Drug Information Journal, Vol. 32, pp. 401–421, 1998
Printed in the USA.

0092-8615/98

BIOSTATISTICAL METHODOLOGY IN CARCINOGENICITY STUDIES*

WILLIAM R. FAIRWEATHER, PHD
Associate Director, Office of Epidemiology & Biostatistics, Center for Drug Evaluation & Research, Food & Drug Administration, Rockville, Maryland

AMIT BHATTACHARYYA, PHD
Senior Statistician, SmithKline Beecham Pharmaceuticals R&D, Hertfordshire, United Kingdom

PETER P. CEUPPENS, PHD
Zeneca Pharmaceuticals, United Kingdom

GUENTER HEIMANN, DR. RER. NAT.
Corporate Biometry, Schering AG, Berlin, Germany

LUDWIG A. HOTHORN, PHD
Professor, University of Hannover, Hannover, Germany

RALPH L. KODELL, PHD
Division of Biometry & Risk Assessment, National Center for Toxicological Research, Food & Drug Administration, Jefferson, Arkansas

KARL K. LIN, PHD
Division of Biometrics II, Office of Epidemiology & Biostatistics, Center for Drug Evaluation & Research, Food & Drug Administration, Rockville, Maryland

HARRY MAGER, PHD
Head, Preclinical and Clinical Biostatistics, Center of Drug Research, Biometry, Bayer AG, Wuppertal, Germany

BRIAN J. MIDDLETON
Zeneca Pharmaceuticals, United Kingdom

WOUT SLOB, PHD
National Institute for Public Health and Environment (RIVM), Bilthoven, The Netherlands

KEITH A. SOPER, PHD
Senior Research Statistician, Merck Research Laboratories, West Point, Pennsylvania

NIGEL STALLARD, PHD
Medical and Pharmaceutical Statistics Research Unit, University of Reading, Reading, United Kingdom

JOHN VENTRE, MS
Nycomed, United States

JANE WRIGHT, PHD
Quintiles, United Kingdom

Reprint address: William R. Fairweather, Associate Director, Office of Epidemiology & Biostatistics, HFD-700, US Food and Drug Administration, 5600 Fishers Lane, Rockville, MD 20857.

*The views expressed herein are those of the authors and not necessarily those of their employers.

This paper addresses the design, conduct, and statistical analysis of carcinogenicity studies, especially in the context of drug products for human use. It contains suggestions concerning the choice of dose levels, number of animals, methods of slide reading, and the ensuing statistical analysis, focusing on the significance testing approach. The purpose of this document is to describe the current thinking of statisticians and others who work in the area of carcinogenicity studies. The authors represent experience gained in the pharmaceutical industry, regulatory agencies, and academia.

Key Words: Carcinogenicity; Study design; Data analysis; Test statistics; Literature

INTRODUCTION

THIS PAPER ADDRESSES the design, conduct, and statistical analysis of carcinogenicity studies, especially in the context of drug products for human use. In these studies laboratory animals are exposed to a chemical compound during a period of time (usually a large fraction of their lifespan), after which the animals are sacrificed and examined for the presence of tumors. These studies are intended to provide insight into the compound's carcinogenic potential in the experimental species used, as a model system for man.

Usually a carcinogenicity study contains one or more control groups and a limited number of dose groups (typically three or four), so that the observed tumour rates in each group can be related to the dose administered. Depending on the agent's application, the question to be answered may be qualitative: "Does the compound have carcinogenic potential?," or it may have a more quantitative nature: "At what dose level can any carcinogenic effects be considered to be absent or negligible?" For example, in testing drugs, the qualitative question is usually posed, while in food additives, pesticides, and environmental pollutants, the quantitative question is more common (quantitative risk assessment). This difference is largely attributable in the United States to differences in law and the nature of the products involved. For example, it is not acceptable that a product has carcinogenic potential if it is intended for use in foods, but if it is intended for use as a drug its overall benefit and risk must be weighed.

This paper discusses considerations and suggestions concerning the design of carcinogenicity studies (eg, choice of dose levels, number of animals) and the ensuing statistical analysis, focusing on the significance testing approach. Nonetheless, one should keep in mind that tests of significance have only limited value with respect to the original question posed, and therefore, should be interpreted with great care. This is the more so in carcinogenicity studies, where the detectable tumor frequency is typically several orders of magnitude higher than what would be considered as "negligible" or "acceptable" in the human population. The limitations of the significance testing approach of data analysis will also be addressed here. The authors acknowledge that because of these limitations the final conclusions to be drawn from carcinogenicity studies should not depend on statistical considerations alone. Underlying biological mechanisms will affect the interpretation of statistical findings. Quantitative risk assessment by regression analysis, aimed at deriving low-risk dose levels for nonthreshold agents, is outside the scope of this paper.

Rodent carcinogenicity studies are deceptively complex. In theory, the experimental design and response variables are well understood from years of use, and there are numerous statistical methods available to summarize and analyze the data. In practice, it is often difficult to conclude on the basis of the study alone that a drug or other chemical is or is not carcinogenic: The doses administered may be insufficient for producing a tumor response or may produce life shortening toxicity which prevents observation of late-developing tumors. The events of most concern from a clinical perspective are apt to occur infrequently. There are multiple theories of tumor formation, and some imply a carcino-

genic finding while others suggest that tumors derive from the toxicity of high doses themselves. Unlike other areas of pharmaceutical science, there is seldom independent replication of a carcinogenicity study. This implies that both positive and negative study results must be interpreted cautiously.

A carcinogenicity study is not a confirmatory study of a suspected outcome, such as a Phase III clinical trial. It is conducted as a screen, with a large set of potential outcomes of the type of interest, almost all of which could be listed in advance. It is possible, even likely, that a "significant" statistical test may result from chance alone. On the other hand, it is also known that corrections that can be made for multiple testing and multiple comparisons could entirely obliterate a weakly positive effect. Thus, the practice of statistical science in this area is very challenging.

The purpose of this document is to describe the current thinking of statisticians and others who work in the area of carcinogenicity studies. The authors represent experience gained in the pharmaceutical industry, regulatory agencies, and academia. It is hoped that the discussion here will highlight the challenges and current efforts to meet those challenges.

DESIGNING AND CONDUCTING THE STUDY

Study Objectives and Limitations

The doses used in carcinogenicity studies are generally higher than those planned for human use, relative to the size or weight of the test species. The probability of observing a carcinogenic response is assumed to increase with dose, length of exposure, and the number of animals exposed. Because of high costs, limited resources, and concerns about the care and use of experimental animals, only relatively small numbers of animals can be justified on a routine basis. Thus, routine carcinogenicity studies employ small numbers of animals at high doses. As a consequence, negative results might reflect only a lack of statistical power to detect real carcinogenic effects. On the other hand, positive results observed using extremely high doses do not predict with certainty that unacceptable risks would be associated with the generally lower human doses. Moreover, in some cases, because of the species differences in metabolism, doses administered may not be directly related to exposure achieved. In most carcinogenicity studies, dose is used as a surrogate for exposure. It is important to determine whether this relationship is linear over the range of doses employed and whether the same relationship holds in humans.

The United States Food and Drug Administration generally separates the evaluation of a drug for carcinogenic potential into two parts:

1. Determine what is the effect on the species studied for the doses administered, and
2. Determine the relevance of these findings to man, considering the nature of the findings, relative pharmacology, and so forth.

This paper is concerned with the first of these issues.

The major limitation of the use of long-term animal experiments to establish the carcinogenic potential of drugs and other chemicals is that direct linkage of carcinogenic processes observed in animals to actual human cancer is rare. There have been instances, however, such as the induction of bladder tumors in dogs exposed to aromatic amines, that have mirrored the human experience. Because most known human carcinogens have been found to be carcinogenic in experimental animals by appropriate testing, this long-term study has become an integral part of the process of ensuring public health in many countries. Nevertheless, there is a great deal of uncertainty in the animal-to-man extrapolation.

The response of interest in these studies is an exposure-related increase in tumor incidence. Because most types of tumors observed in rodent test species are occult by nature, that is, they are not detectable while an animal is still alive, the exact age at onset

for most tumors is not observable. Differential mortality among dose groups from causes other than tumor invalidates the comparison of overall lifetime incidences, in general, and dictates the comparison of age-specific tumor incidence rates (1,2,3). Knowledge of each animal's age at death and the presence or absence of occult tumors, however, is insufficient to establish the identifiability of age-specific tumor incidence rates. Hence, inferences about tumor incidence rates must rely on additional information, such as scheduled sacrifices or cause-of-death data, or they must rely on additional assumptions, such as the likely degree of tumor lethality.

