3. EXPERIMENTAL DESIGN

3.1 Introduction
3.2 Principles of experimental design
3.3 Designs for screening studies
3.4 Designs for dose-response studies
3.5 Designs for studies of mechanism
3.6 Histopathological analysis
3.7 Recording of experimental data
3.8 Summary and recommendations

CHAPTER 3

EXPERIMENTAL DESIGN

3.1 Introduction

Although the primary emphasis in this monograph is on the analysis of carcinogenicity data, several statistical principles underlie the design of all types of experiments. These need to be taken into account in the planning stages of any study, preferably with the involvement at that time of a qualified, experienced statistician. If these principles are ignored, sensible conclusions often cannot be drawn from the data, no matter how sophisticated the statistical method of analysis.

In addition to statistical principles, many other considerations are involved in the planning of a long-term carcinogenicity study. These include the responsibilities of key personnel involved in the conduct of the study, the characterization of the physical and chemical properties of the test substance, the selection of a suitable animal model, the control of the laboratory environment in which animals are housed, the route of exposure and the dosing regimen to be followed, health monitoring and procedures for both gross and histopathological examination, and the methods for accurately recording and storing data for subsequent statistical analysis. This chapter will focus primarily on statistical issues in the planning of experiments; however, this represents only one aspect of good design and cannot be considered in isolation from the many practical concerns just noted.

The need for a well-organized, overall experimental design is well stated in the introductory section of the IARC (1980) report on the conduct of carcinogenicity tests:

'The objective of any chronic study is to ascertain what effect repeated administration of a chemical will have upon tissues or organ systems in animals of either sex of the test species. The attainment of the objective requires:

'(a) a well-devised and explicit protocol, coupled with sufficient supervision to monitor the daily activities of the study, to ensure that all items of protocol and any changes thereof are understood and are being followed. Any deviations from the protocol must be well documented as to the reason(s) for the deviations, their extent and their nature;

'(b) a technical staff who thoroughly understand their responsibilities and duties, as well as a management that recognizes the importance of the technical staff in the conduct of a carcinogenicity study and is supportive of the staff;

'(c) a record-keeping system which is accurate, reliable, secure and complete . . . ;

'(d) a health monitoring programme that will ensure accurate diagnosis of disease or toxic states, with a minimal loss of tissue samples for histological examination; and

'(e) an archive for storing the test data, protocols and specimens to allow for possible reevaluation in the light of future studies.'

Objectives of carcinogenicity testing

A long-term carcinogenicity bioassay may be conducted for one of several purposes; these may include (i) screening for potential carcinogens, (ii) determination of dose-response relationships, and (iii) elucidation of possible mechanisms of carcinogenic action (Clayson *et al.*, 1983). Screening studies may be employed in order to evaluate the carcinogenic potential of the compound on a qualitative basis. Such studies are intended to establish the existence of a carcinogenic hazard. They are usually conducted at relatively high exposure levels that are administered for the major portion of the lifespan of the test species, so that the probability of observing an effect in a relatively small sample of experimental animals is maximized. Although the observation of adverse effects at a single dose can provide positive evidence of carcinogenicity, a second, lower dose is frequently employed, both to confirm the evidence and to protect against the loss of the high-dose group to intercurrent mortality.

A series of increasing dose levels might be employed in a dose-response study intended to delineate in more quantitative terms the shape of the dose-response curve for a carcinogenic agent. This information is useful if some measure of the potency of the agent is to be estimated (Purchase, 1980), or if it is desired to extrapolate results from the high doses actually used down to lower doses more characteristic of the human environment (Crump, 1979; Armitage, 1982).

More elaborate experimental designs may be used in an attempt to define the mechanism of action of the test agent. However, as studies of transplacental carcinogenesis, initiation/promotion systems, variable exposure patterns (including cessation of exposure), and the synergistic/antagonistic effects of mixtures become more commonplace, guidelines need to be developed for the design of these more specialized experiments.

Experimental design

Complete specification of the experimental design requires careful consideration of a number of design parameters. Perhaps the most fundamental parameter is the total number of animals to be used. While the amount of information obtained will clearly increase with the use of additional animals, there comes a point where the value of the incremental information may not be worth the extra cost. (A more precise discussion of this point is given in Section 3.3 below.) From the practical point of view, logistical and budgetary constraints may also serve to limit the size of the experiment.

Another fundamental consideration is the number of treatment groups and the particular dosing regimen to be applied to each group. As was discussed in the previous

section, a variety of possible treatments and dosing regimens may be considered, depending on the objectives of the study in question.

Once the size of the experiment and the structure of the experimental treatments have been determined, there remains the question of how to allocate the available animals to the various groups in the best possible way. The simplest approach is to assign equal numbers of animals to each group. While such a balanced allocation may be reasonable in many cases, it is not always the optimal strategy (see Sections 3.3–3.5).

A crucial part of any experimental design is proper randomization. Without a randomized design, it is not possible to determine whether any observed differences between the treatment groups are more likely to be due to the treatments themselves than to other intrinsic differences between the groups. Proper randomization permits statistical inferences based on the probability of any observed effects being due to chance alone. Thus, randomization both protects against biases due to unsuspected confounding factors present in the laboratory environment and provides a basis for valid statistical inference (Cochran, 1974; Kempthorne, 1977; Gart *et al.*, 1979; Edgington, 1980).

General design guidelines

A number of important factors must be taken into account in the early stages of planning a carcinogenicity bioassay. These have been considered by a variety of expert groups (Health and Welfare Canada, 1975; Food Safety Council, 1978; Interagency Regulatory Liaison Group, 1979; Environmental Protection Agency, 1979; IARC, 1980). Although many of these are nonstatistical in nature, they nonetheless represent important components of the overall experimental design.

Characterization of test substances: Prior to the conduct of the study, the physical and chemical properties of the test substance should be established, including its purity and the nature of any impurities. Knowledge of chemical reactivity is of importance, particularly when the test agent can react with components of the basal diet.

Species and strain of test animal: In theory, the ideal animal model would be one in which there is little or no tumour incidence in the control animals, but in which the effects of the treatment being tested for carcinogenicity are easily seen. It is clear that the choice of a strain highly resistant to tumours is a poor idea. However, it is not so obvious that a species which has a high background incidence of tumours of the same type as the ones being studied is also to be avoided. There are two main reasons for this. One is that, with low spontaneous rates, far fewer animals of a strain with low background incidence are needed to detect a small increase as statistically significant (see Section 3.3). The second reason for avoiding strains with high background incidence rates is that it is not clear whether an increase in the rate of occurrence of a tumour, common in the animal species, but less so in man, really provides biological evidence in support of a carcinogenic effect in man (Clayson *et al.*, 1983).

In practice, limitations of space, time and cost usually dictate the use of small

rodents, particularly the rat, mouse and Syrian hamster, in carcinogenicity bioassays. The cost of the animals, as well as their availability, longevity and familiarity, are all important factors affecting the choice. Because of differences in susceptibility, the use of more than one species or strain can be advantageous.

Route of exposure: In order to facilitate interspecies conversion, it is desirable that the route of exposure correspond to that in the human situation. Alternative routes may be acceptable if they result in equivalent levels of the test material or its metabolites in the target tissue.

Duration of the experiment: The duration of the experiment can markedly affect the conclusions reached. In general, it is desirable to continue the experiment for a period sufficiently long to provide enough time for the development of tumours which occur long after the start of the exposure period. On the other hand, extending the duration of the experiment may result in a high rate of occurrence of spontaneous lesions among the control animals. For example, in the absence of treatment, more than 20% of female B6C3F1 mice develop lymphoma-leukaemia of the haematopoietic system after 104 weeks of age (Ward *et al.*, 1979) and over 90% of male F344 rats develop interstitial-cell tumours of the testis by the same time (Goodman *et al.*, 1979). The high rate of occurrence of these and other lesions in untreated aged animals makes it more difficult to identify significant effects in the exposed groups at such sites. The interpretation of data involving aged rodents may also be complicated by normal geriatric changes which occur within animal populations (International Life Sciences Institute, 1984a).

In the past, experiments of 18 and 24 months have often been used for mice and rats, respectively, with exposure beginning at the time of weaning. Two other criteria for the termination of the study have also been utilized. One is to continue the study for the full lifespan of the experimental animals. With modern animal husbandry, however, it is possible that some rodents may live for three to four years or even longer. Another proposal has been to terminate the experiment when the proportion of animals surviving in the control group falls to 20%, so that an appreciable number of rodents will be available for comparisons at terminal sacrifice. As discussed in Chapter 5, however, a proper evaluation of the experimental results will take into account all animals included in the experiment regardless of whether or not they survived to the end to the study. Adaptive terminal sacrifice methods have been studied by Louis and Orav (1985), but, at present, the proposed methods are not practicable.

Interim sacrifices: In many studies, a number of animals may be sacrificed at predetermined points in time during the course of the study. In a 24-month rat study, for example, ten animals in each treatment group may be sacrificed at both 12 and 18 months. This will facilitate study of the progressive pathogenesis of the lesion of interest and will ensure that both exposed and unexposed animals are available for purposes of comparison at these times. Efficient methods of planning adaptive interim sacrifices have been considered by Bergman and Turnbull (1983) and by Louis and Orav (1985), although these methods require rapid pathological examination.

Dose selection: Dose selection is one of the most controversial and important elements in the development of a protocol for a chronic bioassay. In addition, considerations behind the selection of appropriate dose levels depend to a large extent on the objectives of the study at hand.

In screening studies, the biological ideal would be to test only dose levels comparable to those to which humans are exposed. However, this is not practicable on statistical and economic grounds, unless the substance tested is an extremely potent carcinogen. There are potentially millions of humans exposed to many of the test substances under consideration, whereas it is usually feasible to expose only hundreds or perhaps a few thousand animals. Thus, a substance that causes the rate of some cancer in humans to increase from 1% to 2%, say, might cause tens of thousands of human cancers but might not be detected as a carcinogen even in a relatively large animal experiment. As it is not feasible to carry out tests involving millions of animals, the only solution is to use dose levels that induce measurable rates of response.

It is essential to have more than one dose level, for several reasons. One important purpose is to provide for the possibility that a misjudgement has occurred with the choice of a single high dose, resulting in either few animals surviving long enough for tumours to arise, or such severe toxic effects are seen that the relevance of the findings are doubtful. Secondly, a treatment effect that is dose-related over several levels of exposure is more convincing than one that is demonstrated only in a single-dose group. A third reason is to allow for the possibility that metabolic pathways used at a high dose may differ from those used for lower doses. If this consideration is ignored, a substance causing tumours at very high dose levels by a mechanism that does not occur at lower dose levels may be erroneously deemed unsafe for humans. A fourth reason is that it is reassuring if no large effect occurs at dose levels in the range to which humans are exposed.

