First Draft May 2010

OECD GUIDANCE DOCUMENT FOR THE DESIGN AND CONDUCT OF CHRONIC TOXICITY AND CARCINOGENICITY STUDIES, SUPPORTING TG 451, 452 AND 453

Section 4: Statistical and Dose Response Analysis, Including Benchmark Dose and Linear Extrapolation, NOAELS and NOELS, LOAELS and LOELS

TABLE OF CONTENTS

Preamble

1. This document is intended to provide guidance on the statistical issues associated with the design and analysis of chronic toxicity and carcinogenicity bioassays, the analysis and interpretation of tumour data and their use in the identification of Benchmark doses, linear extrapolation and various NOAEL and LOAEL measures.

2. The basic fundamental point that this document aims to convey is:

"The statistical methods most appropriate for the analysis of results should be established before commencing the study" (OECD 2009 Para 9 Draft TG 451/453).

3. Statistical analysis of biological data is intertwined with the experimental design of studies so this draft also includes discussion of issues related to study design. Consequently there is some overlap with topics discussed in other sections but this can allow relevant linkages and cross-references to be made to other sections of the document.

Introduction

4. The central concept of this document is that the experimental design represents the strategy for answering the question of interest and that the specific statistical analyses are tactical methods used to help answer the questions. It is, therefore critical that the statistical methods most appropriate for the analysis of the data collected should be established at the time of designing the experiment and before the study starts.

5. It is important to appreciate that many of the standard methods in long-term animal experimentation have been in use for many years. There is, therefore, considerable experience of statistical properties of these designs and in the integration of the biological importance of findings with the results of statistical analyses.

6. It remains important, though, that the principles of the formal study design developed over many years should continue to underpin the subsequent pragmatic interpretation often needed. Vigilance is needed to ensure that the quality of studies does not decline because of familiarity with the methods. The long-term studies require a high standard of conduct to ensure "… an unbiased statistical analysis. These requirements are well known and have not undergone any significant changes in recent years." (see paragraph 12 - introduction).

7. Much of this chapter will concentrate on the statistical methods proposed in various documents and guidelines specifically developed for the analysis of chronic toxicity data. Particular attention will be given to a previous OECD Guidance Document (Guidance Note No 35; OECD, 2002) which made some general points about the strengths and weaknesses of statistical methods and listed some common statistical tests (reproduced and augmented here as Appendix 1) and provided references to a number of sources.

8. It is important at this point to stress that there is no single approach to the statistical analysis of data. There are different schools of thought about statistical methodologies. These can raise fundamental, almost philosophical, issues about the role of statistical analysis. (An example, for instance, is the long-running debate between 'Frequentists' and 'Bayesians'). Statistical methods also continue to develop so that new and modified approaches may continue to be proposed. As a consequence there can be alternative approaches to those suggested in various guideline documents. Such methods may, in practice, satisfy the requirements of a regulatory authority but is always recommended that such approaches are discussed in advance with the relevant regulatory authority.

Objectives

9. Initially, the primary objective of a long-term rodent carcinogenicity bioassay (LTRCB) was qualitative hazard identification (i.e. identification of chronic toxicity and the evaluation of the carcinogenic potential of a chemical administered to rodents for most of their lives).

10. The purpose of the long-term rodent cancer bioassay has, however, widened from the original 1960s-70s objective to extend to a number of other objectives. The Test Guideline 453 (para 3) identifies 7 possible objectives. These relate to hazard characterization, describing the dose-response relationship and the derivation of an estimate of a Point of Departure (PoD) such as the Benchmark Dose (BMD) or a No observable adverse effect level (NOAEL) which can then be used to establish 'safe' human exposure.

11. As a consequence the design can become a compromise with a trade-off in the ability to answer competing questions: hazard identification/characterisation on the one hand and characterisation of the dose-response on the other (see paragraph 58 – study design)

Current statements on statistical methods

12. The OECD Test Guidelines 451, 452 and 453 contain sections relating to the statistical analysis of the data. All three guidelines state:

"When applicable, numerical results should be evaluated by an appropriate and generally acceptable statistical method. The statistical methods and the data to be analysed should be selected during the design of the study (paragraph 9). Selection should make provision for survival adjustments, if needed."