The Need for a Protocol

The underlying reason for conducting a carcinogenicity study according to a predetermined protocol is the same as for any other scientific experiment, that is, to ensure sufficient planning and control, and to enable valid inferences with respect to the primary objective or hypothesis of interest. In a long-term carcinogenicity study, the objective is to determine if there is an increase in tumorigenicity associated with exposure to the test agent, the null hypothesis being that there is no such increase. The experiment is designed and analyzed so as to have sufficient power to detect treatment-related increases in tumor incidence above the background rate. The design should also be able to yield useful data in the event that some of the experiment must be discarded or an unusual number of animals die early. These points will be discussed further below.

Ideally, a level of increase in the tumor incidence rate that is considered biologically relevant would be established *a priori*, and then dose levels and sample sizes would be selected that will give a high probability of achieving statistical significance using a predetermined statistical analysis, if the biologically relevant increase actually exists. It should be clear at the outset how the planned statistical analysis will be used to reach a decision about the null hypothesis, so that any statistically significant results will not be questioned for technical reasons. Given that the design of most studies is now well established (including parallel treatment groups, with three treated groups and at least one zero-dose control, and the usual set of organs to be examined in each animal), it has become standard practice to use 50 animals per group in anticipation of the usually observed mortality. This sample size usually results in approximately 25 animals surviving to the end of the study and in nearly all animals alive when three-fourths of the study is completed. More animals should be employed if interim sacrifices are planned or higher mortality is anticipated.

Statisticians may also have input with respect to the number and spacing of dose levels, the inclusion of scheduled kills, early termination, blind slide reading procedure, and other protocol issues. Current scientific issues, such as the advisability of *ad libitum* feeding versus some level of caloric restriction, can also invite statistical input.

Among the important issues that need to be addressed in the protocol are the identification of specific tissues and organs that will be examined histopathologically and analyzed statistically, whether all groups will be histologically examined initially, and the design of histological examination procedures. Answering questions of biological relevance in the protocol design stage can help to limit the number of tumor types for which statistical tests are conducted, thus aiding in the control of the experimentwise false positive error rate. Even if statistical tests must be done for every tissue and tumor type, when biological information is available it should be used to identify a smaller set of tumor types that are especially relevant and to perform analyses for these "primary hypotheses" separately from all others and with a less extreme correction of the experimentwise false positive error rate. As much as possible, it is desirable for the statistician to collaborate closely with the principal investigator, especially when a "standard" design will not be used.

Dose Selection

Typical approaches to dose selection and dose regimens for carcinogencity studies

were defined by the National Cancer Institute (NCI) in the development of a large scale screening program for industrial chemicals in the 1960s and 1970s (4,5). The estimated maximum tolerated dose (MTD) is used as the high dose in rodent bioassays in order to maximize the ability to detect any carcinogenic effect and to compensate for the small numbers of animals tested and the resulting limitations in statistical power to detect modest drug-related increases in tumor incidence. The MTD is typically estimated or predicted from 90-day dose range-finding studies based on gross evidence of systemic toxicity, including failure to gain weight or body weight decreases and pathologic evidence of cellular/tissue injury. The MTD is the largest dose which results in only a minimal increase in any such overt, nontumorigenic toxicity (6). The NCI and the National Institute of Environmental Health Sciences (NIEHS), to which the carcinogenicity test program was transferred in 1979, utilized a low dose or a mid- and low dose that were ½, ⅓, or ¼ the estimated MTD (4,7).

In the intervening 20 years, there have been numerous advances in identification of molecular and chromosomal changes during carcinogenesis, as well as genotoxic and nongenotoxic (epigenetic) mechanisms of cancer induction. Applicability of the NCI/NIEHS testing paradigm to pharmaceuticals, which have highly specific and potent cellular effects, has been questioned. It is widely believed that chronic tissue injury or perturbations in cellular homeostasis often produced by exceeding the MTD can compromise the interpretation of carcinogenicity studies by producing nonpredictive and irrelevant results, for example, via metabolic overload, cytotoxicity, and so forth (8,9,10).

It is now generally accepted that animal pharmacokinetic/toxicokinetic data (eg, absorption, disposition, metabolism, and elimination) are important, if not essential, to developing appropriate dosing regimens (route and frequency of administration, amount and spacing of dose groups) (11). The International Conference on Harmonization (ICH) has proposed a flexible approach to dose selection for carcinogenicity studies, indicating that pharmacokinetic endpoints, saturation of absorption, pharmacodynamic endpoints, or maximum feasible dose may be as important or more important in determining the high and intermediate doses in carcinogencity studies of pharmaceuticals (12). It has been suggested that the selection of the low dose should yield some multiple of the anticipated human exposure; because an objective of these studies is to provide information for human risk assessment, it is logical for the low dose to be at or slightly above the human exposure, based on area under the curve (AUC) of drug plasma concentration over time. The middle dose (assuming three are used) would then be geometrically located between the low dose and the high dose (13). This assumes a linear relationship holds between dose and exposure and also between log(exposure) and effect.

Alternatively, pharmacokinetic data may be the most relevant data for establishing middle and low doses (11). As with the high dose, a flexible approach to setting the other doses is advised. Ultimately, the experience of the carcinogenicity study itself must be used to assess the adequacy of the doses chosen (see below).

Animal Housing

To eliminate possible confounding effects on tumor occurrences, it is usually recommended that each animal be individually housed, and that cages on shelves be rotated systematically. Studies have shown that, without systematic rotation of cages, the incidence of cataracts and retinopathy are increased in those animals close to a fluorescent light source (14).

The change from single-animal housing to group housing can have significant impact on occurrence of certain types of tumor. In one study the significant dose-related increase in skin tumors was actually not drug related, but was caused by wounds received while fighting, caused by hyperactive behavior of the animals taking the test substance. No increase in skin tumors was observed in other studies of the product in which animals were singly housed.

Study Duration and Early Termination

Animals are usually exposed to the test substance for the majority of their normal life span. The appropriate study duration depends on the species and strain of the animals used and the substance tested. Studies conducted by the National Toxicology Program (NTP) using B6C3F1 mice and F344 rats usually last for two years. Charles River CD mice and CD rats, the species and strains most widely used in studies conducted by pharmaceutical companies, may have shorter life spans than those used in NTP studies. In the past, it was recommended that a mouse experiment should last at least 18 months, and a rat experiment 24 months. The current practice, however, is for a mouse experiment to continue for 24 months unless increased mortality is observed.

In special situations, an experiment could be terminated if the number in the control group is reduced to 10–12 animals early in the study. If the mortality of the high-dose group becomes excessive due to obvious toxicity of the tested substance, the high-dose group could be terminated when the survival of the group is reduced to 10 animals. The other treatment groups in the experiment, whose survival is not adversely affected, should not be terminated prematurely. Part of the control group should be sacrificed if it is intended to analyze the tumor findings of the (terminated) high-dose group. Sacrificing a few animals in the control group, however, does not allow comparisons with much power and may adversely affect power for comparisons of late-developing tumors. Such sacrifices should not be undertaken without consultation with appropriate regulatory officials.

The main reason for the early termination of an entire experiment or the high-dose group under those special situations is to have enough animals available from each group for a thorough pathological evaluation. It is also important, however, that a study should last for the major part of the animal's life span to allow occurrence of late-developing tumors.

A dose group is sometimes considered for early termination due to markedly decreased survival, usually due to unanticipated treatment-related toxicity. When a dose level is so high that severe toxicity or even death is induced, the relevance of any neoplastic response to human risk assessment may be questioned. For this reason, regulatory bodies have long defined the MTD to be a dose not causing decreased survival apart from neoplastic response, and do not recommend testing animals above the MTD (15).

When a top dose group is determined to be above the MTD and the study is still ongoing, the investigator has at least four options:

1. Terminate the dose group without histopathologic examination,
2. Terminate the dose group and do histopathologic examination,
3. Drop the dose to a lower level, perhaps zero, and continue follow-up, and
4. Continue dosing at the same level and continue follow-up.

Terminating a dose group without histopathological examination is generally acceptable only when substantial treatment-related mortality is observed early in the study and the dose group is terminated before there is much likelihood of observing a neoplastic response.