The results of a recent survey of published guidelines of experimental designs for carcinogenicity tests are summarized in Table 3.1 (International Life Sciences Institute, 1984b). An examination of these guidelines indicates that a control and two or three positive levels have often been recommended. The highest dose should elicit minimal signs of toxicity to ensure that the test animals have been sufficiently challenged, yet not be so great as to result in appreciably decreased body weight or decreased survival, other than as a result of tumour induction. Lower doses are taken either to be specified fractions of the high dose or to lie within a certain range of the high dose.

The highest dose that satisfies the preceding criteria is often referred to as the 'maximum tolerated dose' or MTD (Munro, 1977). It is possible that the MTD could be different for males and females of the same species. If the difference between sexes is small, a common MTD can be employed. If the difference in sex-specific MTDs is appreciable, separate MTDs may be employed, at the expense of comparability between sexes, in order to maximize the sensitivity of statistical tests within each sex for increased tumour occurrence in the high-dose group relative to the control group.

While the guidelines given in Table 3.1 are generally quite specific, their underlying rationale is often less explicit. Most are vague with respect to recommendations concerning randomization. In this chapter, however, we will attempt to develop recommendations based on sound statistical principles of experimental design.

Table 3.1 Summary of selected recent guidelines on experimental design for carcinogenicity bioassays

Source	No. of dose levels (excluding untreated controls)	No. of animals of each sex per dose	High dose	Low dose	Intermediate doses
EPA[a] (1979)	3+	50	Induces slight toxicity but no substantial reduction in longevity due to effects other than tumours	Less than $\frac{1}{2}$ of intermediate doses but not less than $\frac{1}{10}$ of high dose	$\frac{1}{4}$ to $\frac{1}{2}$ of high dose
IRLG[b] (1979)	2+	No. of animals required to provide adequate assurance of safety if the test failed to detect carcinogenicity	Can be administered for the lifetime of the test animal and not (i) produce clinical signs of toxicity or pathological lesions other than those related to a neoplastic response, (ii) alter the normal longevity of the animals from toxic effects other than carcinogenesis, and (iii) appreciably inhibit normal weight gain		
IARC[c] (1980)	2	50	Elicits some toxicity when administered for the duration of the test period, but does not induce (i) overt toxicity, (ii) toxic manifestations which are predicted materially to reduce the life span of the animals except as the result of neoplastic development, or (iii) 10% or greater retardation of body weight gain as compared with control animals	$\frac{1}{4}$ or $\frac{1}{2}$ of high dose	
OECD[d] (1981)	3	50	Elicits signs of toxicity without substantially altering the normal lifespan due to effects other than tumours. For diet mixtures, the ingested concentration should not exceed 5%	Should not interfere with normal growth, development or longevity of the animal or result in any indication of toxicity. In general, not less than 10% of high dose	Mid-range between high and low doses depending upon the toxicokinetics of the chemical

[a] Environmental Protection Agency
[b] Interagency Regulatory Liaison Group
[c] International Agency for Research on Cancer
[d] Organization for Economic and Cooperative Development

Chapter overview

In Section 3.2, the basic principles of experimental design, including randomization, replication and stratification, are reviewed. Practical considerations involving the design and conduct of carcinogenicity trials are also noted. Experimental designs for screening studies are considered in Section 3.3. In addition to guidelines on sample size requirements for conventional bioassays, consideration is given to experiments involving multiple strains of test animals. Dose-response studies are discussed briefly in Section 3.4; special studies designed to elucidate certain aspects of carcinogenic mechanisms of action are treated in Section 3.5. Design considerations relating to histopathological analysis are considered in Section 3.6, while operational procedures involved in the acquisition and recording of experimental data are discussed in Section 3.7. A brief summary of the recommendations made in this chapter is given in Section 3.8.

3.2 Principles of experimental design

The primary purpose of experimental design is to ensure that the objectives of the study can be met and that valid, meaningful conclusions can be drawn from the results obtained. A good experimental design will also maximize the value of the information obtained by eliminating potential sources of bias, reducing experimental error to a minimum, and providing means of assessing experimental error. This is accomplished through proper application of the techniques of randomization, replication and stratification.

Experimental units

The term 'experimental unit' refers to the smallest unit of experimental material which is treated alike. In some studies, for example, several animals are housed in the same cage. It is thus possible that interactions between animals in the same cage, in a common environment, could result in cage effects. In this event, the cage rather than the individual animal would constitute the experimental unit.

The experimental evidence for or against the existence of cage effects is scant, because such effects have gone largely unassessed in past studies. In one large skin-painting study in which such effects were considered, a highly significant ($p < 0.01$) indication of cage effects was found among two of the three positive-control groups (Gart, 1976, p. 113). However, significant effects ($0.01 < p < 0.05$) were found in only three of 52 tobacco-condensate groups. No cage effect was noted with respect to neoplastic lesions in a bioassay of hexachlorobenzene conducted by Arnold *et al.* (1985).

While it avoids cage effects, individual housing may lead to increased stress in rodents (Hatch *et al.*, 1965; Sigg *et al.*, 1966), although conflicting evidence is available in this regard (Andervont, 1944; Fare, 1965). If multiple housing is used, on the other hand, feed and water consumption must be administered collectively, and animals may

be lost to cannibalism or the more rapid spread of communicable disease. In order to clarify these issues, some basic research on the desirability of group housing and the attendant possibility of cage effects is required from the point of view of proper experimental design.

The presence of other effects may also lead to an experimental unit which does not correspond to the individual animal. In two-generation studies involving exposure *in utero,* for example, animals from the same litter share common genealogical traits and are subjected to similar levels of transplacental exposure. As a result, the entire litter rather than the individual pup may constitute the experimental unit. Another example is that in which an entire column or contiguous group of cages is subjected to the same treatment. If there were a gradient within the laboratory influencing tumour occurrence, the ensemble of cages could conceivably represent the experimental unit. Such effects may be due to differences in susceptibility among the test animals, or to differences in environmental factors such as temperature, lighting, humidity and air flow (Fox *et al.*, 1979).

In a review of data from experiments on the carcinogenic potential of the food-colour additive, FD & C Red No. 40, Lagakos and Mosteller (1981) noted a correlation between the incidence of reticuloendothelial tumours and the animal's position on the shelf-level on racks, which persisted after adjustment for sex, dose and rack column. The tumour rates appeared to be higher in animals on the upper shelves. Animals in successive groups were allocated to successive rack positions. In this particular study, group 1 males went into the top three (of five) shelves of the front of rack 1, group 2 females went into the bottom two shelves of the front of rack 1 and the top shelf of the back of rack 1, and so on. Thus, shelf position introduced some bias into the treatment comparisons, which did not, however, affect the overall negative conclusions. Another example of potential positional effects is the study that has been reported by Greenman *et al.* (1984).

As our knowledge of clustering effects is still somewhat limited, the analysis of most studies continues to treat the individual animal as the experimental unit. This will of course be valid when the experimental design is such that the individual animal is the appropriate unit for purposes of statistical analysis, as in the completely randomized design with individual housing that is discussed in the next section. This may also be reasonable when any cage, litter or positional effects that may be present are negligible. However, empirical evidence to support this assumption would be reassuring.

Randomization of animals

As noted earlier, the purpose of randomization is two-fold. First, it ensures against potential biases in the experimental results. Second, valid statistical inference can be based on the permutation distribution induced by the randomization scheme employed. This avoids the need to make further assumptions concerning underlying statistical models for the experimental data. In order to avoid bias, it is essential that the predisposition to the response of interest be the same in all treatment groups. Bias will

be introduced if the animals in one group are more likely to develop tumours than those in another group.

To avoid bias, animals must be assigned randomly to treatment groups or cages. This is done usually by using so-called 'pseudo-random' numbers generated by a computer. Depending on the number of animals to be used, the number of animals caged together, and the number of experimental groups, a randomization list may be drawn up.

As an example, we consider the situation of 200 animals of the same sex to be assigned randomly to four groups of 50 with five animals from the same group caged together. The 200 animals, thus, have to be assigned to 40 cages and these to the four treatment groups. As a preliminary step, consecutive numbers should be arbitrarily given to all 200 animals; this may well be their order of presentation.

The randomization list will be a random sample without replacement from the 40 cage numbers, each being included exactly five times. For example, the first ten numbers in such a list may be:

Consecutive animal number:	1	2	3	4	5	6	7	8	9	10	...
Random cage number:	37	32	12	8	9	32	17	19	23	16	...

This would mean that animal No. 1 would be placed in cage No. 37, and animal No. 2 in cage No. 32, and so on. (Note that animal No. 6 will also be assigned to cage No. 32.) This procedure ensures that cage mates have been randomly selected.

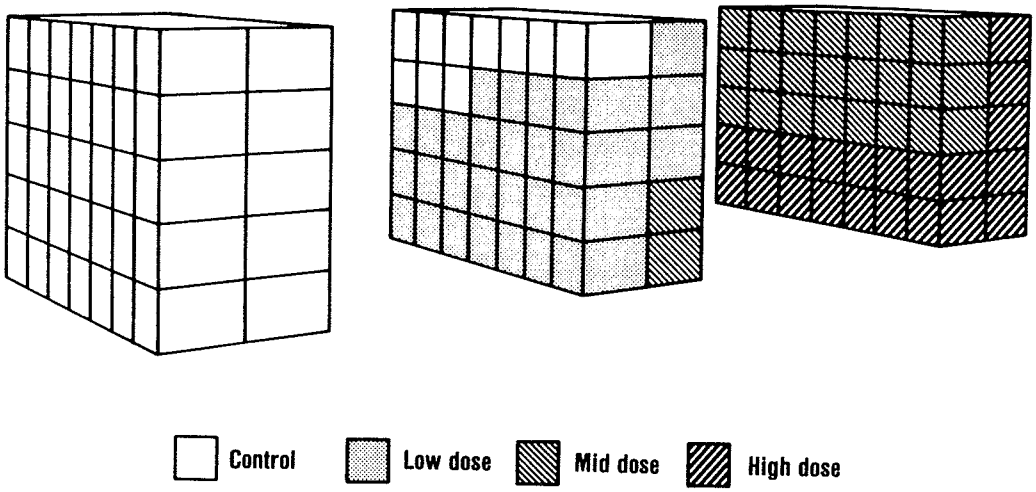
This complete randomization of animals into cages permits distribution of these cages in a deterministic way to the treatment groups. For example, the first ten cages are assigned to group No. 1, the second ten to group No. 2, and so on. However, a randomization list could once again be used, taking a random sample without replacement from the four group numbers, and using each number exactly ten times. This list, when matched to the consecutive cage numbers, would identify the group that was allocated to each cage.

