81, 843-855.

Breslow, N. (1970), "A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship," Biometrics, 57, 579-594.

Chen, J.J., K.K. Lin, M.F. Huque, and R.B. Arani (2000), "Weighted P-Value for Animal Carcinogenicity Trend Test," Biometrics, 56, 586-592.

Chen, J.J., and D.W. Gaylor (1986), "The Upper Percentiles of the Distribution of the Logrank Statistics for Small Numbers of Tumors," Communications in Statistics - Simulation and Computation, 15, 991-1002.

Chu, K.C., C. Cueto, and J.M. Ward (1981), "Factors in the Evaluation of 200 National Cancer Institute Carcinogen Bioassays," Journal of Toxicology and Environmental Health, 8, 251-280.

Cochran, W. (1954), "Some Methods for Strengthening the Common χ^2 Tests," Biometrics, 10, 417-451.

Cox, D.R. (1959), "The Analysis of Exponentially Distributed Life-times with Two Types of Failures," Journal of Royal Statistical Society, Series B, 21, 4121-421.

Cox, D.R. (1972), "Regression Models and Life Tables (with discussion)," Journal of Royal Statistical Society, Series B, 34, 187-220.

Dayan, A.D. (1988), "Biological Assumptions in Analysis of the Bioassay," in Carcinogenicity, The Design, Analysis and Interpretation of Long-Term Animal Studies, an International Life Sciences Institute (ILSI) monograph, edited by H,C. Grice and J.L. Ciminera, Springer-Verlag, New York.

Dempster, A.P., M.R. Selwyn, and B.J. Weeks (1983), "Combining Historical and Randomized Controls for Assessing Trends in Proportions," Journal of the American Statistical Association, 78, 221-227.

Dewanji, Anup and J.D. Kalbfleisch (1986), "Nonparametric Methods for Survival/Sacrifice Experiments," Biometrics, 42, 325-341.

Dinse. G.E. (1988), "Estimating Tumor Incidence Rates in Animal Carcinogenicity Experiments," Biometrics, 44, 405-415.

Dinse, G.E. (1991), "Constant Risk Differences in the Analysis of Animal Tumorigenicity Data," Biometrics, 47, 681-700.

Dinse, G.E. (1994), "A Comparison of Tumor Incidence Analyses Applicable in Single-Sacrifice Animal Experiments," Statistics in Medicine, 13, 689-708.

Dinse, G.E., and J.K. Haseman (1986), "Logistic Regression Analysis of Incidental-Tumor Data from Animal Carcinogenicity Experiments," Fundamental and Applied Toxicology, 6, 751-770.

Dinse, G.E., and S.W. Lagokos (1983), "Regression Analysis of Tumor Prevalence Data," The Journal of the Royal Statistical Society, Series C, 32, 236-248.

Fairweather, W.R., A. Bhattacharyya, P.P. Ceuppens, G. Heimann, L.A. Hothorn, R.L. Kodell, K.K. Lin, H. Mager, B.J. Middleton, W. Slob, K.A. Soper, N. Stallard, J. Ventre, and J. Wright

(1998), “Biostatistical Methodology in Carcinogenicity Studies,” Drug Information Journal, 32, 401-421.

Food and Drug Administration (FDA), (1987), Guideline for the Format and Content of the Nonclinical/Pharmacology/Toxicology Section of An Application.

FDA (1997), Formats and Specifications for Submission of Animal Carcinogenicity Study Data, Divisions of Biometrics I, II, III, and IV, Center for Drug Evaluation and Research. Rockville, Maryland; March 12, 1997.

FDA (1999), Guidance for Industry, Providing Regulatory Submissions in Electronic Formats --- NDAs, Center for Drug Evaluation and Research.

Gart, J.J., D. Krewski, P.N. Lee, R.E. Tarone, and J. Wahrendorf (1986), Statistical Methods in Cancer Research, Volume III - The Design and Analysis of Long-Term Animal Experiments, International Agency for Research on Cancer, World Health Organization.