If a dose group is terminated early and histopathologic examination is done on all animals, a subset of the concurrent control group must be terminated at the same time if statistical analysis is to be done. Little beyond descriptive summaries can be done without a comparable control group. The subset of controls should be selected as a formal random sample of control animals alive at the time of the early sacrifice. If possible, the sacrifice week should be selected at a time when historical control data are available, for example, at one year. Early sacrifice of part of the control group, however, implies a reduced number of controls at terminal sacrifice, resulting in decreased statistical power for all remaining dose

groups. Partial sacrifice of the control group is an available option only if the group is much larger than the standard 50 animals. One solution to the problem of decreased statistical power is to terminate all dose groups simultaneously, but that is feasible only if the study is quite close to the scheduled terminal sacrifice. Statistical analyses for males and females are done separately; therefore, there is no statistical need for sacrifice times to be the same.

One strategy to minimize survival differences is to drop the dose to a level below the MTD in the group experiencing reduced survival. If the statistical analysis employs dose or log dose as the trend score, then some choice between the original and the revised dose must be made for that group. If the revised dose is below the dose for the next highest group, perhaps even zero, then even the order of treatment groups may be unclear. In this case it may be preferable to do two statistical analyses, one analysis comprised of the control and all doses which were found to be at or below the MTD, another separate analysis comprised of the control and the dose group(s) originally above the MTD, but using the revised dose as the trend score.

Interim Sacrifices

In some carcinogenicity studies it may be useful to sacrifice some of the animals in the course of the study, rather than to wait until the scheduled end and to sacrifice the surviving animals. Such study designs may arise for several reasons (16). If one is interested in the tumor onset time (rather than in the time to death from tumor), but expects a rather low lethality during the study, one could sacrifice a certain number of animals at specified time points to estimate the tumor incidence rates. This design may also help if one is interested in a certain tumor type which is not very lethal and typically has early onset. Interim sacrifices then allow one to distinguish between different groups on a time-adjusted basis and not only based upon the crude rates of tumor bearing animals observed at the end of a study.

There are different possibilities to design studies with interim sacrifices. A straightforward method is to specify in advance the sacrifice times, as well as to choose the animals randomly before the start of the study. An alternative would be to randomize at the time of sacrifice. It is important to select the animals randomly, and not to select the moribund cases, since this might bias the data on intercurrent mortality or on lethal tumor types.

False Negatives, False Positives, and Sample Size

Statistical hypothesis testing in carcinogenicity studies usually focuses on testing the null hypothesis, that there is no increase in tumor incidence associated with the test compound against the alternative, that such an increase exists. Two types of error are thus possible. An error of type I, or a false positive, arises from rejecting the null hypothesis when it is true, and an error of type II, or false negative, arises from failing to reject the null when it is false. Because the probability of rejecting the null under the alternative hypothesis depends on the true carcinogenicity, in order to plan sample sizes, doses to employ, and so forth, it is usual to specify the magnitude of a carcinogenic effect above which it is considered important to reject the null hypothesis.

It is usually the case in statistical hypothesis testing that the type I error rate is controlled, commonly fixed at the 5% level, while the type II error rate depends on the sample size chosen and the effect that is to be detected. Given a certain rate of response, a sample size calculation can then be performed to ensure that the false negative rate is sufficiently small, commonly 20%.

By limiting the false positive rate (type I error) the risk to the producer is fixed in advance, while the risk to the consumer from false negatives is determined by the number of animals that can be studied, how long they live on study, the number and spacing of dose groups, the quality control of slide reading, and so forth. In the analysis of carcinogenic-

ity studies where a negative result is indicative of lack of carcinogenicity, this approach corresponds to assuming that a product is safe until it is proven otherwise. Because in this case the consumer risk is associated with false negatives, it has been argued that it is the false negative rate which should be limited *a priori* with the false positive rate determined by the sample size (17–22). Such an argument leads to a different testing approach in which the objective of the study is not to test whether the data are consistent with a lack of any carcinogenic effect, but to assess whether any plausible carcinogenic effect is less than some specified upper limit. The sample size for such a study would be determined so as to limit the false positive rate, that is, to conclude lack of carcinogenicity with high probability if the product is truly noncarcinogenic (23).

Although perhaps desirable, it has not proven feasible to define a level of carcinogenic risk that would be considered acceptable, as required by this approach. Note that elimination with high probability of risks of one in 100 are well beyond what could be accomplished by studies of reasonable size and yet such risks are far too common to be considered "acceptable." Thus, the problem continues to be formulated in the more classical manner.

Use of Two Control Groups

Current practice often includes the use of two control groups or a single, double-sized control group. Dual control groups may receive the same treatment or different treatments, that is, there may be no difference between the control groups (nominal difference) or a real difference.

Although doubling the size of the control group is not the most efficient use of additional animals for a dose-response assessment based only on the current study, it does come closer to the ideal (equal allotment) for purposes of comparing all treated groups versus the control. It is also closer to the ideal for some multiple comparison procedures which emphasize the controls. More importantly, it gives a better internal assessment of how rare each tumor type is, and it provides more accurate comparison to historical control data.

When the two control groups are treated differently (eg, one is untreated and one receives the vehicle), important information on the effects of the vehicle can be obtained. For this purpose, the two controls are compared directly without reference to the groups treated with active ingredient. The dose-response comparison should involve only the vehicle control if there is a significant difference between the two control groups, but may include both control groups if there is no significant difference between them for a given tumor type.

If there is no dose response with the vehicle control group alone but incidence is higher than expected in all groups on the basis of historical experience, an equally high incidence in the untreated control would lead one to conclude that the vehicle is not responsible. It is advisable (or required) that sponsors include both types of controls if there is no information on the carcinogenic potential of the vehicle alone or on its ability to affect that of the active ingredients.

Lastly, the use of nominally different control groups is not recommended. In such designs, the animals are assigned at random to two groups and no difference in treatment, handling, caging, sample preparation, slide reading, and so forth is intended. If a significant difference occurs between these groups, it is not clear what the cause could be. Of course, the difference may represent mere chance. If more than chance is involved, the implication is that there are unknown factors operating on the study and that they have affected the two control groups unequally. It is not known whether the same factors have affected the other treatment groups. Consequently, one might have to repeat the study. Dual controls are treated more fully by Haseman (24). If dual control groups are contemplated, their purpose and the intended analytical approach should be stated in the protocol.

As a variant of the above, it has been

suggested that a second nominal control group and a second nominal high-dose group could be used. The organs of these additional animals would be collected and frozen but tissues would not be prepared for microscopic examination unless there is a significant finding in the main study. This serves to provide an immediately available confirmatory study at modestly increased cost. This approach is rarely used.

Slide Reading

Pathologists generally do not read the slides completely blinded as to treatment group. In some cases, the dose group is apparent by the effect of the stain used in preparing the slides or by the toxic effect of the test substance. In other cases, no effort is made to blind the reader. In order to establish a general understanding of the scope of tissue damage occurring in the study and the appearance of undamaged tissues, the high dose and the controls are often read in whole or in part unblinded. Then the slides are read in blocks of four or eight, with each block containing one or two examples of tissues from each treatment group. All tissues from one organ are read together. All of the tissues of one sex (or all of the tissues from one organ of one sex) are read first.

If the same reader is not able to read all of the slides for the study, he or she should read all slides for a single organ. Infrequently, multiple readers may read each slide and strive for consensus. More frequently, a peer review system is employed after the primary review, and possibly only to resolve disputes or when a significant finding is observed.

Although these procedures are designed to economize on scarce pathology resources, there are obvious problems in credibility of results. Results that appear adverse to the product cannot be dismissed simply because it is known that the slide readings are overly conservative. If regulatory authorities accept the results at face value and must, therefore, restrict or deny approval of the product, sponsors may attempt to reread the slides. The objectivity of this procedure and of the final results may be questioned. If it is necessary to reread the slides, credibility of the results may depend on the ability to perform a blinded and randomized reading.

Statistical analyses should account for multiple readers or multiple reading of slides. It is likely futile to attempt to determine reader effects for these rare events. In any case, the procedures to be employed should be spelled out in the protocol and should conform to current good practice, minimizing biases and providing balance to this aspect of the design wherever possible.

Some pathologists read the low- and mid-dose groups only if significant results are observed between the high-dose and the controls. This practice distorts the false positive and false negative error rates of any subsequent tests for trend. This is an area of current statistical research. The study would almost surely have to be enlarged to provide the same level of statistical confidence. The practice is not acceptable if there is high mortality in the high-dose group or there is inadequate exposure of the high-dose group. It is questionable to employ this procedure in other cases.