Random location of cages

Randomization of the animals to cages and into experimental groups does not yet ensure that the predisposition to the response of interest is the same in all treatment groups. Care must be taken also that potential treatment effects are not confounded with environmental factors. An example of a design with potential for serious bias is shown in Figure 3.1 (Bickis & Krewski, 1985). Even if the animals have been randomized to cages, the existence of an environmental gradient, say, from left to right, would bias any comparisons between the various treatment groups. In particular, if tumour occurrence were enhanced along this gradient and an agent with no carcinogenic potential whatsoever were administered to the test animals, then the increasing trend in tumour incidence going from left to right would give the illusion of a dose-related effect.

A somewhat better design that has been used often in the past is the systematic cyclic design (Figure 3.2A), in which the doses are assigned to the cages in rotation.

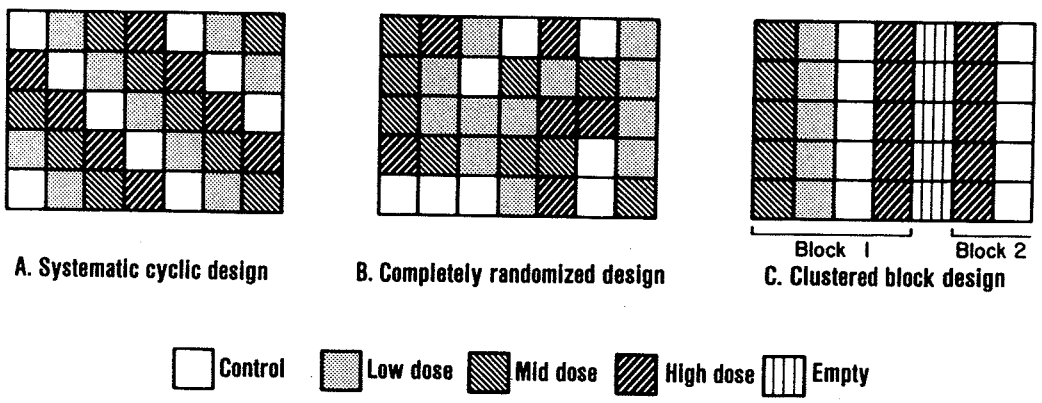
Fig. 3.1 Cage layout for a systematic block design (after Bickis & Krewski, 1985)



Although large biases may be avoided in this way, the use of a fixed sequence is potentially vulnerable to small biases. Another approach would be to assign the treatments to cages completely at random (Figure 3.2B). This procedure should provide adequate protection against environmental gradients with a moderate number of cages, and it has an element of simplicity that is readily exploited at the analysis stage (see Chapter 5). An objection to complete randomization that has been raised is that it increases the chance of misapplication of treatments; this problem may be controlled through careful record-keeping and the use of well-trained personnel. Visual devices, such as different coloured labels for the cages of different treatment groups, can be very helpful in this respect, although it does not allow for 'blind' delivery of treatment.

A more serious problem may be the potential for cross-contamination with volatile agents or through spillage of feed. In this case, the clustered block design in Figure 3.2C, in which all cages in the same column are treated identically, may be considered. With this last design, however, the column rather than the cage may form the most appropriate experimental unit.

Fig. 3.2 Cage layout for three experimental designs (front view of first bank of cages) (after Bickis & Krewski, 1985)



More complicated, Latin-square designs have been proposed by Lagakos and Mosteller (1981) as a means of balancing positional effects. However, their implementation requires that certain relationships be satisfied between the number of rows and columns in the cage rack and the number of dose groups. The symmetry which eliminates positional effects, moreover, would be disrupted easily by missing observations that cannot be avoided in long-term studies. Technically, such designs also require certain assumptions concerning the lack of interactions between rows, columns and treatments.

Another method of balancing positional effects and providing a more uniform environment is to rotate the cage positions during the course of the experiment. Although this procedure is sometimes used, it has not been subjected to the same careful study as the other designs. Nonetheless, some form of cage rotation may prove useful in designs with a high degree of clustering, such as the clustered block design. In this case, rotation of the columns would tend to minimize the effects of horizontal gradients. Similarly, frequent rotation of the positions of the three banks of cages in the systematic design shown in Figure 3.1 would serve to reduce the potential for serious biases.

While not necessary for purposes of bias reduction, the rotation of cages in a completely randomized design ensures a more uniform environment for all animals. By reducing this source of variation, it is possible that the sensitivity of any comparisons between treatments could be improved. For the same reason, consideration could be given to the periodic rotation of the cage positions within columns in the clustered block design or within banks in the systematic design. Until the contribution of environmental factors to the overall experimental error is more clearly defined, however, the potential gains from cage rotation remain unclear.

In any cage rotation scheme, some cages will become vacant due to deaths prior to the termination of the study. While the manner in which empty cages should be handled has received little attention so far, the inclusion of these cages in subsequent rotations will maintain the advantages of the original plan in terms of counterbalancing positional effects.

All of the preceding randomization schemes for cages are directly applicable when only one sex of the test animal is to be used. With protocols involving both sexes, there may be problems in housing males and females in adjacent cages. Because of this, it has been a common practice in the past to keep male and female animals apart, often in separate rooms. When this is done, any comparisons between sexes may reflect environmental as well as sex differences. With certain designs, however, it may be feasible to mix males and females to a limited extent. With the completely systematic design, for example, banks of males could be interspersed among banks of females. This would be advantageous when comparisons between sexes are of interest.

Regardless of the experimental lay-out and randomization scheme actually employed, care should be taken to document the precise placement of each cage and its corresponding treatment. Any subsequent alteration to this initial configuration should also be recorded meticulously. This information is essential in order to characterize the experimental design and to permit a valid statistical analysis of the experimental data.

Replication

The significance of any experimental finding will be greatly enhanced if the results can be reproduced. In order to assess reproducibility, some form of replication is necessary. Replication requires that several independent experimental units be subjected to precisely the same experimental conditions in order to provide an estimate of experimental error. This allows the statistical significance of any observed differences between treatment groups to be determined by comparing the variations between units in the same group.

In the completely randomized design with single caging, the individual animal is the experimental unit, and replication is achieved by increasing the number of animals allocated to each treatment. More information may be obtained if some grouping of subjects into homogeneous strata or blocks is done prior to the allocation (see section below). In this case, animals are assigned deliberately to the blocks to reduce variation, and then the animals within each block are assigned randomly to the treatments.

Another form of replication that has been used frequently in agricultural research, although not in carcinogenicity testing, is to repeat the entire experiment. In order that the individual experiments be comparable, they should all follow the same protocol. Aside from this constraint, however, there is no need to keep the repetitions uniform. It may, in fact, be advantageous to make them as diverse as possible. They could, for example, be carried out in different laboratories at different times and, if possible, use different suppliers of animals and feed and different batches of the test compound. If studies are replicated in this manner, it is possible to assess treatment-replicate interaction or the extent to which treatment effects may differ between replicates. The absence of such interaction provides additional assurance of the reproducibility of the result.

The use of replicates also has the advantage that the chance of discovering a susceptible subpopulation is increased (Haseman & Hoel, 1979). It is possible that the toxic effects of a compound are manifested only if certain other factors are present. These may be a genetic predisposition to toxic effects, the presence of certain other compounds in the diet, or, possibly, microbial or environmental influences. If only one replicate is used, then the chance of encountering the required set of conditions is less than if a number of replicates are employed under different conditions.

The demand for replication of experimental findings is inherent in the process of scientific evaluation. For example, in the *IARC Monographs* (IARC, 1982b), sufficient evidence for carcinogenicity in experimental animals is established only if 'there is an increased incidence of malignant tumours: (a) in multiple species or strains; or (b) in multiple experiments (preferably with different routes of administration or using different dose levels); or (c) to an unusual degree with regard to incidence, site or type of tumour, or age at onset.' Evidence based on replication across sexes and species is stronger than that based on experiments conducted under identical conditions.

Stratification

Many different factors, both genetic and environmental, can influence the process of tumour induction. For comparisons between groups to be as precise as possible, the

animals in different treatment groups and their environments must be as similar as possible, other than with respect to the treatment of interest. If factors affecting tumour occurrence can be identified before the conduct of the study, the test animals can be divided into more homogeneous strata, defined in terms of different levels of such blocking factors. Comparisons between treatments may then be made within blocks or strata, thereby eliminating interblock variation from the comparisons of interest.

Animals can be stratified by litter so that genetic variation is reduced within a block (Mantel *et al.*, 1977; Mantel & Ciminera, 1979). In some studies, animals have been divided into different weight classes before being assigned to different treatments, thereby stratifying by initial body weight. If environmental effects in the laboratory are of concern, cage position might be used to define strata with, for example, one bank of cages constituting a block. Other variable factors that can influence the tumorigenic process include the age of the animals at the start of the study and the source of the animals, including the particular shipment from the supplier.

While stratification represents a potentially useful tool for increasing analytical precision, it is not without its disadvantages. Since blocks are different by design, the variation between blocks provides no information on experimental error. If the test subjects are few, stratification can actually diminish the amount of information on the magnitude of the experimental error even if it actually reduces the error. In the extreme case in which no two animals in the same block are treated alike, experimental error is estimated from the inconsistency of treatment differences across the blocks, or treatment-block interaction. If effects of treatment do differ among blocks, then this inconsistency can lead to an overestimation of experimental error.

Control animals

Regardless of the objective of the study, some form of reference or control group, against which the effects of the treatment of interest can be judged, is essential. The nature of the control group may, however, differ depending on the study protocol. For screening studies involving exposure to one or more levels of the test agent, the appropriate control is simply a group of unexposed animals. For more elaborate mechanistic studies, one or more different types of control group involving different treatments may be required in order to isolate the effects of interest (see Section 3.5).

It is essential that the treated groups differ from the control group only with respect to the treatment of interest, and not with respect to any other aspect, such as diet, husbandry or observation. Any comparisons between the treated and control groups will reflect all differences between these two groups. Thus, in an experiment with two groups – one in which animals are exposed to cigarette smoke in smoking machines and the other an untreated control group – the only assessment that can be made is the effect of smoking combined with the stress due to the the animals' being placed in the machines. To test the effect of smoking only, one needs control animals that are 'sham-smoked', that is, they are placed in machines for the same length of time as the treatment animals but not exposed to smoke. Similarly matched control animals will

also be required when the treatment is applied by injection, with oral dosing by gavage or in a vehicle such as corn oil. Thus, if a chemical is administered in corn oil, the control animals should be administered corn oil without the chemical. Such control animals are termed 'vehicle control' animals to indicate that they have been subjected, as nearly as possible, to the method and route (i.e., the vehicle) of exposure experienced by the treated animals.