Gehan, E.A. (1965), "A Generalized Wilcoxon Test for Comparing K Samples Subject to Unequal Patterns of Censorship," Biometrika, 52, 203-223.

Goldberg, K.M. (1985), "An Algorithm for Computing An Exact Trend Test for Multiple 2 x K Contingency Tables," a paper presented at Symposium On Long-Term Animal Carcinogenicity Studies.

Haseman, J.K. (1983), "A Reexamination of False-Positive Rates for Carcinogenesis Studies," Fundamental and Applied Toxicology, 3, 334-339.

Haseman, J.K. (1984a), "Statistical Issues in the Design, Analysis and Interpretation of Animal Carcinogenicity Studies," Environmental Health Perspective, 58, 385-392.

Haseman, J.K. (1984b), “Use of Historical Control Data in Carcinogenicity Studies in Rodents,” Toxicologic Pathology, 12, 126-135.

Haseman, J.K. (1985), "Issues in Carcinogenicity Testing: Dose Selection," Fundamental and Applied Toxicology, 5, 66-78.

Haseman, J.K. (1991), a personal communication to Robert Temple, M.D., CDER, FDA.

Haseman, J.K., J. Huff, and G.A. Boorman (1984), "Use of Historical Control Data in Carcinogenicity Studies in Rodents," Toxicologic Pathology, 12, 126-135.

Haseman, J.K., J.S. Winbush, and M.W. O’Donnell (1986), “Use of Dual Control Groups to Estimate False Positive Rates in Laboratory Animal Carcinogenicity Studies,” Fundamental and Applied Toxicology, 7, 573-584.

Haseman, J.K., G. Hajian, K.S. Crump, M.R. Selwyn, and K.E. Peace (1990), "Dual Control Groups in Rodent Carcinogenicity Studies," in Statistical Issues in Drug Research and Development, K. E. Peace, Editor, Marcel Dekker, New York.

Haseman, J.K. (1999), personal communication to the author.

Hoel, D.G. (1983), "Conditional Two Sample Tests with Historical Controls," in Contributions to Statistics, P.K. Sen, Editor, North-Holland Publishing Company.

Hoel, D.G., and T. Yanagawa (1986), "Incorporating Historical Controls in Testing for a Trend in Proportions," Journal of the American Statistical Association, 81, 1095-1099.

Hoel, D., and H. Walburg (1972), "Statistical Analysis of Survival Experiments," Journal of the National Cancer Institute, 49, 361-372.

Iatropoulos, M.J. (1988), "Society of Toxicologic Pathologists Position Paper: "Blinded" Microscopic Examination of Tissues from Toxicologic or Oncogenic Studies," in Carcinogenicity, The Design, Analysis, and Interpretation of Tong-Term Animal Studies, edited by H.C. Grice and J.L. Ciminera, ILSI Monographs, Spring-Verlag, New York.

ICH (1995), S1C Dose Selection for Carcinogenicity Studies of Pharmaceuticals, ICH - S1C.

ICH (1998), S1B Testing for Carcinogenicity of Pharmaceuticals, ICH - S1B, Federal Register, vol. 63, 8983-8986, 1998.

Ibrahim, J.G., and L.M. Ryan (1996), "Use of Historical Controls in Time-Adjusted Trend Tests for Carcinogenicity," Biometrics, 52, 1478-1485.

Jovanovic, B.D. and M.A.G. Viana (1996), "Upper Confidence Bounds for Binomial Probability in Safety Evaluation," American Statistical Association 1996 Proceedings of the Biopharmaceutical Section, 140-144.

Jovanovic, B.D. and P.S. Levy (1997), "A Look at the Rule of Three," The American Statistician, 51, No.2, 137-139.

Kodell, R.L., and H. Ahn (1996), "Nonparametric Trend Test for the Cumulative Tumor Incidence Rates," Communications in Statistics - Theory and Methods, 25, 1677- 1692.