DATA ANALYSIS

The discussion here will be limited to the most commonly used methods in carcinogenicity assays of pharmaceuticals, namely those used to analyze rates of tumor-bearing animals and those used to perform time-adjusted analyses. The cases of equal and unequal mortality patterns across the groups will be addressed.

Descriptive Statistics

Descriptive statistics are important in characterizing the distinctive or essential features of the study and should be reported for males and females separately, by treatment group. For continuous variables such as body weight and food consumption, the number of observations, range, arithmetic mean, median, standard deviation, and standard error of the

mean should be sufficient. Other statistics, such as skewness, kurtosis, and percentiles, can be added if the data set is large. For discrete variables such as occurrences of neoplastic and nonneoplastic lesions, the number of animals at risk and examined, the number of animals with each type of lesion, and the observed to expected ratio should be reported. The number expected is calculated in the usual way from the 2xk contingency table for that lesion. The number of tissues autolyzed (unusable) should be provided by tissue type for each animal. Graphics are useful for displaying the individual animal data and summary statistics over time.

Assessment of Challenge

Intercurrent mortality data should be evaluated to see if the survival distributions of the treatment groups are significantly different and whether increased dose is associated with increased mortality from nontumor causes. Adequate food consumption but failure to gain weight may also be indicators of the nearly toxic nature of the (highest) doses administered. Failure to achieve a sufficient challenge can jeopardize the conclusions of a study which does not show a significant increase in tumors with dose. This is especially true if the MTD was not determined using an established procedure.

Analysis of Tumor Rates

The simplest assessment of the effect of dose or treatment on the rate of tumor formation is obtained by ignoring the time of tumor occurrence and the number of deaths from causes other than tumors. Fisher's exact test may be used to compare two groups, and a trend test (25,26), or contrast tests (27) may be used for multiple groups. The number of animals at risk can be the number starting the study or the number alive at the time of the first tumor (15).

For these methods, there are both permutational and asymptotic distributions available. Permutation tests may be very conservative for low tumor rates (10% or less) because of the discrete nature of the distribution, that is, for a predetermined alpha level, there may be no critical value that yields an alpha-level test and the next lower critical value yields a test with far lower alpha level. This problem can be overcome by using unconditional tests to compare the rates (27,28, 29). The mid-p-value approach according to Lancaster (30) is simple, but does not yield a correct alpha-level test (31).

Deaths on study are common, and they modify the number of animals at risk. The effect may be unequal between treatment groups. If there is unequal intercurrent mortality between groups of animals, an age-specific comparison of tumor rates is needed (32). Indeed, some statisticians have argued that an age-specific approach should always be used, regardless of the apparently equal mortality rates (3,33). The authors generally agree with this opinion.

Incorporating Tumor Lethality and Speed of Development

For occult tumors discovered only at necropsy, when precise tumor onset times are unknown, additional information or assumptions are needed. The most commonly available additional information, at least for studies of pharmaceutical compounds, is classification of tumor lethality. For "fatal" tumors (that is, rapidly developing tumors), the onset time may be approximated by the time of death and comparison may generally be made using the log-rank test (34). For "nonfatal" tumors, death is assumed to occur randomly and independently of the presence of tumors, and tumor prevalence (not incidence) rates can be compared using the Mantel-Haenszel test as described by Hoel and Walburg (35). The primary difference in these methods is the definition of the set of animals at risk (1). The special case of infrequent fatal (or palpable) tumors in the presence of unequal intercurrent mortality is addressed later in this section.

When tumor lethality is known, the most common analyses are those described by Peto et al. (33). Animals are divided into

those which are considered to have had fatal tumors and those which are considered to have had nonfatal, or "incidental" tumors. For these analyses, each tumor type is treated separately. Tumor lethality is graded on an ordinal scale ranging from one, if the tumor is definitely incidental, to four, if the tumor is definitely fatal. Usually tumors graded one or two are classified as incidental and those graded three or four as fatal.

The sensitivity of the test outcome to the classification of tumors may be assessed by dichotomizing at different points on the scale and comparing the results obtained. The data may also be analyzed both under the assumption that all tumors are fatal and that all tumors are incidental to assess the impact of this classification for the current study. In practice, most occult tumors are considered to be incidental, that is, that onset times are not necessarily or generally close to death times. Similarly, the periods in which deaths are observed may be perturbed in order to assess the robustness of the outcome to this aspect of the study. Further details are provided, for example, by (15,36,37,38). When information on tumor lethality is unavailable, several new alternatives to the Peto procedure are worth considering (39–42).

If tumor onset times are known more accurately, for example, palpable tumors, an age-specific analysis is available using the tumor onset times in a time-to-event or survival analysis. In this case, tumor onset times may be compared using the log-rank or Gehan's test (43). Prentice and Marek (44) discuss the relative merits of Gehan's test, the logrank test, and related censored data rank statistics. For the sake of brevity, consideration here is limited to the randomly right-censored model and to procedures such as the logrank test or Gehan's test to compare two groups, to corresponding extensions to the k-sample case (45,46), and to corresponding trend tests (47).

Infrequent Fatal or Palpable Tumors

The test statistics in Peto et al. (33) are always referred to their asymptotic distributions to assess significance of the trend in tumor rates. It is tempting to use a permutation approach when the tumor of interest occurs in only a few animals. This is not a straightforward decision, though, and depends upon the pattern of mortality from all other causes (equal across the groups or unequal).

When mortality is equal across treatment groups, the permutation distribution is recommended for assessing the statistical significance of the logrank test, Gehan's test, or corresponding methods. The respective asymptotic versions are applicable if the number of tumor-bearing animals is large (48). Heimann and Neuhaus (49) recommend use of asymptotic distributions when the number of tumor-bearing animals exceeds 10% of the total.

For incidental tumors Ali (50) describes the use of exact permutation tests for assessing trend when the number of tumors is small to moderate. His procedure provides the exact P-value for the test statistic regardless of the underlying intercurrent mortality. Although a similar permutational analysis is often used for fatal or palpable tumors (54), care must be exercised here. The limitations of this approach are described by Heimann and Neuhaus (49), where the arguments follow those of Breslow (45). This permutation distribution considers the margins of all tables comprising the logrank or Peto's test to be fixed. Each table is taken to follow a central hypergeometric distribution under the null hypothesis, and the overall permutation distribution is computed as if the tables were independent. The pattern of fatal (or palpable) tumors with early detection affects the margins at all subsequent times, and the tables are, therefore, not truly independent.

When mortality is unequal across treatment groups, the permutational distribution is no longer the exact conditional distribution for tests of fatal or palpable tumors. The respective asymptotic distributions may still be used, however, if the number of tumor-bearing animals is large (49,51).

Calculation of significance levels for Peto's test or the logrank test is more difficult

when there are unequal mortality patterns among dose groups. In brief, unless the censoring distributions are equal, the permutations of the data are not equally likely, so the hypergeometric distribution does not provide "exact" probabilities for the various tables (52,53). A number of discrete permutation distributions have been proposed (31,49,51, 54,55). At best, they may be considered to be approximations to the correct discrete distribution.

The classical permutation tests (43,45,51) may be obtained by considering all possible assignments of animals to treatments as equally likely, while fixing the rest of the information obtained in the experiment. These tests do not treat the margins of the tables as fixed. Under the null hypothesis this solution is an exact conditional distribution in case of equal intercurrent mortality patterns across the groups, and it is asymptotically correct, in case of unequal mortality patterns (49,51). This permutation distribution may be computed by exhaustive enumeration or by sampling from the set of all permutations by Monte Carlo methods (56,57).

Even with no intercurrent mortality, these different permutational methods produce different P-values. The degree of difference in the intercurrent mortality distributions also affects the quality of the approximation in each case. There is not general agreement on the use of these tests when fatal or palpable tumors occur rarely. Research in this area is continuing.

Trend Test Versus Pair-wise Comparisons

The test statistic should be defined to be sensitive to alternative hypotheses of interest (eg, increasing tumor rate with increased dose) (31). The proportion of animals developing tumors during the course of a long-term study generally increases with dose in case of a compound-related effect (15). In order to evaluate the carcinogenic potential of a compound, it is frequently of interest to test for a generally increasing dose-response trend of the incidence of a tumor type. An additional objective of a carcinogenicity experiment is to compare the rate of affected animals of the control group(s) to that of each dose group (58,59).