The question whether to include a positive control group, involving a known carcinogen, has been discussed extensively. However, it is generally agreed that in routine screening no positive control animals need be used (IARC, 1980). One reason is that carcinogens act by different mechanisms on different tissues, so that one would not necessarily know which positive control substance to choose when testing an agent of unknown carcinogenic potential. Furthermore, the inclusion of positive control groups introduces hazards for personnel and the risk of cross-contamination. However, when the objective is to assess the relative carcinogenicity of a range of treatments known to be carcinogenic (for example, cigarette-smoke condensate) and when the testing must be carried out in a series of studies, positive control animals are required, as it is otherwise impossible to compare reliably the carcinogenic potency of different treatments.

There has been some discussion about the use of two identical control groups (Society for Toxicology, 1982). With a completely randomized design, any difference between the two groups would be due solely to chance, so that, in effect, the two groups form one large control group. From the statistical point of view, differences between the two control groups could indicate systematic departures from complete randomization. Thus, while two control groups serve no useful purpose in a properly randomized experiment (other than to increase the total number of control animals), this practice could act as a quality control mechanism in terms of identifying unsuspected biases in design.

Haseman (1985) compared the response rates in the dual control groups used in a series of 16 bioassays of food-colour additives and found no significant difference between the response rates in the pairs of control groups used within the same study.

Criteria for evaluating experimental designs

The most important requirement of any experiment is that it should provide the imformation needed to meet the study objectives. In particular, it should be free from biases which may exist in the absence of randomization, and there should be a sufficient number of treatment groups to enable identification of the quantities of interest. Since there are generally many designs that satisfy these conditions, considerations of sensitivity and efficiency can be used to choose among them.

The sensitivity of an experiment is its ability to detect small differences. In screening studies, sensitivity is often quantified in terms of the false-negative rate, or the probability of not detecting a carcinogen of a given potency (see Section 3.3). Equivalently, sensitivity may be expressed in terms of the power, that is, the probability of detecting a carcinogen of a given potency. The expressions are equivalent because power equals one minus the false-negative rate. For dose-response

studies, where the objective is usually the estimation of some parameter of the dose-response curve, sensitivity may be measured by the standard error of the estimate.

In general, experiments with more animals tend to be more sensitive. There may exist one design that will be the most sensitive among all designs of a given size. However, such optimal designs are not always practical, for the following reasons. First, the optimal design may depend on parameters that cannot be determined until the experiment is completed. Secondly, an experiment may have several objectives, and the design that is optimal for one objective may not be optimal for another. Finally, the optimal design may not be feasible because of operational or other constraints. Nonetheless, the optimal design may still be used as a yardstick for gauging the efficiency of the actual design relative to the optimal one.

3.3 Designs for screening studies

Conventional studies

In a screening study, the purpose of the experiment is to arrive at a decision regarding the carcinogenicity of the test compound. Two types of errors are possible in making such a decision: an innocuous chemical may falsely be declared carcinogenic (Type-I error), or a carcinogen may incorrectly be considered harmless (Type-II error). The probabilities of these two types of errors occurring are termed the 'false-positive' and 'false-negative' rates, respectively (Table 3.2). These error rates depend on both the experimental design and the decision procedure used. Once the experimenter decides on a false-positive rate, that is, determines the risk he is prepared to accept for making the first type of error, the decision rule will be derived from this mathematically. The predetermined value of the false-positive rate is termed the 'nominal significance level'. It is then the function of the experimental design to minimize the false-negative rate.

In order to gain some idea of the sensitivity of a screening bioassay, consider a simple experiment in which $n = 50$ animals are assigned to both a control and a single test group. Suppose that there is no difference between the groups with respect to intercurrent mortality and, thus, that the proportions of animals with tumours in the two groups are compared using Fisher's exact test at a nominal significance level of

Table 3.2 False-positive and false-negative rates in carcinogenicity screening tests

Experimental evidence for carcinogenicity	Carcinogen	
	No	Yes
No	Correct decision	False negative
Yes	False positive	Correct decision

Table 3.3 False-negative rates for a simple carcinogenicity screening test[a]

Excess tumour incidence in test group (%)[b]	Tumour incidence in control group (%)				
	0	1	5	10	20
5	90	88	87	88	90
10	43	49	61	69	77
15	11	18	34	46	58
20	2	5	15	25	36
25	<1	1	5	11	19

[a] Based on Fisher's exact test ($\alpha = 0.05$) with 50 animals in each of a control and a test group and assuming that all animals respond independently

[b] Difference between the response rates in the test and the control groups, respectively

$\alpha = 0.05$. (For details of this statistical test, see Chapter 5.) As indicated in Table 3.3, the false-negative rate for compounds inducing an increase of 25% over the background incidence rate is less than 1% whenever the spontaneous response rate is low. These results also suggest that a carcinogenic compound tested at a dose level inducing only a 5–10% increase over the background incidence rate might well go undetected. However, the use of high doses tends to maximize the carcinogenic potential of the test compound and thereby minimize the risk of a false-negative result.

Similar results for group sizes of $n = 25$, 50, 75 and 100 are shown in Table 3.4. When the background incidence rate is low, the use of more than 100 animals per group will result in moderate false-negative rates, with compounds inducing tumours in as few as 10% of the exposed animals, but at nearly double the cost. The use of 25 animals per group would be effective only for compounds responsible for an increased risk well in excess of 25% in exposed animals.

Minimum sample sizes required to detect a carcinogenic effect of a given magnitude with Fisher's test procedure may be calculated by summing the probabilities of those outcomes that would result in a significant result. This approach has been used by Haseman (1978) to obtain sample sizes for select values of p_0 and p_1, the response probabilities in the unexposed and exposed groups, respectively.

A simple approximation to the minimum value of n required to achieve specified error rates for two given response probabilities, p_0 and $p_1 = p_0 + \delta$, has been developed by Walters (1979). (A detailed comparison of this and other approximations has been made by Chen, 1984.) This particular result is based on the standardized difference

$$\Delta = \sqrt{2n}[\sin^{-1}\sqrt{p_0 - (2n)^{-1}} - \sin^{-1}\sqrt{p_1 + (2n)^{-1}}]$$

between the two proportions following the application of a continuity correction and an arc sine transformation. In testing at a nominal α level of significance, the power $1 - \beta$ of the Fisher test will be approximately

$$1 - \beta = (2\pi)^{-\frac{1}{2}} \int_{z_\alpha}^{\infty} \exp\left\{-\frac{(x - \Delta)^2}{2}\right\} \mathrm{d}x,$$

Table 3.4 False-negative rates for a simple carcinogenicity screening test[a]

Excess tumour incidence in test group (%)[b]	No. of animals per group (n)	Tumour incidence in control group (%)				
		0	1	5	10	20
10	100	2	10	29	43	56
	75	12	21	41	55	66
	50	43	49	61	69	77
	25	90	89	85	84	87
15	100	<1	1	6	15	27
	75	1	3	15	27	41
	50	11	18	34	46	58
	25	68	67	67	71	77
20	100	<1	<1	1	3	9
	75	<1	<1	4	9	19
	50	2	5	15	25	36
	25	42	42	48	55	65
25	100	<1	<1	<1	<1	2
	75	<1	<1	1	2	7
	50	<1	1	5	11	19
	25	21	23	31	40	51

[a] Based on Fisher's exact test ($\alpha = 0.05$) with n animals in each of a control and a test group and assuming that all animals respond independently

[b] Difference between the response rates in the test and the control groups, respectively

where z_α denotes the $100(1-\alpha)$ percentile of the standard normal distribution. By iterating on n, the minimum size required to achieve power $1-\beta$ can be readily evaluated, given p_0 and p_1. This approximate procedure is computationally simpler than the direct approach, yet it yields results in excellent agreement with the exact results (Walters, 1979), as does the related closed-form expression of Dobson and Gebski (1986).

The minimum sample sizes required to achieve a false-negative rate of $\beta = 0.10$, using a nominal significance level of $\alpha = 0.05$ based on this procedure, are shown in Table 3.5 for selected values of p_0 and p_1. These results indicate that, when the background incidence rate is low, the use of 50–60 animals will permit the detection of effects involving about 15% of the exposed animals, subject to the specified error rates α and β. More animals would be required to detect a smaller effect or the same effect in the presence of a higher spontaneous response rate.

False-positive and false-negative rates may also be calculated while testing for increasing linear trends in proportion (see Chapter 5). Tests for linear trend may be based on large-sample chi-square statistics (Armitage, 1971, pp. 363–365) or on exact permutation tests (Cox, 1958; Thomas *et al.*, 1977). Chapman and Nam (1968) obtained an explicit form for the asymptotic power of the former test, and Nam (1984) provides an expression for the exact unconditional power of the latter test.

Because the computations required for these exact results are extensive, some

Table 3.5 Minimum group sizes required to ensure a false-negative rate of 10% or less[a]

Excess tumour incidence in test group (%)[b]	Tumour incidence in control group (%)				
	0	1	5	10	20
1	819	2661	9084	16 287	28 110
5	162	243	503	783	1 232
10	80	100	166	233	339
15	53	61	90	119	163
20	39	44	59	75	98
25	31	34	43	53	67

[a] Based on Fisher's exact test ($\alpha = 0.05$) with n animals in each of a control and a test group and assuming that all animals respond independently

[b] Difference between the response rates in the test and the control groups, respectively

simpler, approximate results are desirable. Nam (1984) has derived a modified formula for sample size determination when testing for linear trend, based on a normal approximation with a continuity correction. For the special case of three equally spaced doses (including the control at dose zero) with n animals per dose, the minimum value of n required to result in Type-I and Type-II errors of α and β, respectively, is given by

$$n = A[1 + \{1 + [2(p_2 - p_0)/A]\}^{\frac{1}{2}}]^2/[4(p_2 - p_0)^2].$$

Here, p_0 and p_2 denote the response probabilities for the control and high-dose groups, respectively (p_1 is not involved in this term because of the equal-dose spacing) and

$$A = [z_\alpha(2\bar{p}\bar{q})^{\frac{1}{2}} + z_\beta(p_0q_0 + p_2q_2)^{\frac{1}{2}}]^2,$$

where $\bar{p} = \sum_{i=0}^{2} p_i/3$, $\bar{q} = 1 - \bar{p}$, and z_α denotes the $100(1-\alpha)$ percentile of the standard normal distribution.

The number of animals, n, needed in each group to obtain 90% power with three equally spaced doses is given in Table 3.6. These results indicate that experiments with 50–100 animals per group will be effective in detecting a linear trend involving an increase of about 20% or more above the background incidence rate in the high-dose group. As with Fisher's exact test discussed above for pairwise comparisons, note that smaller sample sizes are required when the background rate is low.