Kodell, R.L., and H. Ahn (1997), "An Age-Adjusted Trend Test for the Tumor Incidence Rate," Biometrics, 53, 1467-1474.

Kodell, R.L., B.A. Pearce, A. Turturro, and H. Ahn (1997), "An Age-Adjusted Trend Test for the Tumor Incidence Rate for Single-Sacrifice Experiments," Drug Information Journal, 31, 471-487.

Kodell, R.L., K.K. Lin, B.T. Thorn, and J.J. Chen (2000), "Bioassays of Shortened Duration for Drugs: Statistical Implications," Toxicological Sciences, 55, 415-432.

Lin, K.K. (1988), "Peto Prevalence Method Versus Regression Methods in Analyzing Incidental Tumor Data from Animal Carcinogenicity Experiments: An Empirical Study," in the 1988 American Statistical Association Annual Meeting Proceedings (Biopharmaceutical Section), New Orleans, Louisiana.

Lin, K.K. (1995), "A Regulatory Perspective on Statistical Methods for Analyzing New Drug Carcinogenicity Study Data," Bio/Pharam Quarterly, Vol. 1, Issue 2, 18-20.

Lin, K.K. (1997), "Control of Overall False Positive Rates in Animal Carcinogenicity Studies of Pharmaceuticals," presented at 1997 FDA Forum on Regulatory Sciences, December 8-9, 1997, Bethesda, Maryland.

Lin, K.K. (2000), "Carcinogenicity Studies of Pharmaceuticals," in Encyclopedia of Biopharmaceutical Statistics, edited by S.C. Chow, Marcel Dekker, New York, 88-103.

Lin, K.K., and M.W. Ali (1994), "Statistical Review and Evaluation of Animal Tumorigenicity Studies," in Statistics in the Pharmaceutical Industry, Second Edition, Revised and Expanded, edited by C.R. Buncher and J.Y. Tsay, Marcel Dekker, Inc., New York.

Lin, K. K. and M. A. Rahman (1998a), "Overall False Positive Rates in Tests for Linear Trend in Tumor Incidence in Animal Carcinogenicity Studies of New Drugs," Journal of Pharmaceutical Statistics, with discussions, 8(1), 1-22.

Lin, K. K. and M. A. Rahman (1998b), "False Positive Rates in Tests for Trend and Differences in Tumor incidence in Animal Carcinogenicity Studies of Pharmaceuticals under ICH Guidance S1B," unpublished report, Division of Biometrics 2, Center for Drug Evaluation and Research, Food and Drug Administration.

Lin, K.K. (1998), "CDER/FDA Formats for Submission of Animal Carcinogenicity Study Data," Drug Information Journal, 32, 43-52.

Lindsey, J., and L. Ryan (1993), "A Three-State Multiplicative Model for Rodent Tumorigenicity Experiments," Journal of Royal Statistical Society, Series C, 42, 283-300.

Lindsey, J., and L. Ryan (1994), "A Comparison of Continuous- and Discrete-Time Three-State Models for Rodent Tumorigenicity Experiments," Environmental Health Perspectives, 102 (Suppl. 1), 9-17.

Louis, T.A. (1981), "Confidence Intervals for a Binomial Parameter After Observing No Success," The American Statistician, 35, No. 3, 154-154.

Malani, H.M., and J. Van Ryzin (1988), "Comparison of Two Treatments in Animal Carcinogenicity Experiments," Journal of American Statistical Association, 83, 1171-1177.

Mantel, N., and W. Haenszel (1959), "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," Journal of National Cancer Research, 22, 719-748.

Malani, H.M., and Y. Lu (1993), "Animal Carcinogenicity Experiments with and without Series Sacrifices," Communications in Statistics - Theory and Methods, 22, 1557-1584.