13. The specific passages dealing with statistics are:

Paragraph 9 of the draft TG 451 (and also of TG 453) states:

"The statistical methods most appropriate for the analysis of results, given the experimental design and objectives, should be established before commencing the study. Issues to consider include whether the statistics should include adjustment for survival, analysis of cumulative

tumour risks relative to survival duration, analysis of the time to tumour and analysis in the event of premature termination of one or more groups. Guidance on the appropriate statistical analyses and key references to internationally accepted statistical methods are given in a Guidance Document on the design and conduct of chronic toxicity and carcinogenicity studies (7), available on the OECD public website on Test Guidelines, and also in Guidance Document No.35 on the analysis and evaluation of chronic toxicity and carcinogenicity studies (20)."

Paragraph 8 of TG 452 states:

"The statistical methods most appropriate for the analysis of results, given the experimental design and objectives, should be established before commencing the study. Issues to consider include whether the statistics should include adjustment for survival and analysis in the event of premature termination of one or more groups. Guidance on the appropriate statistical analyses and key references to internationally accepted statistical methods are given in a Guidance Document on the design and conduct of chronic toxicity and carcinogenicity studies (7), available on the OECD public website on Test Guidelines, and also in Guidance Document No.35 on the analysis and evaluation of chronic toxicity and carcinogenicity studies (15)."

14. Other organizations have made suggestions for the statistical methods to be used: For instance, the US EPA's Proposed Guidelines for Carcinogen Risk Assessment (EPA, 1996) advises that:

"Statistical analysis should be performed for each tumour type separately. The incidence of benign and malignant lesions of the same cell type, usually within a single tissue or organ, are considered separately and are then combined when scientifically defensible (McConnell et al., 1986). Trend tests and pairwise comparison tests are the recommended tests for determining whether chance, rather than a treatment-related effect, is a plausible explanation for an apparent increase in tumour incidence. A trend test such as the Cochrane-Armitage test (Snedecor & Cochrane, 1967) asks whether the results in all dose groups together increase as the dose increases. A pairwise comparison test such as the Fisher exact test (Fisher, 1932) asks whether an incidence in one dose group is increased over the control group. By convention, for both tests a statistically significant comparison is one for which $p<0.05$ that the increased incidence is due to chance. Significance in either kind of test is sufficient to reject the hypothesis that chance accounts for the result. A statistically significant response may or may not be biologically significant or vice versa. The selection of a significance level is a policy choice based on a trade-off between the risks of false positives and false negatives. A significance level of greater or less than 5% is examined to see if it confirms other scientific information. When the assessment departs from a simple 5% level, this should be highlighted in the risk characterisation. A two-tailed test or a one-tailed test may be used".

Study designs

15. Strategic issues relevant to study design include the number of doses, their spacing, the choice of the top dose, the group sizes, the length of study and the choice of control groups. In particular, the choice of the number of dose levels and the dose spacing is crucial to achieving the objectives of the study (e.g. hazard identification or dose-response/risk assessment) and is important for subsequent statistical analysis.

16. Guidelines 451, 452 and 453 have a core design consisting of three treatment groups (with different dose levels) and one or more negative control groups for each sex. Each group should be at least 50 animal of each sex for TG 451 and 453 and at least 20 animals for each group for TG 452.

17. The updated OECD TG453 recommends, for the chronic phase of the study, at least three dose groups and a control group, each group containing at least 10 males and 10 females per group.

18. There are different strategies for the allocation of resources (i.e. the animals) to the groups in the design depending upon the objective of a study. These could range from the equal allocation to a single negative control and high dose group to maximize the power to detect a difference (hazard identification) to the allocation of single units to a large number of different doses across the whole dose range. The first case would be an anova-style analysis while the latter would be a regression analysis with a test for the lack of fit of the dose-response relationship. Intermediate designs would be a number of animals allocated to at each of a number of dose groups. This design can be analysed by analysis of variance methodology where the between group comparison can be broken down into linear, quadratic, other components and a lack of fit component. This reflects the continuum that exists between the anova and regression modelling approaches.

19. There has been debate over whether a 4 group 50 animal/group design should be replaced with an 8 group 25 animals/ group design. Unpublished work examined the power of the different designs, using three different scenarios to detect a linear trend in the proportions together with other dose selection issues using the nQuery Advisor software. These calculations made no assumptions about differential survival. This analysis showed that the power of the 8 group design was between 5 and 22% lower than the 4 group design. To achieve comparable power with the 8 group design the sample sizes would need to increase by about 40%.