For a carcinogenicity study, trend can be defined as a progression (for increasing incidence) of tumor incidence with increasing dose, including the controls as zero dose. For such a trend test, a dose metric has to be chosen depending on biological considerations. Mantel et al. (60) suggested the use of three dose metrics: (a) arithmetic (ie, original dose levels) can be used when a linear dose-response is expected; (b) ordinal (ie, 0, 1, 2, . . .) when a monotone dose-response is expected; and (c) arithmetic-logarithmic (ie, logarithms of the original doses) when a log-linear dose-response is suspected. Park and Kociba (61) expressed concern regarding false positive rates and sensitivity of the dose metric used. They show that false positive errors for Peto's trend test, which is based on the normal approximation, can exceed the nominal level to a varying extent depending on the dose metric used. Their example compares arithmetic and ordinal dose units, with some false positive rates (at the nominal rate of 5%) increasing to as much as 6.5% in the arithmetic units over the ordinal. An alternative is to use multiple contrast tests, for example, in a permutative version (27).

If the experimenter suspects a dose-response other than monotone (eg, due to saturation of absorption or to metabolism) or wishes to determine which doses differ from the concurrent controls, pair-wise tests along with tests for departure from trend can assist in answering these queries. The pair-wise comparisons will test the equality of each dosed group to the control group. This raises the problem, however, of multiple comparisons and increasing the number of false positives. Care in interpretation is required by the experimenter when a lower-dose comparison is significant while the high-dose comparison is not. Consideration can be given to stepwise trend tests satisfying the closure principle (62,63).

It is tempting to read into the pattern of incidence rates more than can be rationally

justified. Biologists may interpret a strictly increasing pattern of incidence rates as the dose increases to be more biologically significant than a pattern containing one or two lower rates as the dose increases, even if both patterns were statistically significant. It should be clear that the trend tests discussed here do not differentiate between such patterns and that a resulting p-value is not more significant if it is accompanied by a strictly increasing set of responses.

Ali (50) and others noted the advantages of tests for trend over homogeneity tests and pair-wise comparisons. If two or more dose levels are studied, statistical tests for positive trend with respect to the actual dose levels will usually be more sensitive than the standard pair-wise comparisons (33).

One-Sided Versus Two-Sided Procedures

Koch (64) points out that ". . . the objectives of a study determine whether its inferential posture is one-sided or two-sided. To be effective, however, an intended inferential posture must have clear documentation for its nature and rationale in the protocol for a study. . . ." It can be agreed that investigating dose- or treatment-related increases in tumor incidence is usually the intended posture in protocols written for long-term carcinogenicity studies. For certain tumors, however, tumor combinations and groups of "negatively correlated" lesions, testing for selected decreases in tumor rates may be informative to an investigator.

Modelling and Parametric Methods for Time-to-Event Data

Nonparametric time-to-event analysis of occult tumor onset, such as the Peto test described above, requires knowledge of tumor lethality (16). In the absence of such data, parametric assumptions are needed. Since tumor lethality data are often unreliable or unknown, a number of alternative semi- or fully-parametric methods have been developed.

Fully-parametric models in which tumor onset and death rates are assumed to follow a Weibull distribution known except for the values of a small number of parameters have been proposed by Kodell and Nelson (65), Dewanji et al. (66), and Omar et al. (67). Semi-parametric models have also been proposed by Dinse (68,69) and Portier (70). In these models minimal parametric assumptions are made, for example, that the tumor onset and death rates differ by a constant, with nonparametric methods used to estimate the death rate. The semi-parametric models are based on fewer assumptions than the fully-parametric methods but provide no opportunity for validation of the assumptions made. For fully-parametric methods, by contrast, some validation of the appropriateness of the parametric form used is possible (67). Animal tumorigenicity studies in drug development are seldom large enough to permit reliable assessment of model parameters beyond those directly reflective of dose response.

Adjustments for Multiple Comparisons and Multiple Endpoints

In carcinogenicity studies there are two major sources of multiplicity: multiple comparisons among the treatment groups and multiple tumor sites, both of which operate to increase the experimentwise type I error. The ultimate objective of a carcinogenicity study is control of the consumers' risk (type II error). Multiplicity adjustments to limit experimentwise type I error increase type II error dramatically.

To solve this problem either no adjustment (all comparisons based on level-alpha tests) should be applied or a multiple comparison procedure which ensures only a small increase in type II error should be used. Such a procedure for the one-way layout "control versus k doses" should take into account the type of endpoint (tumor rate), the restriction of the alternative to an increase (one-sided testing), increasing tumor rates with increasing doses but with the possibility of downturns at high doses, the unknown shape of the dose-response relationship, different patterns

of censoring, different sex-, site- and group-specific number of animals at risk, and the small sample sizes. The objective of such a procedure is to reveal the existence of a generally increasing trend. Beyond this, procedures for analyzing the multiple, correlated tumor sites should be taken into account (71,72).

If analyses are presented with no adjustment for multiplicity, the familywise type I error should be estimated. Because the familywise error depends on the number and type of comparisons involved, several calculations based on a variety of reasonable definitions of the "family" may be more informative that a single estimate in which the family is defined to be all possible comparisons.

A separate analysis is ordinarily done for males and females on each distinct tumor type observed in at least two animals, often resulting in 50 or more analyses per sex. Given such a large number of statistical tests, the probability of a false-positive result for at least one tumor type (the familywise error rate, or FWE) is greatly increased. Formal P-value adjustment methods to control the FWE necessarily result in a substantial decrease in power, but can balance a tendency to give excess weight to an isolated statistically significant finding having no biological support.

Several tumor types in a study may be statistically significant at the $P = 0.05$ level, raising the question how to separate any true treatment effects from false positive results arising solely by chance. At least four approaches can be useful: consistency of effect (an effect is observed in both males and females, or in two species, for example), consideration of known mechanism of action, comparison with historical control data, and adjustment for multiplicity of endpoints.

Haseman (73) proposed an ad hoc rule for adjusting P-values for multiplicity of endpoints: a P-value is considered "significant" at the 0.05 level if it is significant at the $P = 0.01$ level and it is a common tumor (with historical control rate above 1%), or if it is significant at the 0.05 level and it is a rare tumor (with historical control rate 1% or below). Assuming two groups with 50 animals per group and use of a one-sided Fisher exact test, Haseman estimated a familywise error rate (FWE) of 1.3–2.4%, depending on sex and species. This calculation was based on empirical results, using the standard list of tissues examined in the NTP studies and their usual rare and common outcomes. This same rule results in a FWE of 7.3–7.5% if the false positive error rate for a two-sex, two-species series of experiments is to be controlled. This rule has the advantage that power is stable regardless of the number of additional rare tumor types analyzed. If one does not have two sexes and two species, a different rule would be appropriate.

Carcinogenicity studies using trend tests and with 200 animals per sex have FWE greatly above that reported by Haseman (73). If an analogous ad hoc multiplicity adjustment rule is used, some estimate of its FWE should be provided. The United States Food and Drug Administration currently uses a 0.005 level (0.01 level) for the one-sided trend test of a common (rare) tumor in order to limit the FWE to approximately 0.1 (74). Note that this larger type I error rate is tolerated in order to keep the type II error rate reasonably low.

If a formal multiplicity adjustment is used, it should take account of discreteness for tumor types seen in few animals. The Bonferroni adjustment is extremely conservative and should not be used. Modifications of the Bonferroni procedure, however, such as that of Holm (75), are less conservative and may be useful.

Permutation or bootstrap resampling methods can be used for formal multiplicity adjustment (57). The resampling methods of Westfall and Young (76) provide multiplicity-adjusted P-values including age-adjustment by stratification, and are available in proc MULTTEST of the Statistical Analysis System. If all tumors are incidental then correlations among tumor types are accounted for. If some tumors are lethal then age and multiplicity adjusted P-values can be calculated if tumor types are assumed independent.

If P-values are formally adjusted for multiplicity, they should supplement, not replace, unadjusted P-values. A statistical report with multiplicity-adjusted P-values must always include the unadjusted P-values as well. Biological interpretation must be combined with statistical significance results to obtain a balanced interpretation of study results.

Use of Historical Controls

The term historical controls as used in preclinical pharmaceutical research refers to control groups in previous animal experiments that can be considered to be comparable to the control group in a current experiment. An essential prerequisite for including controls from different studies is that the studies were conducted under similar or identical protocols with regard to the control group, the species and strain, the vehicle (if any), animal health, age, and husbandry. For the purpose of any analysis of tumor incidence, the duration of the follow-up of the animals up to the particular time under risk considered in the analysis as well as the sacrifice scheme have to be similar. Although there may be several sources for inhomogeneity among the control groups included in the database—for example, genetic variations, environmental factors, methods of animal handling, diet, and different pathology techniques—historical controls have their chief value in providing information on the spontaneous tumor rate and variability in the species and strain concerned under reasonably standardized conditions.