McKnight, B., and J. Crowley (1984), "Tests for Differences in Tumor Incidence Based on Animal Carcinogenesis Experiments," Journal of American Statistical Association, 79, 639-648.

Moon, H., H. Ahn, and R.L. Kodell (2000), "Testing Incidence of Occult Tumors by Attributing Tumor Lethality in the Absence of Cause-of-Death Information," submitted to Biometrics.

OFR (Office of the Federal Register), (1985), "Chemical Carcinogens; A Review of the Science and Its Associated Principles," in Part II, Office of Science and Technology Policy, Federal Register, March 14, 1985, 47-58.

OFR, (1995), "Dose Selection for Carcinogenicity Studies of Pharmaceuticals," in Part III, Department of Health and Human Services, Food and Drug Administration, Federal Register, March 1, 1995, Vol. 60, No. 40, 11278-11281.

OFR, (1997), "Electronic Records; Electronic Signatures," Federal Register, March 20, 1997.

Peto, R., M.C. Pike, N.E. Day, R.G. Gray, P.N. Lee, S. Parish, J. Peto, S. Richards, and J. Wahrendorf (1980), "Guidelines for Simple, Sensitive Significance Tests for Carcinogenic Effects in Long-term Animal Experiments," in Long-term and Short-term Screening Assays for Carcinogens: An Critical Appraisal, World Health Organization.

Portier, C.J., and G.E. Dinse (1987), "Semiparametric Analysis of Tumor Incidence Rates in Survival/Sacrifice Experiments," Biometrics, 43, 107-114.

Prasse, K. (1986), "Letter to the Editor (on blinded microscopic evaluation of slides from toxicity and carcinogenicity studies)," Toxicology and Applied Pharmacology, 83, 184-185.

Prentice, R.L., R.T. Smythe, D.Krewski, and M.Mason (1992), "On the Use of Historical Control Data to Estimate Dose Response Trends in Quantal Bioassay," Biometrics, 48, 459-478.

Society for Toxicology (1982), "Animal Data in Harzard Evaluation: Paths and Pitfalls," Fundamental and Applied Toxicology, 2,101-107.

Sontag, J.A., N.P. Page, and U. Saffiotti (1976), Guidelines for Carcinogen Bioassay in Small Rodents, Carcinogenesis Technical Report, DHEW Publication (NIH), 76-801.

Tamura, R.N, and S.S. Young (1986), "The Incorporation of Historical Information in Tests of Proportions: Simulation Study of Tarone's Procedure," Biometrics, 42, 343-349.

Tamura, R.N, and S.S. Young (1987), "A Stabilized Moment Estimator for the Beta-Binomial Distribution," Biometrics, 43, 813-824.

Tarone, R.E. (1975), "Tests for Trend in Life Table Analysis," Biometrika, 62, 679-682.

Tarone, R.E. (1982), "The Use of Historical Control Information in Testing for a Trend in Proportions." Biometrics, 38, 215-220.

Temple, R.T., W.R. Fairweather, V.C. Glocklin, and R.T. O'Neill (1988) "The Case for Blinded Slide Reading," Comments on Toxicology, 2:99-109.

Thomas, D.G., N. Breslow, and J.J. Gart (1977), "Trend and Homogeneity Analyses of Proportions and Life Table Data," Computer and Biomedical Research, 10, 373-381.

U.S. Interagency Staff Group on Carcinogens (1986), "Chemical Carcinogens; A Review of the Science and Its Associated Principles," Environmental Health Perspectives, 67, 201-282.

Vollset, S.E. (1993), "Confidence Intervals for Binomial Proportion," Statistics in Medicine, 12, 809-824.

Williams, P.L., and C.J. Portier (1992), "Analytic Expressions for Maximum Likelihood Estimation in a Nonparametric Model of Tumor Incidence and Death," Communications in Statistics - Theory and Methods, 21, 711-732.

Woodruff, R.S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," Journal of the American Statistical Association, 66, 411-414.