Control groups and length of study

20. The control groups can be either an untreated or a vehicle control group. The animals in these groups are expected to be treated in an identical fashion to those in the test groups. A discussion of the implications for the statistical analysis of the inclusion of more that one control group can be found in a later section. A control group of pair-fed animals may be included if the palatability of the substance administered in the diet (i.e. a reduction of 20% of more in food intake) is of concern.

21. All animals should be treated identically throughout the length of the study (see section on study duration). At the end of the study a full detailed gross necropsy should be carried out on all the animals from the control and test groups.

22. Planned interim kills may be part of the design. It is suggested in the combined OECD Guideline (OECD TG453) that a group of 10 male and 10 females per group (reduced from the original 20 per sex per group) should be included and that the terminal kill of these animal could act as an interim kill for the main carcinogenicity study.

23. Some consideration of the implications of including such a small sub-group for the power of the study is needed. These animals should be allocated to the treatment group before starting the study based upon some randomization process (see Bannasch et al, 1986). Indentifying these animals before the start of the study could, however, lead to them experiencing slightly different test environments such as this sub-study being maintained in a different room under slightly different conditions which might introduce biases into the statistical analysis.

Purpose of statistical analysis

24. The objective of the statistical analysis of the data generated in long-term toxicity tests is to assess whether exposure to the test chemical is associated with toxicological effects such as, for instance, an increased tumour incidence.

25. Statistical analyses can address this aim in two ways. On the one hand, the objective may be test a hypothesis that one or more treated group is different from the concurrent group; alternatively, the objective may be to estimate the size of an effect in a comparison between groups and provide some indication of the precision or confidence that can be ascribed to that estimate.

26. As mentioned previously there are different 'schools of thought' about the statistical analysis of data. Much of the work in toxicology has been carried out based upon the traditional frequentist approach particularly around the concept of hypothesis testing. While recognizing that alternative viewpoints exist and this is a controversial area, most of the emphasis in this document will be on the traditional approaches.

27. There is a need to remain aware of the distinction between statistical significance and biological importance. The increasing emphasis in the statistical community on estimation over hypothesis testing is a crucial development in the distinction between these two concepts with statistical analysis being a part of the interpretation of the biological importance, not an alternative.

28. The concept of statistical significance is an important component of the hypothesis testing approach. In a test of a null hypothesis the P value is a measure of how likely a result that has been obtained, or one more extreme, might have arisen if there were no difference between, say, the two groups. The P value is dependent upon a number of factors: endpoint, variability, sample size, experimental design and statistical method. Conventionally, certain critical values ($P < 0.05$, $P < 0.01$ and $P < 0.001$) have been considered as denoting specific levels of statistical significance although many statisticians dislike this approach. Importantly, denoting something as statistically significant does not mean it is biologically important. (The use of the term biologically or clinically important is an attempt to avoid misunderstandings as the word 'significant' has a specific and precise meaning for statisticians.)

29. A critical distinction needs to be made between statistical analysis and statistical significance. The two concepts are not synonymous. The reporting of the statistical significance of a hypothesis test may be one, often small component, of the much larger component of the design and analysis of an experiment. Many statisticians argue against the reporting of significance levels arguing

instead that the emphasis should be on emphasising the size of effects and the confidence in them. This avoids the problem of a small biologically unimportant effect being declared statistically significant and the artificiality of trying to dichotomise a result into a positive or negative finding on the basis of a P value of, for instance, either 0.051 or 0.049.

30. Similarly, declaring a result non-significant (often designated as P>0.05 or NS, again a nomenclature not favoured by statisticians) should not be interpreted as meaning the effect is not biologically important or that the null hypothesis is correct. Rather it means that there is not sufficient evidence to reject the null hypothesis.

31. A 'debate' concerns the nature of the strategy used to carry out statistical analyses. Different statistical methods will produce different results (i.e. significance levels) when they are applied to the same data sets. The specific tests applied will have different results because they are testing different hypotheses. A trend test will be testing whether there is a linear trend with a slope greater than zero; a pairwise comparison will comparing whether a treated group is significantly different from the controls. In general, testing a trend which is a more specific hypothesis has greater power than a pair-wise comparison. It is also a single test compared with the 3 pair-wise comparisons between the dose groups and the negative controls. This introduces the concept of the use of corrections for multiple comparisons. These are sometimes used to address concerns that when a large number of comparisons (e.g. between pairs of treatments) are made that there is a risk of Type 1 errors. (A Type I error is the risk of wrongly rejecting the null hypothesis in a statistical test when, in fact, it is true and thus declaring results significant when they are not).