Genetic variations and other, often unknown factors may be important sources of variation, at least for certain tumors. For a subset of the NTP database it has been shown that the laboratory and even the month may substantially contribute to variation (77). Because there may also be remarkable shifts in tumor rates over time, a reasonable recommendation is that historical comparisons be limited to control data for the past 3–4 years, preferably from the same laboratory (78). The factors to be recorded, controlled, and standardized for building up an historical database include: species and strain, source of animals, sex, age, extent of pathological examination, dietary and environmental conditions, diagnostic criteria and nomenclature, time at risk (follow-up time) for the particular event considered, survival time, animal health and husbandry, quality assurance, and peer review procedures (78,79).

Roughly speaking, variations in historical databases may be due to factors that operate between laboratories, between trials of the same laboratory, and within trials. It is recommended that drug developers seek statistical advice on how to analyze these different sources of variation and to assess the comparability of the trials included in the database from the statistical point of view.

Historical control group data have been used as a safeguard against false negatives and false positives when performing statistical tests. Since the usually applied statistical tests rely entirely on the current experiment at hand they may fail to detect toxicologically relevant increases in the case of rare tumors. For example, an increase from 0/50 tumors in concurrent controls to 3/50 in treated animals yields a p-value of 0.12 when using the one-sided Fisher test. Against a background rate of 1/1000 in the historical database, however, this statistically nonsignificant result may well be viewed as pharmacologically relevant. On the other hand, an uncommonly low incidence in the concurrent control group accompanied by data obtained under treatment that are within historical control variation is a situation where statistically false positive results may occur. Provided that there is no dose-related trend or other evidence of carcinogenic potential it might be assumed that the low incidence appeared by chance and that the events in the treated groups give no indication of a carcinogenic effect. Obviously, there is some ambiguity in such a conclusion, which should be taken seriously since the study results indicate an increased risk to consumers. The reduction of the incidence in the control group may well be caused by some condition that also operated on the treated groups. In addition, due to various sources of variability the incidence

ranges in an historical database will increase with the number of studies included. Although this effect will be much more pronounced for tumors that have relatively high spontaneous control rates than with rare tumors, any comparison of tumor rates in treated animals to the range in historical controls should be based on solid statistical arguments. Comparing the maximum rates from all previous controls to that encountered in the treated groups of the current study is not a defensible procedure (78).

Several parametric and nonparametric statistical procedures have been proposed to incorporate historical control information (80–83). The basic principle is that they combine information both from historical controls and the concurrent control with weighting that takes account of the corresponding variabilities.

At present, no statistical procedure clearly outperforms all others. Some of the parametric approaches suggested tend to be unstable against minor changes of the historical database (83) and may give rise to biased estimates (84) resulting in inflated false positives. Formally incorporating historical information into a statistical procedure has been challenged with regard to having any justification at all (85). Nevertheless, analysis across several studies (meta-analysis) has found increased use in various clinical situations (86). It is concluded that using historical control information for routine statistical testing should be avoided, except in cases of equivocal results or those involving rare tumors. An essential prerequisite is a sound knowledge of the historical database and the underlying study outlines. The statistical method to be used for incorporating historical data should be prespecified in the study protocol.

INTERPRETING AND REPORTING RESULTS

Relating Tumors to In-life Events

Concomitant in-life information, such as food consumption and body weight, is usually collected during the course of a carcinogenicity study and may be considered in parallel with the analysis of tumor incidence. In-life data may provide valuable information to aid the interpretation of a carcinogenicity endpoint and should not be ignored in the biological interpretation of effects. It must be recalled that in-life events may also reflect treatment or dose effects.

Dealing with Serendipitous Tumors

It may be that during the course of a macroscopic examination of all the animals, a tumor is noted in an organ that was not included on the list of protocol organs and hence not routinely examined microscopically. For example, a compound may induce an extremely rare type of tumor. Alternatively there may be some reason for not including all animals in the microscopic examination of particular organs (economic reasons, for example). Three possible courses of action in this situation are:

1. Count any animal for which that organ was not examined as negative,
2. Exclude unexamined animals from the analysis, or
3. Exclude from the analysis the group containing such unexamined animals.

Which of these options is used depends on prevailing circumstances.

The first option is relevant when it can safely be assumed that had the condition been present, then it would have been noticed during necropsy. This may be a reasonable assumption if the tumor is usually evident macroscopically when present, and also taking into account the experience and competence of the technicians responsible for the necropsy. If there is doubt as to whether this assumption is reasonable then this type of analysis would be invalid.

The second option is of limited use, particularly if the reason for microscopic examination is related to the presence of tumor as described above. If the reason for examination is unrelated to the presence of tumor and

if a similar number of animals were examined in each of the groups, then this option may be useful. Sample size could become an issue if many animals were to be excluded.

In certain protocols there are different examination schedules for different groups. For example, certain organs may be routinely microscopically examined in the control and high-dose groups but not in the intermediate groups. There may be a case here for assuming that groups not examined are negative, although the validity of such an assumption would need to be justified. If this assumption were in doubt, then it would be preferable to exclude the unexamined groups (option 3) from the analysis altogether and consider only the fully examined groups.

If there is doubt as to whether the incidence of neoplasia in nonprotocol organs is correctly and adequately represented, then further experimental work may be necessary, including microscopic examination of all groups. Examples may include particularly small organs where equally small tumors may be difficult to identify macroscopically, or organs in which developing neoplasia may be wholly contained and externally invisible.

Biological Reasoning

Statistical analysis of the tumor incidence in a carcinogenicity study should be interpreted in light of the overall study findings. These include survival rate, cause of death contributing to lack of survival (tumor or nontumor), known pharmacological effects of the compound, toxic effects of the compound, other tumor-exciting factors such as changes in husbandry, diet (eg, protein content, caloric content), background health status of the animals, and so forth.

Compounds such as those which are hormonally active are known to produce an increase in certain types of hormonally dependent tumors. For example, testosterone receptor antagonists produce an increase in interstitial tumors in the rat, which is not considered to be of biological significance and has not precluded compound registration.

Cytochrome p450 inducers can cause liver enlargement, increased thyroxine metabolism resulting in an increased production of thyroid hormone and thyroid tumors. This phenomenon has been shown to be specific to the rat and is not believed to be biologically relevant in terms of human risk assessment. Compounds which produce tissue damage (eg, in the liver) have been shown to result in increased tumor production as a result of increased cell turnover. It is well known that animals on high protein and high calorie diets put on excessive weight. This can result in reduced survival and increased tumor incidence.

Historical Information

Any increase in tumor incidence should also be interpreted in the light of the historical incidence of that tumor. The occurrence of an unusual/rare tumor, even at statistically nonsignificant levels, may be viewed as being of biological significance. Similarly, a tumor causing an increase in premature deaths, although not showing an overall statistically significant increase, may be considered to be of biological significance.

All historical data should be interpreted in light of genetic drift with time. Thus, new data should be compared with recently derived historical data.

Documentation of a Study

The documentation of a study should include all information needed for regulatory pharmacologists/toxicologists and statisticians to conduct their reviews. The information should include a statement on data quality assurance; individual animal data on lesions, mortality, body weight, and so forth; summary of the study; description of the nature of the drug and its treatment indication; general information containing study location, study identification, duration, persons involved and their responsibilities; information regarding material, test system, and methods used, such as test substance and auxiliaries, mixes and analyses of test substance in the

feed, dosage form, doses, study groups, rationale for the dose selection, test animals and housing conditions, inspection of the animals, determinations of body weight, of feed and test substance, and water intake, hematological investigations, enzymes, electrolytes, substances and metabolic products in blood, urinalyses, toxicokinetic investigations, ophthalmological investigations, necropsy, organ weights, and histopathological (both gross and microscopical) examinations; historical data; computer-assisted collection of data and processing, statistical analyses and presentation of results; conclusion of evaluation of the results; references; and attached tables and figures containing the information mentioned above.

Electronic Submission of Data for Regulatory Review

Because carcinogenicity studies contain large amounts of data, additional statistical analyses are not practical if the data must be entered into computers from the sponsor's paper dossier. Machine-readable files are generally submitted to regulatory authorities along with the paper dossier to facilitate necessary statistical review and analyses.