Two-generation studies

One of the major considerations in the design of a two-generation bioassay is the selection of the second-generation animals. Studies on transplacental exposure using saccharin, styrene and amaranth have revealed considerable intralitter correlation (Grice *et al.*, 1981). Although actual bioassay data are required to determine whether or not litter effects are to be expected also for tumours in second-generation animals, it seems clear that the litter rather than the individual pup should be considered as the experimental unit for the purposes of statistical analysis (see Section 7.6).

Table 3.6 Number of animals per group required to obtain false-positive rates of 5% and false-negative rates of 10% based on tests for linear trend with three equally spaced doses

Tumour response rates			Number of animals per group
p_0 (control)	p_1 (low dose)	p_2 (high dose)	
0.02	0.04	0.06	420
0.02	0.07	0.12	112
0.02	0.12	0.22	44
0.10	0.12	0.14	1150
0.10	0.15	0.20	224
0.10	0.20	0.30	70
0.20	0.22	0.24	1860
0.20	0.25	0.30	328
0.20	0.30	0.40	93

In the presence of appreciable litter effects, the statistical power of a two-generation study will depend on the number of pups selected from each litter. In order to illustrate the effects of intralitter correlation on statistical sensitivity, consider a simple hypothetical experiment in which 48 animals of the same sex are to be selected from the second generation in both a control and a single test group. The required number of animals could be obtained on the basis of one per litter, or two per litter from 24 litters. Fewer than 24 litters would be required if more than two pups per litter were chosen.

To illustrate the power of such a study, the probability of detecting a carcinogenic compound that induces a tumour incidence rate of 50% at a site where the background rate is 10% is shown in Table 3.7 for selected values of the intralitter correlation ceofficient in the test group (Krewski *et al.*, 1984a). These results reveal that, for a fixed sample size (in this case, 48), the statistical sensitivity is reduced with increasing intralitter correlation when more than one pup is selected from each litter. The goal of maximum sensitivity may thus be achieved by selecting only one pup per litter.

Table 3.7 Probability of significance in a hypothetical two-generation bioassay in the presence of intralitter correlation

No. of pups selected	No. of litters	Probability of significance (%)		
		Intralitter correlation in test group		
		0.1	0.5	0.9
1	48	95	95	95
2	24	93	88	83
3	16	90	79	68
4	12	87	68	49

For reasons of economy, Mantel (1980) has suggested that two or three pups be selected from each litter. However, since the cost of a two-generation study is due primarily to maintaining the second-generation animals for a period of about two years and to the associated histopathological diagnoses, the savings involved in breeding fewer litters over a 100-day period in the first generation may represent only a small fraction of the total cost of the study.

Multistrain studies

Another consideration in screening studies is the use of several strains of the test species (Haseman & Hoel, 1979). In order to assess the relative merits of single-strain and multiple-strain experiments, consider the following simplified genetic model. Suppose that a certain stock of animals consists of ten homogeneous subgroups of equal size, each with a spontaneous tumour incidence rate of 5%. Suppose further that a certain chemical will increase the tumour incidence rate to 25% in a certain number of these subgroups, but that it will have no effect on the other subgroups. In this model, the entire stock is intended to represent a single outbred strain, while each of the subgroups is either highly resistant or highly susceptible to the chemical under test, with the entire stock less sensitive overall than the susceptible subgroups.

Suppose now that an experiment is to be conducted in which 150 animals are to be assigned to a test group and an additional 150 animals to a control group. One strategy would be to choose these animals from the entire stock available. Alternatively, one or more of the ten subgroups might be selected at random, and a separate experiment conducted for each of the subgroups chosen. If two subgroups were selected, for example, two experiments with 75 animals in each of a treated and a control group would be involved. The first strategy corresponds to the case of a single outbred strain while the second strategy corresponds to the case of several inbred strains, with the total number of animals fixed at 150 in both cases.

The probabilities of detecting the carcinogenic potential of the chemical under test with these alternative strategies are shown in Table 3.8. These probabilities represent

Table 3.8 Probability (%) of detecting a carcinogen in multistrain experiments of equal size, assuming strains are chosen randomly from ten different strains

No. of strains	No. of animals per strain in each test group	No. of susceptible strains[a]			
		0	1	2	3
0[b]	150	3	12	30	53
2	75	4	23	39	54
3	50	3	28	48	64
5	30	2	30	51	66
6	25	2	32	54	70
10	15	3	32	53	67

[a] Tumour incidence increased from 5% to 25%
[b] Single outbred strain

the chances of observing a statistically significant increase in tumour incidence in any of the strains tested. When none of the ten subgroups is susceptible, the probabilities shown represent the chances of a false-positive result. Even though the multistrain experiments involve one statistical test for each strain, the overall false-positive rates are comparable regardless of the number of strains tested.

In all cases considered, the power of the strategy using multiple inbred strains exceeds that of the strategy which uses a single outbred strain. When two susceptible subgroups are present in the population, for example, the chances of detecting the carcinogen by using a single outbred strain are about 30%. With a multistrain experiment involving three randomly selected inbred strains, however, the chances of detecting the carcinogen are increased to about 48%.

The assumptions involved in the preceding example are no doubt over-simplifications. In practice, it is unlikely that the inbred strains used would be selected at random. Moreover, the model of the outbred population as a mixture of homozygous subpopulations does not provide for the heterozygosity present in actual outbred stocks.

While the choice of the incidence rates of the spontaneous and induced tumours is not critical, the use of a large number of inbred strains in a multistrain experiment will result in an insufficient number of animals per strain on which to base a meaningful analysis. If, for example, the design was constrained to 50 rather than 150 animals, the use of ten strains would result in only five animals per group. Another concern of a more practical nature is the difficulty of finding ten unrelated strains. Thus, while carcinogenic effects might be detected more readily by selecting several inbred strains for study when susceptible subpopulations do exist, some experience with this approach is required before it may be recommended as standard practice.

Sequential designs

In any carcinogenesis screening programme of a large pool of chemicals using animal experiments, some compounds will induce large tumour increases and can thus be classified easily as carcinogens, some compounds will induce no tumour increase of note and can thus be classified easily as noncarcinogens, but other compounds will give equivocal results which lead to no clear-cut classification. In the last case, decisions must often be made on the basis of the equivocal results, even when further testing would be advisable. Multistage experimental designs have been proposed to alleviate the problem of making decisions based on equivocal experimental results (Elashoff & Beal, 1976; Elashoff & Preston, 1977; Elashoff *et al.*, 1979). After each stage in a multistage experiment, one can either stop the experiment, if the classification of the compound as a carcinogen or noncarcinogen is clearly indicated by the data, or go on to the next stage, if the evidence regarding carcinogenicity is equivocal. By using a multistage design, it should be possible to lower the false-positive and false-negative rates by reducing the number of decisions made on the basis of equivocal experimental results.

Elashoff *et al.* (1979) considered a particular two-stage design in some detail; they compared the operating characteristics of a screening programme based on their

two-stage design with those of a typical one-stage design, in which a control group of 50 animals is compared to two groups, each of 50 animals, exposed to different dose levels of the test compound. The first stage of their two-stage design is identical to the one-stage design, except that each group has only 35 animals. If the results of the first stage lead to a clear-cut classification of the test compound, then the experiment is terminated. If, however, the first stage gives equivocal results, then the second stage is conducted. The second stage compares a control group of 35 animals to a single exposed group of 35 animals. The evaluation of the second stage is simplified by the fact that any target organs have been identified in the first stage. Thus, the analysis of the second stage is restricted to the identified target organs, and nominal significance levels can be increased somewhat. Elashoff *et al.* (1979) showed that, using their two-stage design, a particular (hypothetical) large pool of chemicals could be screened in a shorter time and with a greater savings of animals than using the one-stage design. These savings were accomplished without increasing the false-positive or false-negative rates of the screening programme.

Although more compounds may be tested in a given period of time using such designs, individual multistage experiments will take longer to complete than single-stage experiments whenever a decision is not reached at the end of the first stage. As with the single-stage designs, moreover, it is still possible that a clear-cut classification concerning carcinogenicity may not be obtained following the completion of a two-stage test. For these reasons, multistage designs are more likely to find application in large-scale bioassay programmes involving many compounds than in studies in which a single substance is to be assessed.

3.4 Designs for dose-response studies

While a screening study is a useful tool for assessing carcinogenic potential on a qualitative basis, a dose-response study is required in order to describe the characteristics of the dose-response curve in more quantitative terms. This is particularly important when an identified carcinogenic hazard cannot be removed readily from the environment.

The evaluation of dose-response data is often done following a specified period of exposure, such as 24 months for rats and 18 months for mice. More generally, the probability of tumour induction may be considered to be a response surface $P(t, d)$ depending on the exposure time t and the dose d (Figure 3.3). Generally, the response probability $P(t, d)$ may be expected to increase with both dose d and time t.

The dose-response curve thus depends on the exposure time, as illustrated in Figure 3.4 by the data on liver and bladder tumour induction as a result of exposure to 2-acetylaminofluorine (2-AAF) from the ED_{01} study conducted by the US National Center for Toxicological Research (Littlefield *et al.*, 1980a). This study included large groups of mice exposed to increasing doses of 2-AAF for periods of 18, 24 and 33 months. While the general shapes of the dose-response curves for either bladder or liver tumours are similar at each of the three exposure times, the dose-response curve does vary somewhat with time. The results also indicate that the risk at all dose levels increases with the period of exposure to 2-AAF.