32. Another issue is the use of parametric and non-parametric tests. Non-parametric methods 'shadow' the similar parametric tests: the Mann-Whitney, the t-test; the Kruskal-Wallis, the one-way anova; the Jonkheere-Terpstra trend test, the linear dose-response trend test. Non-parametric tests are slightly less powerful that their parametric equivalents but give potentially more accurate Type I error rates when the assumptions underlying parametric tests are violated.

33. Importantly, while non-parametric tests may be distribution free they are not assumption free so are probably as vulnerable, if not more so, to differences in the distributions between the groups. Non-parametric tests aim to ensure that correct Type I errors are derived but are less suitable for more complex designs, estimation and model fitting. Small sample sizes (e.g. 4 or 5 experimental units per group) also mean that comparisons using non-parametric tests may have low power even when there are quite large treatment effects.

34. It should always be appreciated that a statistical analysis has its limitations. Statistical analysis cannot rescue poor data resulting from a flawed design or a poorly conducted study. It cannot be stressed enough that good experimental design, again the 'strategy', is the critical part of statistical input into a study. An appropriate statistical analysis will follow directly from a correct experimental design.

Types of data

Qualitative endpoints

35. Different types of data are collected in the course of the LTRCB. Endpoints can be qualitative or quantitative. The power (see below) associated with the detection of biologically important effects can be very different between a qualitative and a quantitative endpoint.

36. Qualitative data can be binary, categorical or ordinal. Examples of binary data are where the classification can take one of two (binary) forms: an animal can be dead or alive or have a tumour or not. There can be controversial issues over whether a tumour is classified as benign or malignant or as incidental or fatal (see later).

37. There are important issues about how pathological findings are described. Sometimes these can be considered categorical (no ordering) or ordinal (where there is some ranking or ordering of the types). Issues arise as to whether findings are split into separate categories such as hepatocarcinoma, adenomas, nodules and hyperplasia or are combined such as benign and malignant tumours or the combination over tumour sites and the inclusion or exclusion of metastatic tumours. The general convention is that metastatic tumours should not be included in statistical analyses. Haseman et al. (l989) discuss the problems created by different histopathological nomenclature. There are also issues associated with splitting into separate categories. This can result in the background incidence of the tumour types being altered which affects power considerations. Splitting into a number of different separate categories of tumours also raises multiple comparison issues. When the disease is progressive, then the identification in the pathological examination of the presence of the more severe form should automatically mean that the earlier (precursor) form is or has been present. (An example is a comparison of the proportion of animals with adenomas between groups when there are already animals where there has been a progression to hepatocarcinomas.) McConnell et al (1986) and the EPA (1996) have recommended that the statistical analysis of each tumour type is carried out separately. Benign and malignant tumour of the same cell type and tissue or organ should be analysed separately and, if it is considered scientifically defensible, analysed using the combined numbers.

38. Pathology data are categorical but can be converted into semi-qualitative or ordinal data such as histopathology grading where the grading is into a number of categories, from, for instance, no through mild to more severe, with the assumption that there is increasing severity but with no assumption that the differences between classes are on a linear scale. (For instance, a change from a category I to category II may not be directly comparable to a change from Category II to Category III).

Quantitative endpoints

39. A considerable amount of quantitative data is collected during the course of a long-term bioassay. Much of this is continuous data such as body and organ weights, clinical chemistry and haematological data. In the case of the LTRCB the length of time either to the death or the identification of a tumour is a quantitative; measure (used in, some cases, as a surrogate measure of the time until a tumour arises). There can also be quasi-continuous data where although the data are discrete counts, such as numbers of various types of blood cells, these are such large numbers that they can be considered as if they are continuous data,

40. The specific statistical methods or tests used to analyse qualitative and quantitative endpoints are different. These have been represented as various times in a decision tree format with different flowcharts. An example developed by the OECD (OECD GD 35) previously will be discussed below. It is important to appreciate that while the statistical methodologies and algorithms differ, the underlying statistical concepts associated with the interpretation of the tests are basically the same. However, the power associated with the different endpoints within the context of the same basic experimental design can be very different.