In the United States, several regulatory authorities commonly require carcinogenicity studies as part of their responsibility to regulate human exposure to potential carcinogens. Carcinogenicity studies are often performed by contract laboratories, which may not know what kind of chemical is under study and, therefore, which regulatory agency will ultimately review the results. Some companies develop products that fall under the authority of different regulatory agencies. Until the mid 1980s, each regulatory agency had its own format for submission of data. At that time, these agencies collaborated to develop a common format, called the Submitters Toxicological Uniform Data Information Exchange Standard (STUDIES), to simplify submission of carcinogenicity study data (87).

The STUDIES format is a series of file structures, each covering one aspect of a lifetime carcinogenicity study. One file contains food consumption data; another contains tumor data; another contains organ weights; and so forth. Although the dozen files comprising the STUDIES format were designed to capture all data generated in a carcinogenicity study, not all files are needed for analysis of tumor findings. The needed files are those identifying the animals, their time of death, and their tumor pathology. Two files decode the tissues and specific lesions observed. In order to assess whether the animals were suitably challenged or to explain unusual findings, data on weight gain and food consumption may also be needed.

Predating the STUDIES formats, the Division of Biometrics, Center for Drug Evaluation & Research, of the United States Food & Drug Administration had developed a format (the "Biometrics" format) in which drug sponsors submitted data of carcinogenicity studies. These formats were used as the model for the files related to tumors that are part of the STUDIES formats, so there is considerable overlap between these sets of files. The "Biometrics" formats have been recently modified to capture comparable information as in the STUDIES formats.

Sponsors may submit data to the United States Food & Drug Administration in either the "new Biometrics" or the STUDIES format. Modifications to these formats to accommodate variations in study design, outcomes, and so forth are acceptable if well documented. Data should be submitted on 3½″ diskettes. Although the formats anticipate ASCII text files, files which contain the same information in other commercially available structures (SAS, Excel, SPlus, etc.) are acceptable. Details on the STUDIES and "new Biometrics" formats can be obtained by writing to the Office of Epidemiology & Biostatistics (HFD-700), Center for Drug Evaluation & Research, United States Food & Drug Administration.

REFERENCES

1. Lagakos SW. An evaluation of some two-sample tests used to analyze animal carcinogenicity experiments. *Utilitas Mathematica.* 1982;21B:239–260.
2. Ryan LM. Efficiency of age-adjusted tests in animal

carcinogenicity experiments. *Biometrics*. 1985;41: 525–531.
3. Bailer AJ, Portier CJ. An illustration of dangers of ignoring survival differences in carcinogenicity data. *J Appl Toxicol*. 1988;8:185–189.
4. Sontag JM, Page NP, Saffiotti U. *Guidelines for Carcinogen Bioassay in Small Rodents*. DHHS Publication (NIH) 76–801. Bethesda, MD: National Cancer Institute; 1976.
5. Page N. Concept of a bioassay program in environmental carcinogenesis. In: Kraybill HF, Mehlman HA, eds. *Advances in Medical Toxicology*. New York, NY: Wiley and Sons; 1977:87–171.
6. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *Guideline for Industry: Dose Selection for Carcinogenicity Studies of Pharmaceuticals*. 1995.
7. Huff JE, Moore JA. Carcinogenesis studies design and experimental data interpretation/evaluation at the National Toxicology Program. In: *Industrial Hazards of Plastics and Synthetic Elastomers*. New York, NY: Alan R. Liss Inc.; 1984:43–64.
8. Munro IC. Considerations in chronic toxicity testing: The chemical, the dose, the design. *J Environ Pathol Toxicol*. 1977;1:183–197.
9. Ames BN, Gold LS. Too many rodent carcinogens: mitogenesis increases mutagenesis. *Science*. 1990; 249:970–972.
10. Cohen SM, Ellwein LB. Genetic errors, cell proliferation and carcinogenesis. *Cancer Res*. 1991;51: 6493–6505.
11. Morgan DG, Kelvin AS, Kinter LB, Fish CJ, Kerns WD, Rhodes G. The application of toxicokinetic data to dosage selection in toxicology studies. *Toxicologic Pathology*. 1994;22:112–123.
12. Kipling J. *Dose Selection for Carcinogenicity Studies of Pharmaceuticals FINAL DRAFTs—ICH Harmonised Tripartite Guidelines*, The Association of the British Pharmaceutical Industry; 1994.
13. Faccini JM, Butler WR, Friedmann JC, Hess R, Reznik GK, Ito N, Hayashi Y, Williams GM. IFSTP Guidelines for the design and interpretation of the chronic rodent carcinogenicity bioassay. *Exp Toxicol Pathol*. 1992;44(8):443–456.
14. Greenman DL, Bryant P, Kodell RL, Sheldon W. Influence of cage shelf level on retinal atrophy in mice. *Lab Animal Sci*. 1982;32:353–356.
15. Gart JJ, Krewski D, Lee PN, Tarone RE, Wahrendorf J. *Statistical methods in cancer research Volume III. The design and analysis of long-term animal experiments*. Lyon; International Agency for Research on Cancer; 1986.
16. McKnight B, Crowley J. Tests for differences in tumor incidence based on animal carcinogenesis experiments. *J Am Stat Assoc*. 1984;79:639–648.
17. Bross ID. Why proof of safety is much more difficult than proof of hazard. *Biometrics*. 1985;41:785–793.
18. Millard SP. Proof of safety vs proof of hazard. *Biometrics*. 1987;43:719–725.
19. Holland B, Ordoukhani NK. Balancing type I and type II error probabilities: Further comments on proof of safety vs proof of hazard. *Comm Stat-Theory Meth*. 1990;19:3557–3570.
20. Erickson WP, McDonald LL. Tests for bioequivalence of control media and test media in studies of toxicity. *Environ Toxicol Chem*. 1995;14:1247–1256.
21. Hauschke D. Statistical proof of safety in toxicological studies. *Drug Inf J*. 1997;31(2):357–361.
22. Hauschke D, Hothorn L. Safety assessment in toxicology studies: Proof of safety versus proof of hazard. In Chow S-C and Liu J-P eds. *Design and Analysis of Animal Studies in Pharmaceutical Development*. New York, NY: Marcel Dekker, Inc; 1998.
23. Stallard N, Whitehead A. An alternative approach to the analysis of animal carcinogenicity studies. *Regul Toxicol Pharmacol*. 1996;23:244–248.
24. Haseman JK, Hajian G, Crump KS, Selwyn MR, Peace KE. Dual control groups in rodent carcinogenicity studies. In Peace KE, ed. *Statistical Issues in Drug Research and Development*. New York, NY: Marcel Dekker, Inc.; 1990.
25. Armitage P. Tests for linear trends in proportions and frequencies *Biometrics*. 1955;11:375–386.
26. Portier CJ, Hoel D. Type 1 error of trend tests in proportions and the design of cancer screens. *Comm Stat A*. 1984;13:1–14
27. Neuhaeuser M, Hothorn LA. Trend tests for dichotomous endpoints with application to carcinogenicity studies. *Drug Inf J*. 1997;31(2):463–469.
28. Little, RJA. Testing the equality of two independent binomial proportions. *Am Statis*. 1989;43(4):283–288.
29. Storer BE, Kim C. Exact properties of some exact test statistics for comparing two binominal distributions. *J Am Stat Assoc*. 1990;85:146–155.
30. Lancaster HO. Significance tests in discrete distributions. *J Am Stat Assoc*. 1961;56:223–234.
31. Chen JJ, Kodell RL, Pearce BA. Significance levels of randomization trend tests in the event of rare occurrences. *Biom J*. 1998.
32. Lin KK, Ali, MW. Statistical review and evaluation of animal tumorigenicity studies. In: Buncher CR and Tsay JY eds. *Statistics in the Pharmaceutical Industry*. New York: Marcel Dekker; 1994.
33. Peto R, Pike MC, Day NE, Gray RG, Lee PN, Parish S, Peto J, Richards S, Wahrendorf J. Guidelines for simple sensitive significance tests for carcinogenic effects in long-term animal experiments. In: *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, supplement 2: Long-term and Short-term Screening Assays for Carcinogens: A Critical Appraisal*. Lyon: International Agency for Research on Cancer; 1980:311–346.
34. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *J R Stat Soc A*. 1972;135(2):185–206.
35. Hoel DG, Walburg HE. Statistical analysis of survival experiments. *J Nat Cancer Inst*. 1972;49:361–372.
36. Lagakos SW, Louis TA. Use of tumor lethality to