Fig. 3.3 The probability of tumour induction $P(t, d)$ represented as a response surface depending in time t and dose d (after Krewski *et al.*, 1983)

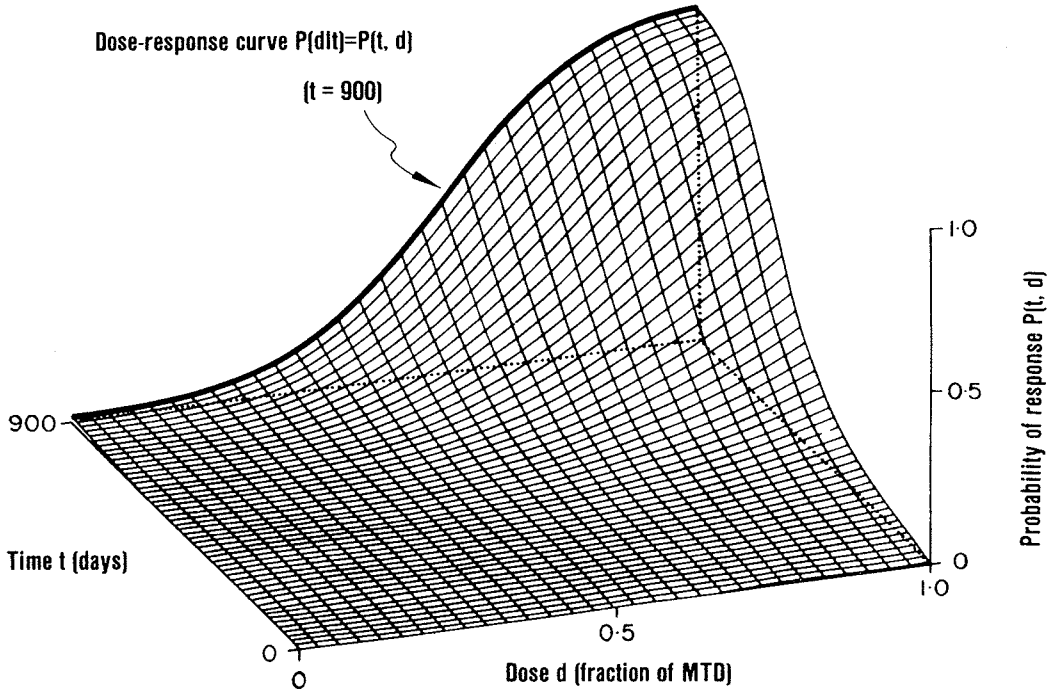
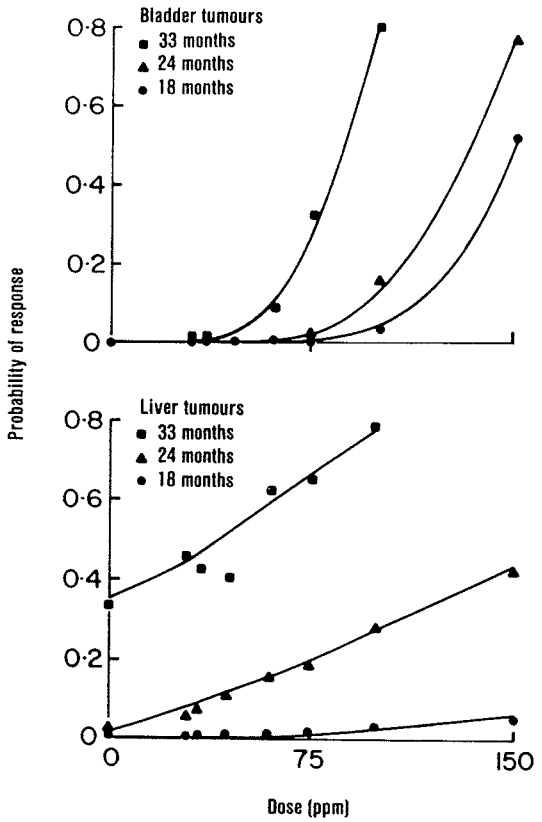


Fig. 3.4 Dose-response curves for bladder and liver tumours induced by 2-AAF following 18, 24 and 33 months' exposure (after Littlefield *et al.*, 1980a)



In the absence of any prior information on the shape of the dose-response curve, any reasonable series of increasing nontoxic dose levels may be used. Generally, the more dose levels, the greater the resolution of the dose-response curve. The largest dose-response study conducted to date, the ED_{01} study at the US National Center for Toxicological Research (Cairns, 1980), involved seven doses of 2-AAF ranging from 30–150 ppm in the diet, and an untreated control group, with between about 100 and 1700 Charles River BALB/c female mice at each dose. In total, 24 192 mice were used. Although the lowest and highest doses were separated by a factor of only five, the shape of the dose-response curves for bladder and liver tumours was well determined. (This compact design was possible due to the conduct of a smaller, pilot study.)

Another study of comparable size was commissioned by the UK Ministry of Agriculture, Fisheries and Food (Peto *et al.*, 1984). This involved a total of 5000 rodents (mice, hamsters and rats) administered different nitrosamines in the drinking-water. The main experiment involved one control and 15 dose groups of 48 males and 48 females each, with doses ranging from 0.033 to 16.896 ppm. In Section 4.3, we give the data from this study on the occurrence of pituitary tumours in male Colworth rats exposed to *N*-nitrosodimethylamine.

The two experiments described above are atypical because of their magnitude. Cost considerations normally dictate the conduct of smaller studies involving one control and three or four dose levels (see Table 3.1). Although less definitive, even these smaller studies can provide valuable information on the shape and nature of the dose-response relationship.

Because human exposure to most environmental carcinogens is low, data from dose-response studies are sometimes used to estimate the probability of tumour induction at low-dose levels. Since direct estimates of small probabilities are not feasible with small experiments (Kalbfleisch *et al.*, 1982), indirect estimates based on the extrapolation of results obtained at high doses are necessarily obtained using a particular dose-response model (Krewski & Van Ryzin, 1981; Kalbfleisch *et al.*, 1983). (Possible models are discussed in detail in Section 6.2.)

The problem of selecting the most suitable experimental design for puposes of low-dose extrapolation has been the subject of several investigations. However, the optimal design depends to some extent both on the assumed model and on the criterion used to compare competing designs. Experimental designs which minimize the asymptotic variance of maximum-likelihood estimates of risk, using a variety of three-parameter response models, have been investigated by Krewski *et al.* (1984b). With only three parameters in the dose-response model, the optimal design will involve only three dose levels (Chernoff, 1972). In Table 3.9 we give the response rates at three dose levels for different mathematical dose-response models (details of these are given in Section 6.2). This table shows the corresponding proportional allocation of animals which would lead, for the given three-parameter model, to the most precise estimation of the dose corresponding to a risk of 10^{-5}. Note that, for many agents, the response rate at the high dose in the optimal design may exceed that found at the MTD.

Taking the MTD into consideration, the optimal designs were generally found to involve three treatment groups, including one group at the MTD and one control

Table 3.9 Three-dose unrestricted optimal design for low-dose extrapolation to a risk of 10^{-5} under the three-parameter probit, logit and Weibull models

Model	Optimal response rates[a]			Optimal allocation (%)[b]		
	p_0	p_1	p_2	c_0	c_1	c_2
Probit	0.01	0.129	0.953	13	52	35
	0.05	0.222	0.963	19	45	36
	0.10	0.296	0.968	21	43	36
Logit	0.01	0.154	0.931	10	47	43
	0.05	0.244	0.945	15	40	45
	0.10	0.316	0.953	16	38	46
Weibull	0.01	0.199	0.976	11	57	32
	0.05	0.296	0.980	18	49	33
	0.10	0.370	0.983	20	46	34

[a] Response rates at the three optimally selected doses (In all cases, p_0 was found to correspond to the spontaneous response rate)

[b] Percentage of animals to be allocated to each of the three optimal doses

group. The dose given to the intermediate dose group depends on the curvature of the dose-response curve, with greater curvature requiring a larger fraction of the MTD. The allocation of animals among the dose groups depends on a number of parameters, including the acceptable risk and background response rate. However, a 1:2:1 allocation, with half of the animals on the low dose and the other half divided evenly between the control and high-dose groups, appears to result in a reasonably efficient design. An interesting property of these designs is that they are practically independent of the particular model assumed. In addition, designs for which dose placement and animal allocation differ moderately from those of the optimal design maintain high efficiency. One disadvantage of such optimal designs is that one has to know something of the shape of the dose-response curve in order to determine the optimal design.

For this reason, both Krewski *et al.* (1984b) and Portier and Hoel (1983b) have considered the efficiencies of suboptimal designs that may be expected to perform reasonably well in a variety of situations. In spite of the different approaches and models, both of these investigations have yielded similar conclusions. The former investigators proposed a design with one control and three equally spaced groups, at 0, $\frac{1}{3}$, $\frac{2}{3}$ and 1 times the MTD or at 0, $\frac{1}{4}$, $\frac{1}{2}$ and 1 times the MTD. Both 1:1:1:1 and 1:2:2:1 animal allocations for these doses were found to perform well. Portier (1981) has recommended a design with similar dose levels and a 2:3:3:2 animal allocation. These designs are again similar to those recommended by Gaylor *et al.* (1985a), who attempted to obtain the tightest possible confidence limits rather than to minimize the variance of the low-dose risk estimates.

The question of which mathematical model to use is of great concern in low-dose extrapolation (see Section 6.1). Both Chambers and Cox (1967) and Crump (1982) have developed optimal designs for discriminating between two specified dose-response models. (Related results for assessing the goodness-of-fit of multistage models have

been given by Portier & Hoel, 1984a.) Unfortunately, their results are not particularly encouraging. Even with such optimal designs, several thousands of animals are required to permit reasonable discrimination between two plausible models.

Because of the uncertainty regarding the shape of the dose-response curve in the low-dose region, some form of linear extrapolation is often employed in an attempt to obtain an upper limit on risk. Low-dose linearity may occur with carcinogens that interact directly with genetic material (Hoel *et al.*, 1983). These authors also demonstrate that nonlinearity at higher doses may be due to saturation effects in the metabolic activation process. Another argument leading to low-dose linearity is that of dose-wise additivity (Crump *et al.*, 1976). Under this hypothesis, spontaneous lesions may be modelled as arising from an effective 'background dose' of the test agent.

A simple procedure for linear extrapolation (Van Ryzin, 1980) involves extrapolating linearly from the 1% or 10% response point, based on a suitable model. An important feature of the above designs is that they are also nearly optimal for this procedure, at least in terms of minimizing experimental error. Other designs may be required in order to minimize systematic errors in model specification as well as random experimental error (Lawless, 1984).

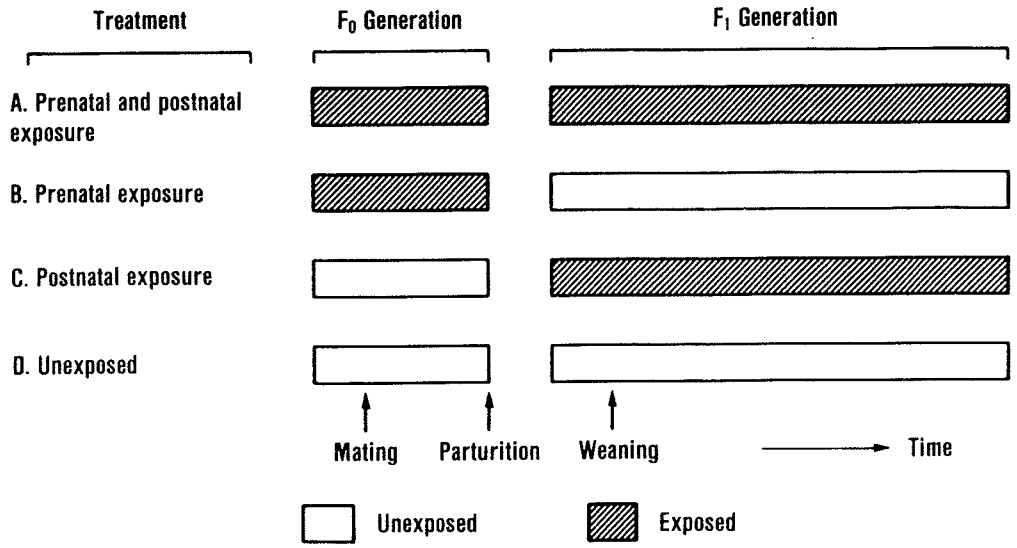
3.5 Designs for studies of mechanism

Designs for studies that examine certain hypotheses concerning the mechanisms of carcinogenesis generally need to be tailored to a particular experiment. Because most past data have been collected for the purpose of screening chemicals, however, the statistical procedures required to assess adequately the data from such experiments still remain to be established. Nonetheless, some general observations on the design of specific types of studies are given below.