Sample size and power considerations

41. The power of a study is the probability of detecting a true effect of a specific size or larger using a particular statistical test at a specific probability level. The power is $(1-\beta)$ where β is the Type II error associated with a hypothesis test. (The Type II error is the probability of wrongly accepting the null hypothesis as true when it is actually false.)

42. The power of a study for a qualitative trait depends upon 5 factors: the sample size (n), the significance level (α), whether the test is one- or two-sided, the size of effect of interest ($d = q-p$) and the control incidence (p). In the case of quantitative data; the proportions are replaced by the size of effect of interest and a measure of the inter-individual variability such as the standard deviation. Numerous software packages and programs are available for carrying out these calculations.

43. The OECD Guidelines indicate the appropriate sample sizes for each group. In the LTRCB this is usually at least 50 animals of each sex at each dose level. This group size reflects a trade-off between the statistical power of the design and economic practicalities of the design. In practice, the LTRCB has low power in the sense that effects that in other experimental contexts would be considered biologically important effects cannot easily be distinguished from the Type 1 errors which reflect chance fluctuations between groups.

44. It is recognized that the power of the study can only be increased modestly by increasing sample sizes. Similarly, allocating animals differently between the test groups can increase the statistical power of detecting, for instance, low dose effects. Portier & Hoel (1983, 1984) suggested, for instance, that 20 animals should be moved from the low to the medium group to improve the power of the study (a 50:30:70:50 split instead of 4 groups of 50). However, this proposal does not seem to have been adopted to any extent.

45. A common feature is that the power associated with qualitative data is less than that associated with quantitative data. See, for instance, FDA (2001) para 62 and 63 which relate to the low power of comparisons of the low dose group in the LTRCB. The design, for instance, is only able consistently to detect increases of about 10% over the negative control incidence with the power being reduced (further) if there is a high control incidence.

Statistical Flowcharts

46. Flowcharts for conducting statistical analyses based decision rules to make choices have been developed and used extensively in the analysis of statistical data in the biological sciences. There are obvious practical advantages in having a set of standard methodologies with the choice of particular methods made at key decision points based upon data. A number of examples can be found in textbooks. Gad (2006) for example has developed some for the analysis of toxicological data. The OECD produced a flowchart (OECD GD 35) in a previous document which is reproduced here (Figure 1).

47. The choice of the statistical method to use is based upon whether the data are qualitative or quantitative and based upon the assumptions required by the test being met. The choice of route through the flow chart is based upon the results of answers to queries higher in the chart. For example, in the event of a test for non-normality of the data a non-parametric test may be chosen in preference to a parametric one. This methodology speeds the analysis and reduces the amount of valuable time a statistician spends on an analysis.

48. Gad (2000), for instance, suggested that statistical analysis of endpoints such as body and organ weight data are "universally best analysed by ANOVA followed, if called for, by a post hoc test." He suggests Bartlett's test is performed first to ensure homogeneity of variances. With smaller sample sizes he suggests that Kruskal-Wallis test may be more appropriate. In the case of clinical chemistry he points to the limitations of univariate analyses when there is a battery of parameters.

49. Critics point out that, although there are efficiency gains, there is a 'deskilling' of the task, an over-emphasis on significance testing for decision making and vulnerability to artefactual results. There is also the philosophical problem of the use of hypothesis testing methodology to choose another hypothesis test where the use of a multiple testing procedure can complicate quantifying the true probability values associated with various comparisons.

50. Some concern has been expressed over whether tests for normality or, for heterogeneity of variances are over-sensitive and, as a consequence, unnecessarily rule out the use of robust statistical methods such as the analysis of variance and, thus, potentially reduce the power of the design.

51. In the case of tests for normality, the null hypothesis is that the data are normally distributed. In practice, this can mean that a small sample which may be quite non-normal may fail to 'trigger' a significant result while a large sample with a slight deviation from normality will be called as non-normal and lead to a switch in the subsequent statistical test. Such test behaviour is counter-intuitive to what in practice is required in the selection of statistical tests.