interpret tumorigenicity experiments lacking cause-of-death data. *App Stat.* 1988;37:169–179.
37. McKnight B, Wahrendorf J. Tumor incidence rate alternatives and the cause-of-death test for carcinogenicity. *Biometrika.* 1992;79:131–138.
38. Kodell RL, Chen JJ, Moore GE. Comparing distributions of time to onset of disease in animal tumorigenicity experiments. *Comm Stat-Theory Meth.* 1994; 23:959–980.
39. Boldebuck DH, Neuhaus G, Heimann G. Analyzing carcinogenicity assays without cause of death information. *Drug Inf J.* 1997;31(2):489–507.
40. Dinse GE. A comparison of tumor incidence analyses applicable in single-sacrifice animal experiments. *Stat Med.* 1994;13:689–708.
41. Lindsey JC, Ryan LM. A comparison of continuous- and discrete-time three-state models for rodent tumorigenicity experiments. *Environ Health Perspect.* 1994;102 (Suppl. 1):9–17.
42. Kodell RL, Pearce BA, Turturro A, Ahn H. An age-adjusted trend test for the tumor incidence rate for single sacrifice experiments. *Drug Inf J.* 1997;31(2): 471–487.
43. Gehan E. A generalized Wilcoxon test for comparing arbitrarily singly censored data. *Biometrika.* 1965; 52:650–654.
44. Prentice RL, Marek P. A qualitative discrepancy between censored data rank tests. *Biometrics.* 1979; 35:861–867.
45. Breslow N. A generalized Kruskal Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika.* 1970;57:579–594.
46. Koziol JA, Reid N. On multiple comparisons among K samples subject to unequal patterns of censorship. *Comm Statist A.* 1977;12:1149–1164.
47. Tarone RE, Ware J. On distribution free tests for equality of survival distributions. *Biometrika.* 1977; 64:156–160.
48. Chen JJ, Gaylor D. The upper percentiles of the distribution of the logrank statistic for small sample sizes. *Comm Stat-Simul.* 1986;15(4):991–1002.
49. Heimann G, Neuhaus G. The permutational distribution of the logrank statistic under random censorship with applications to carcinogenicity assays. University of Hamburg; 1996.
50. Ali MW. Exact versus asymptotic tests of trend of tumor prevalence in tumorigenicity experiments: A comparison of P-values for small frequency of tumors. *Drug Inf J.* 1990;24:727–737.
51. Neuhaus G. Conditional rank tests for the two sample problem under random censorship. *Ann Stat.* 1993;21:1760–1779.
52. Gilbert JP. *Random Censorship.* A dissertation submitted to the Department of Statistics, University of Chicago; 1962;34–73.
53. Miller RG, Jr. *Survival Analysis.* New York, NY: John Wiley & Sons; 1981;85–103.
54. Mantel N. Assessing laboratory evidence for neoplastic activity. *Biometrics.* 1980;36:381–399.
55. Soper KA, Tonkonoh N. The discrete distribution used for the log-rank test can be inaccurate. *Biom J.* 1993;35:291–298.
56. Edgington E. *Randomization Tests.* 3rd ed. New York, NY: Marcel Dekker, Inc.; 1995.
57. Heyse JF, Rom D. Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. *Biom J.* 1988;30(8):883–896.
58. Dinse GE, Haseman JK. Logistic regression analysis of incidental tumor data from animal carcinogenicity experiments. *Fund App Toxicol.* 1986;6:44–52.
59. Chen JJ, Moore GE. Impact of surviving time on tests for carcinogenicity *Comm Stat-Theory Meth.* 1994;23(5):1375–1386.
60. Mantel N, Tukey JW, Ciminera JL, Heyse JF. Tumorigenicity assays including use of the jackknife. *Biom J.* 1982;24:579–596.
61. Park CN, Kociba RJ. Statistical evaluation of long-term animal bioassays for carcinogenicity. In: Milman HA, Weisburger EK, eds. *Handbook of Carcinogen Testing.* Park Ridge, NJ: Noyes Publications; 1985:358–371.
62. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 1976;63:655–660.
63. Kodell RL, Chen JJ. Characterization of dose-response relationships inferred by statistically significant trend tests. *Biometrics.* 1991;47:139–146.
64. Koch GG. One-sided and two-sided tests and p-values. *J Biopharm Stat.* 1991;1:161–170.
65. Kodell RL, Nelson CJ. An illness-death model for the study of the carcinogenic process using survival/sacrifice data. *Biometrics.* 1980;36:267–277.
66. Dewanji A, Krewski D, Goddard MJ. A Weibull model for the estimation of tumorigenic potency. *Biometrics.* 1993;49:367–377.
67. Omar RZ, Stallard N, Whitehead J. A parametric multistate model for the analysis of carcinogenicity experiments. *Lifetime Data Analysis.* 1995;1:327–346.
68. Dinse GE. Constant risk differences in the analysis of animal tumorigenicity data. *Biometrics.* 1991;47: 681–700.
69. Dinse GE. Evaluating constraints that allow survival-adjusted incidence analyses in single-sacrifice experiments. *Biometrics.* 1993;49:399–407.
70. Portier CJ. Estimating the tumor onset distribution in animal carcinogenesis experiments. *Biometrika.* 1986;73:371–378.
71. Finkelstein DM, Schoenfeld DA. Analysis of multiple tumor data from a rodent carcinogenicity experiment. *Biometrics.* 1989;45:219–230.
72. Chen JJ. Global tests for analysis of multiple tumor data from animal carcinogenicity experiments. *Stat Med.* 1996;15:1217–1225.
73. Haseman JK. A reexamination of false-positive rates for carcinogenesis studies. *Fund App Toxicol.* 1983; 3:334–339.
74. Lin KK, Rahman MA. Overall false positive rates in tests for linear trend in tumor incidence in animal carcinogenicity studies of new drugs. *J Biopharm Stat.* 1998;8(1):1–15.

75. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6:65–70.
76. Westfall PH, Young SS. *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment.* New York, NY: John Wiley and Sons; 1993.
77. Haseman JK, Huff J, Boorman GA. Use of historical control data in carcinogenicity studies in rodents. *Toxicol Pathol.* 1984;12:126–135.
78. Board of Scientific Counselors, National Toxicology Program, US Department of Health and Human Services. *Report of the NTP ad hoc panel on chemical carcinogenesis testing and evaluation*; 1984.
79. VanZwieten MJ, Majka JA, Peter CP, Burek JD. The value of historical control data. In: Grice HC, Ciminera JL, eds. *Carcinogenicity: The Design Analysis and Interpretation of Long-Term Animal Studies.* New York, NY: ILSI Monograph Series Springer-Verlag; 1988:39–52.
80. Tarone RE. The use of historical control information in testing for a trend in proportions. *Biometrics.* 1982;38:215–220.
81. Hoel DG, Yanagawa T. Incorporating historical controls in testing for a trend in proportions. *J Am Stat Assoc.* 1986;81:1095–1099.
82. Krewski D, Smythe RT, Dewanji A, Colin D. Tests for trend in carcinogen bioassay with historical controls. In: *Proceedings of the Biopharmaceutical Section ASA.* Washington, DC; American Statistical Association; 1986:248–253.
83. Smythe RT. The use of historical controls in bioassay for carcinogens. In: Krewski D Franklin C, eds. *Statistics in Toxicology.* New York, NY: Gordon and Breach Science Publishers; 1991:593–609.
84. Tamura R, Young SS. The incorporation of historical control information in tests of proportions: Simulation study of Tarone's procedure. *Biometrics.* 1986; 42:343–349.
85. Tamura R. Discussion to the paper by Krewski D, Smythe RT, Burnett RT. The use of historical control information in testing for trend in quantal response carcinogenicity data. *Symposium on Long-Term Animal Carcinogenicity Studies: A Statistical Persepctive.* Bethesda, MD, March 4–6, 1985.
86. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med.* 1991;10:1665–1677.
87. Environmental Protection Agency. *STUDIES/ Chronic: Data formats for chronic/oncogenicity rodent bioassays.* Springfield, VA: National Technical Information Service (PB90–213885); 1990.

ADDITIONAL SUGGESTED READING

Krewski D, Brown C. Carcinogenic risk assessment: A guide to the literature. *Biometrics.* 1981;37:353–366.
Portier CJ. Design of animal carcinogenicity experiments: Dose allocation, animal allocation and sacrifice times. *Proceedings of the Symposium on Long-Term Animal Carcinogenicity Studies: A Statistical Perspective.* 1995:42–50.
Portier CJ. Biostatistical issues in the design and analysis of animal carcinogenicity experiments. *Environ Health Perspect Supp.* 1994;102(1):5–8.