Two-generation studies may be used in cases in which the test agent may exert its effect after exposure *in utero* (Grice *et al.*, 1981). Although prenatal exposure to certain nitroso compounds can induce tumours in young animals (Ivankovic, 1973), other compounds may require two-generational exposure in order to exert their full carcinogenic potential. The latter form of exposure was long thought to be necessary to obtain bladder lesions with saccharin (Arnold *et al.*, 1983a), although recent results (Shubik, 1985) have demonstrated that exposure from birth onwards may be sufficient. Generally, however, in order to distinguish between effects induced prenatally and postnatally, two-generation dosing regimens need to be included in the study protocol (Figure 3.5).

Another form of study of mechanisms addresses the multistage nature of carcinogenesis. In particular, there is now a substantial body of literature which suggests that tumour formation is a multistage process (Pitot & Sirica, 1980; Clayson, 1981). Initiation is thought to involve direct interaction between the proximate carcinogen and cellular DNA, although subsequent promotional events may be required for tumour development. In order to test a particular initiator/promoter pair in this model system, however, several dosing regimens involving various applications of the initiating and promoting agent may be required (Williams *et al.*, 1981). The primary endpoint in most two-agent designs, such as those described in Figure 3.6, is the development of

Fig. 3.5. Possible dosing regimens in a two-generation cancer bioassay (after Bickis & Krewski, 1985)



papillomas. More complicated dosing regimens, often using more than two test agents, are required to investigate theories regarding different stages of promotion in papilloma formation and the number of genetic alterations required to transform a normal cell to a carcinoma (Hennings, 1986).

A third type of study design would be necessary to investigate the effect of discontinued exposure on the development of neoplastic or preneoplastic lesions (Day

Fig. 3.6 Possible dosing regimens in an initiation/promotion study (after Bickis & Krewski, 1985)

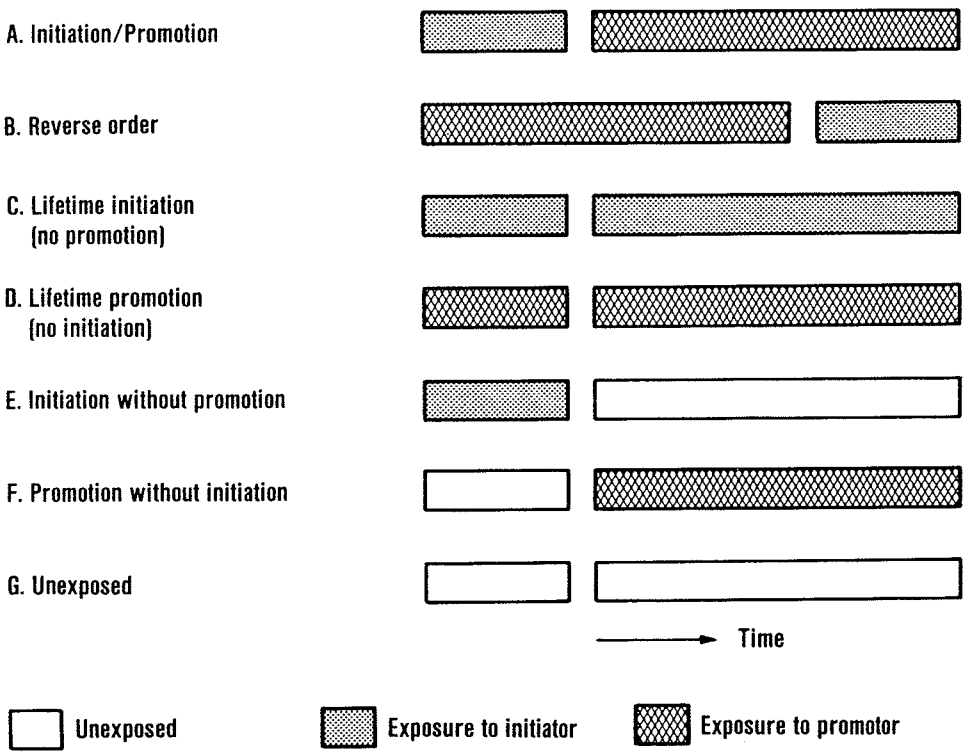
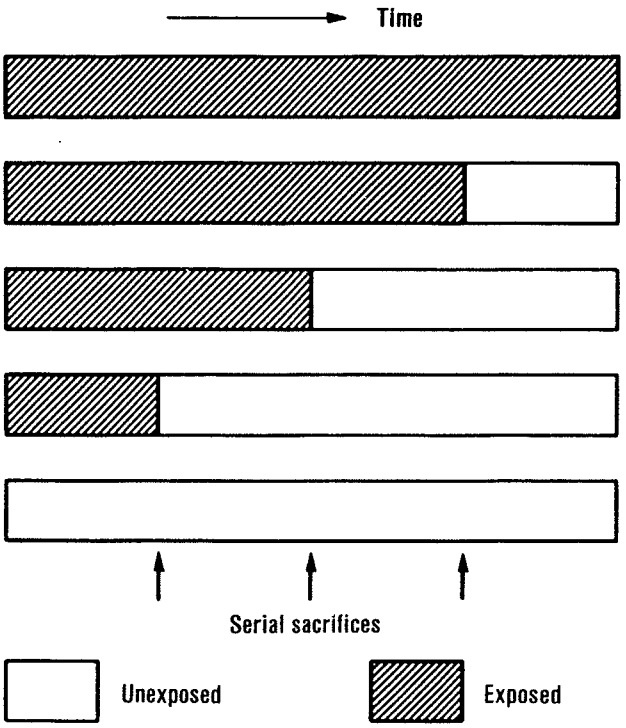


Fig. 3.7 Possible dosing regimens in a discontinued exposure study (after Bickis & Krewski, 1985)



& Brown, 1980; Arnold *et al.*, 1983b). In this case, dosing regimens involving exposure to the test agent followed by a return to control conditions would be employed, as illustrated in Figure 3.7. Both exposure and non-exposure periods of varying duration could be employed to explore the reversibility hypothesis in greater detail, as in the ED_{01} study (Cairns, 1980). The results of that study indicated that bladder neoplasms induced by 2-AAF occurred early in the study but were dependent on continuous exposure. In addition, moderate or severe hyperplasia apparently regressed when 2-AAF feeding was discontinued (Littlefield *et al.*, 1980b). In contrast, liver neoplasia occurred late in the study and did not require continuous exposure to 2-AAF (Littlefield *et al.*, 1980).

While the preceding designs focus on different aspects of the effects of exposure to a single toxicant, it is well known that the effect of certain carcinogens is much greater in combination than singly, as is the case with lung cancer mortality observed among asbestos workers who also smoke (Hammond *et al.*, 1979). Because of the ever-increasing number of potentially toxic agents present in the human environment, mixed exposures need to be evaluated for safety (Freundt, 1982). In order to explore the possibility of interactions between chemicals and other test agents, experiments involving exposures to mixtures of toxicants are required in order to evaluate synergistic or antagonistic potential (Wahrendorf & Brown, 1980; Wahrendorf *et al.*, 1981; Abdelbasit & Plackett, 1982; Métivier *et al.*, 1984).

In designing such multifactorial studies, investigators should clearly identify and distinguish the relevant factors which can be or must be manipulated in the design. Such factors include those which are essential for biological and technical reasons, such

as sex, batch of animals, solvents of the exposure, or others. The different factors that are intended to be manipulated deliberately must then be defined very concisely. These factors may be two chemicals, one chemical and one radiation exposure, or individual constituents of a commonly-found, complex mixture, such as betel quid or tobacco-smoke condensate.

The number of levels at which each factor can be investigated meaningfully has to be decided upon. A complete factorial design would then include as many experimental groups as the product of the number of levels of all the considered factors. Note that an untreated or appropriately vehicle-treated control level of each factor has to be included. For example, Métivier *et al.* (1984) report an experiment in which four dose levels (that is, unexposed and three doses) of exposure to an aerosol of [^{239}Pu]-plutonium oxide were combined with two (that is, vehicle treatment and one dose) levels of intratracheal benzo[*a*]pyrene instillation, resulting in $4 \times 2 = 8$ experimental groups.

In general, however, the number of experimental groups and the corresponding number of animals required in such studies may be large, even if only two factors are being studied. In order to permit a clear analysis of the individual and combined effects, an unexposed level should be included for each factor, and all possible combinations should be maintained in the experimental design.

3.6 Histopathological analysis

Histopathological analysis forms perhaps the most critical component of the long-term carcinogenicity bioassay. Since important conclusions concerning the determination of carcinogenicity are based on comparisons of patterns of tumour occurrence between exposed and unexposed animals, it is imperative that pathological examinations follow the most stringent standards of quality, uniformity and objectivity (Ward *et al.*, 1978; Ward & Reznick, 1983). This requires adherence to proper procedures at the time of necropsy including gross tissue examination, the selection of tissues to be examined, the sectioning and histological preparation of tissue samples and the evaluation of tumour morphology.

Although many lesions may be detected readily following gross examinations, it is also essential that all tissues of concern be subjected to histological examination, since many microscopic lesions not visible at the time of necropsy can be detected using modern histological procedures (Kulwich *et al.*, 1980). For example, more than 50% of the neoplastic lesions present in organs such as the thymus, lung, adrenal, Harderian gland and urinary bladder were not apparent during the gross examination in studies done at the US National Center for Toxicological Research (Frith *et al.*, 1980).

Generally, past practice in evaluating histopathological data has been to indicate only the presence or absence of particular lesions at specified sites. This can be refined in two ways. First, since many carcinogenic effects progress through a series of stages involving minimal to advanced neoplastic changes, it is often possible to assign a grade to such changes. For example, tumour status might be categorized on a scale from 0 to 5, indicating the absence of a lesion, or minimal to severe effects. Second, morphometric techniques may be used to quantify the extent of such changes (Kuiper-Goodman *et al.*, 1976). The objective of both grading and morphometric

analysis is to provide more detailed information on which to base statistical analyses, and, hence, to obtain stronger conclusions concerning the carcinogenic potential of the test agent. Nonetheless, these two procedures require considerable additional time and effort, and are thus not yet widely applied.