52. Other decision rules based upon trying to find an optimal transformation can lead to a very heterogeneous set of analyses based upon decision rules which lead on to another different transformation. The FDA Redbook (FDA 2000), for instance, indicated that unnecessary transformation should be avoided as should the use of a mixture of parametric and non-parametric methods on the same endpoint.

53. From a philosophical point of view mechanical statistical analysis differs from the modern way of carrying out analyses using statistical models. In this there is an iterative process of fitting and testing models and checking assumptions. Visual representation of data is also an important aspect of the analysis, relying on inspection of the data for outliers, trends, goodness of fit and checks of assumptions. Care should, therefore, be taken in carrying out statistical analyses using flowcharts. The standard provisos in the interpretation of statistical analyses should always be kept in mind.

Description of the OECD flowchart

54. In this section the OECD (OECD Document 35) flowchart (Fig 1) will be worked through to illustrate the points that arise in its use. This will then be followed with some of the issues that arise if statistical analyses are totally dependent upon such an algorithmic approach.

55. The flowchart developed previously for the OECD is discussed below and is similar to an approach used by the US NTP. The objective, here, is to provide a brief overview of the methods used. The individual tests are described briefly in the glossary.

56. For simplicity the description of the flowchart will be by working from left to right. This is for convenience and there is no priority implicit in this ordering. When an option is reached, the left choice initially will be described. Once the end of the 'tree' has been reached the description will move back to the previous node/decision point and continue down that. The same 'zig zag' procedure will be carried until the tests at the far right of the flowchart are reached.

57. In general, (but not precisely) the boxes at the top are concerned with aspects of the nature of the data, moving down there are tests of the assumptions underlying the methods followed by 'omnibus' tests (which test for overall differences between groups irrespective of the specific design), these are followed by tests of whether there is a linear trend across the treatment groups and/or between treatment groups and the negative control group. In some cases there is a circular/interactive path when, for instance, following rejection of the assumption of normality the data are transformed to logarithms (natural or to the base 10) and the test for normality made again. The number in brackets relates to the numbers representing specific points on the flowchart. The assumption is that a decision is made to choose one or other route if the P value associated with the test statistic is <0.05. (Note that the use of hypothesis tests, the specific choice of P values, the choice of parametric or non-parametric and the use of multiple testing procedures are controversial issues and the particular routes outlined in this particular flowchart would not necessarily be agreed on by all statisticians.) Note that some of the statistical methods, particularly those related to the analysis of tumour data are discussed in more detail elsewhere in the document.

58. The top level of the flowchart (1) relates to a check of the data for overall quality and the identification of the type of data, whether quantitative continuous or qualitative or discrete. As part of the data checking an optional test for outliers such as the Dixon & Massey test is suggested. Moving to the methods included for continuous data a test for the assumption of normality (either the Kolmogorov-Smirnov test or the Shapiro-Wilk) test is identified (2). If the test is significant indicating that the data are not normally distributed the option is that the data are

logarithmically transformed (3) and the test for normality carried out again as well as perhaps testing for outliers using the Extreme Studentized Deviate (ESD) statistic (4). If neither transformation or identifying outliers results in normality the suggestion is to assume the distribution of the data is not normal and move (5) to the use of non-parametric methods (in the centre of the flowchart). If the data are assumed to be normal then a further test for homogeneity of variances (6) is suggested. In the case of a two group comparison this is an F-test(7). If the F test is not significant then the two groups are analysed by a standard Student's t-test (8); if the F test is significant then the comparison is by the modified t-test using Satterthwaite's method for unequal variances (9). Returning to point (6): either Levene's test or Bartlett's test are used to test for homogeneity of variances (10) when there are three or more groups. If the variances are considered heterogeneous the flowchart directs the analysis to non-parametric methods (11). If the variances are considered homogeneous then the comparison of all groups is suggested to be by a one-way analysis of variance followed by the multiple comparison procedure Duncan's multiple range test or Tukey's Honest Significant Difference tests (12). In the case of pair-wise comparisons between the control and the dosed groups, the flowchart suggests Dunnett's test (13).