The question whether pathological evaluation of histologically prepared tissues should be performed without knowing the treatment group of the subject has been the topic of considerable debate over the years (Weinberger, 1973, 1979; Fears & Schneiderman, 1974). From the statistical point of view, it is desirable that pathology, both gross and histological, should be done 'blind'. This does not mean that the pathologist should be given simply a numbered slide and asked to identify the lesions present. All tissues of the animal should be evaluated as a unit, and the entire clinical history of the animal should be available to the pathologist. Nowhere on this record, however, should there be any indication of the animal's treatment group; in fact, the pathologist need not know if the individual animals he is diagnosing come from different treatment groups. All animals should be considered equivalent, except from what he can observe. For this reason, the animals should be presented in a random order.

The argument is sometimes raised that, unless the pathologist knows which is the high-dose group, he has difficulty diagnosing effects since he does not know what kind of lesion to expect from the treatment. In this case, a pilot experiment might be advisable, or else a satellite group of high-dose animals may be included for the pathologist to examine. In order to obtain an unbiased assessment, however, the results from these animals would not be used in assessing the significance of those from the main study.

Another possible approach is to have the pathologist examine a selection of control and high-dose lesions in order to familiarize himself with the nature of the lesions to be diagnosed. After this initial examination, these same slides would then be re-read 'blind' and in a random order along with those from the remaining animals.

While blind reading is clearly preferable, the avoidance of 'diagnostic drift' or time-related changes in the evaluation of the slides is perhaps of more serious concern. Thus, the difference between blind studies and those read non-blind but in a random order may be less important than the difference between studies read non-blind but in random order and those read non-blind in some systematic order.

Because of the large number of tumours examined in a typical bioassay, methods for reducing the histopathological workload by examining only a sample of the slides have been considered by Fears and Douglas (1977a,b). According to their proposal, a complete set of slides would be read only if it appeared necessary on the basis of the sample results. Although this offers some potential savings in costs, current practice is to subject all slides to a thorough histopathological examination in order to obtain as much information as possible.

3.7 Recording of experimental data

An adequate system for collecting, processing, reporting and storing large amounts of data is an essential part of the design of any long-term bioassay and is most easily organized using a computerized data storage and retrieval system (Naylor, 1978; Cranmer *et al.*, 1978; Konvicka *et al.*, 1978; Felsky *et al.*, 1979; Lawrence *et al.*, 1979; Herrick

et al., 1983). This could involve separate subsystems for the different data sets generated during the conduct of a long-term study, such as those for feed and water consumption, body weight, haematology and pathological findings. Another approach is to use one master system to acquire and integrate data from these different sources and to integrate them into a common data base. This latter approach is advantageous in laboratories where the studies are all of a similar nature, but it is less flexible in situations where the experiments are more variable.

The pathologist should record data in a systematic way so that the information is readily transferable to the computer *via* a system suitable for this purpose (Frith *et al.*, 1977; Naylor, 1978; Faccini & Naylor, 1979). Care should be taken to ensure that the same lesion is always described in the same way, and that quantitative assessments are given, where possible, of size, number and severity of observed lesions. Organs for which sections are not available, or which are too autolysed for examination, should be clearly marked. Unless indicated to the contrary, failure to mention a particular lesion described in another animal should always imply that the lesion was not present. Methods by which the pathologist's report and the statistical analysis are both generated by computer from the same input data source are to be preferred over systems in which the report is dictated and the extraction of data represents the first step in statistical analysis.

Systems are now available by which the pathologist can enter data directly *via* a computer terminal. However, it is not clear that this is the most practical method, as it requires the pathologist to move continually from the microscope to the terminal. Roe and Lee (1984) have developed a complete system for recording, reporting and statistical analysis of histopathological data from animal studies. Pathologists may enter data directly into the computer, or onto *pro forma* (which can be modified according to requirements) which are then subsequently processed for computer input.

Care should be taken also to guard against variation in standards of diagnosis both among different pathologists and in different time periods for each pathologist. Where the experiment is so large that more than one pathologist is required, it is important to avoid systematic differences in recording lesions which may result in a misleading indication of treatment-related effects. One simple procedure is to ensure that each pathologist has the same proportion of animals from each treatment group. A better method may be to assign different pathologists to different tissues across the board. Whatever approach is adopted, the identity of the pathologist responsible for a specific diagnosis should be recorded.

It is also important to guard against changes in standards with time. For example, if a pathologist starts with the first animal in the first group and works systematically through to the last animal in the last group, dose-related trends in the severity of some lesions might not reflect any true treatment effect. To avoid bias due to diagnostic drift, the slides should be read in random order, or in an order which avoids any systematic tendency for all the animals in one dose group to be read in advance of the remaining animals. This is particularly important with respect to lesions which pose diagnostic difficulties or those scored according to their degree of severity. Whatever scheme is used, the date and time of the pathologist's determination should be recorded.

Care should also be taken to ensure consistency between the microscopic and

macroscopic findings. Sections of suspect tumours noted macroscopically *post mortem* should be available for examination microscopically, and missing organs should be noted.

The data base for statistical analysis should include not only the pathologist's findings, but also the time at which the tumour was first noticed. This is feasible for visible or palpable lesions, and, in other cases, will correspond to the time of necropsy. For some experiments, the time at which lesions first reached different, specified sizes will also be recorded.

Ideally, it would be desirable also to record whether each tumour was the underlying cause of death. Sometimes this is an easy process, but more often it is not possible to give a reliable answer. One practicable compromise is to devise a four-point scale in which tumours are categorized as being 'definitely not', 'probably not', 'probably', or 'definitely' responsible for the death of the animal. A first estimate should be obtained at gross necropsy, with the possibility of revision later by histological examinations.

Finally, as we have noted, it is important that any relevant design features of the experiment are available for statistical analysis if required. These include not only details of the dose and the duration of treatments applied, but also details of cage placements, the name of the pathologist, and number of the animal and the time it was examined.

3.8 Summary and recommendations

In this chapter, we have provided an overview of the design of long-term animal experiments that examine potential carcinogenic effects of a test agent. It is clear from this discussion that the design of animal carcinogenicity experiments is a complex issue and that no hard and fast rules for their conduct can be laid out in advance. Nonetheless, a number of general statistical and other principles may be identified which should assist in the design of individual studies.

The preferred design in specific cases will be strongly dependent on the study's objectives. In this chapter, we have identified three broad categories of designs appropriate for screening, dose-response and mechanistic studies. Many of the studies conducted to date were intended to identify the presence or absence of carcinogenic activity in qualitative terms only; these fall into the first category. Screening studies of this type may involve only one or two dose groups and an unexposed control group. In order to maximize the chances of observing a carcinogenic effect in a relatively small population of test animals, moderate to high doses are used generally, with exposure continuing throughout the major portion of the animals' normal lifespan. The highest dose is however subject to the criteria for the MTD, particularly with respect to the requirements for minimal effects on body weight and overall survival (with the exception of increased mortality attributable to tumour induction). Lower doses are included often for confirmatory purposes and to ensure against the possibility that the high dose may have been exceeded. Commonly used sample sizes of about 50 animals per group are reasonably sensitive to effects involving 20% or more of the exposed population in cases where the background rate of occurrence of the lesion of interest is low.

A number of modifications to these conventional designs for screening studies, including the use of two-generation studies incorporating perinatal exposure, have been discussed in the literature. Multistrain studies involving the use of a variety of tester strains (generally with fewer animals per strain) have been proposed to increase the chances of selecting susceptible test groups. Sequential designs have been considered also as a means of decreasing the average cost and time involved in assessing a large number of test agents. These involve the conduct of a smaller study at the first stage, followed, when necessary for purposes of clarification in equivocal cases, by a second, confirmatory stage. While both of these latter proposals have merit, little past experience exists on which to base a sound evaluation of their properties. Thus, their use is recommended with caution.

Dose-response studies differ from screening studies in that additional dose levels are usually employed in an attempt to define more clearly the shape and nature of the dose-response relationship for the test agent. While elaborate studies of this type have been conducted with a limited number of compounds, such as 2-AAF and various nitrosamines, the use of one control and only three or four exposed groups is more common. Although optimal designs for low-dose risk assessment can be developed under specific parametric assumptions, they are fairly robust in terms of the efficiency of the resulting estimates, so that any reasonable design should be suitable.

Even more elaborate designs are required for studies of mechanism, depending on whether the objective is to study the effects of prenatal and postnatal exposure, to explore initiation/promotion hypotheses, to assess the impact of cessation of exposure on the carcinogenic process, or to evaluate the effects of joint exposures to different agents. All these studies require a variety of specialized treatment combinations in order to isolate the effects of interest; in addition, these assays need the use of one or more specialized control groups for comparison purposes.

Regardless of the nature of the study, there are a number of fundamental statistical principles that must be taken into account in developing a suitable experimental design. Randomization is essential in order to ensure unbiased comparisons between the treatment groups of interest and to provide a basis for valid statistical inference in terms of probability statements concerning evidence against the null hypothesis of no effect. Replication is essential in order to provide an estimate of experimental error against which to gauge the significance of any apparent related effects. Stratification may be used to decrease the magnitude of the experimental error by making treatment comparisons within homogeneous subgroups in order to increase the precision of the overall analysis.

The concept of an experimental unit is essential to the understanding of experimental design. While, in the past, the individual animal has been treated as the basic unit of information for purposes of statistical analysis, clustering due to cage or litter effects or environmental gradients existing within the laboratory suggests the possibility of one unit being comprised of two or more animals, falling within the same cluster. Since the available evidence on this effect is somewhat equivocal, it requires clarification.

More attention needs to be paid to the development of randomization schemes for the location of cages in carcinogenicity testing. While complete randomization has a desirable element of simplicity and offers protection against positional effects, it also

poses potential problems in terms of cross-contamination of the treatment groups. Cross-contamination can be avoided by assigning animals on the same treatment to the same rack or column, although consideration should then be given to adjusting for possible environmental effects at the analysis stage. The possibility of rotating cage positions may be helpful in this regard but the most suitable rotation scheme is unclear.

Care needs to be taken at the design stage to develop a good system for the recording of experimental data. Different computer systems may be used for this purpose and have the advantage of increased accuracy and speed over manual systems. Special consideration also needs to be given to the standardization of the diagnostic criteria used in reviewing histopathological data and the development of an unbiased system for this review. In addition to the pathology data itself, it is recommended that, whenever feasible, some attempt be made to determine cause of death.