59. Returning to the qualitative data (1) a distinction is made between data which are ranks or discrete counts (14) and those which are pathology findings (15) and those related to death or survival (16). It is suggested that ranked data, together with quantitative data which were either determined to be non-normally distributed (5) or have heterogeneous variances (11) are analysed by non-parametric methods (17). The suggested methods for comparisons are the Kruskal-Wallis test for comparisons between the groups and Jonckheere's test for a trend in the data. Methods identified for comparisons between the control and test groups are the Kolmogorov-Smirnoff test and the Wilcoxon Rank Sum Test (which is equivalent to the Mann-Whitney U-test). If these tests are significant then further testing using distribution free multiple comparison tests can be carried out using tests such as Dunn's or Shirley's tests (18). Returning to pathology findings (15), if an interim kill is carried out comparisons between the proportions of animals with pathological findings in a treated group can be compared with the proportion in the control group using Fisher's exact test (19). At the end of the study a choice is made (20) between tests which take into account information on how long the animals lived without a tumour (21) and those that do not such as the Cochran-Armitage test for a trend in proportions (22). The survival adjusted tests (21) are the Peto analysis (which requires information about whether the tumour is 'incidental' or 'fatal' and the poly k-test which does not need this information. (These methods are described in more detail elsewhere in the document.)

60. Returning to qualitative data (1) and to data on survival/death (16), the flowchart identifies survival analysis approaches such as the Kaplan-Meier non-parametric methods (23) followed by comparison of the graphs of the survival curves (24) followed by analysis using the log rank test (25).

Intercurrent mortality

61. Inter-current mortality is death that arises during the course of a study from anything other than a tumour. Chronic studies can last up to two years. Animals can and do die in the course of the studies for a variety of reasons both related and unrelated to the treatment they have received.

Inter-current mortality is different from the planned interim sacrifices or kills specifically included in some designs. Inter-current mortality complicates the statistical analysis of comparisons between test groups. For instance, older animals are more likely to develop a tumour than younger ones. The risk of getting a tumour and of dying because of a tumour increases with age. Consequently, the probability that an animal that dies unexpectedly during the course of a study also has a tumour will depend upon the animal's age at death. The test chemical may also affect the survival of different groups by causing either more deaths (through non-tumour related toxicity) or fewer deaths (by, for instance, reducing food intake and making the animals less susceptible to obesity- related morbidity and mortality). Peto et al (1980) pointed out that comparisons based upon crude tumour rates in the presence of differential mortality can produce serious errors in the analysis.

62. A simple statistical analysis which does not account for inter-current mortality (described below) can underestimate the carcinogenic effects if the treatment decreases survival. Conversely, if the treatment increases survival then the tests may overestimate the carcinogenic effects. Failure to take intercurrent mortality into effect can, therefore, produce serious biases in the interpretation of results. Peto et al argued that to avoid this occurring, adjustments are needed for differences in survival between the groups and this correction should be routinely used. They state "the effects of differences in longevity on numbers of tumor-bearing animals can be very substantial, and so, whether or not they (the effects) appear to be, they should routinely be corrected when presenting experimental results" Peto et al (1980).

63. Inter-current mortality is a serious problem for the interpretation of chronic studies because a high proportion of animals in the groups need to survive for a successful analysis. Traditionally, sample sizes per treatment group have been between 50 and 60 of each sex. The FDA gave 'a rule of thumb' that 50% survival at about weeks 80 to 90 was considered adequate (FDA 2001) with the key point that about 20-30 animals need to be alive (in each group) at these times (Lin & Ali, 1994). The actual percentage is considered less important and depends upon the size of the treatment groups/sex.

64. The FDA (2001 p4) has suggested there are occasions when inter-current mortality may not be a major problem. However, it is important that any organization carrying out such an analysis should contact the relevant regulatory authority for advice in the event of problems with survival so as to try to ensure that the optimum amount of information is obtained from any study which has to be stopped early.

Context of observation (COO)

65. Information on what is called the 'context of observation' (COO) (or sometimes 'cause of death' (COD) (Peto, 1974)) may be important for subsequent statistical analysis of a study. Peto et al (1980) argue that a distinction needs to be made between whether a tumour is called 'fatal' or 'incidental' for a correct statistical analysis to be carried out. A 'fatal' tumour is one that is considered to have caused the death of the animal; an 'incidental' tumour is found when the animal died from an unconnected reason or was found at the terminal kill at the end of the experiment. Distinguishing between these two 'contexts' can be both difficult and controversial.

However, this distinction is a critical feature of the 'Peto' analysis described in the 1980 IARC monograph (Peto et al, 1980).

66. Complications arise because this dichotomy is not always simple. Information on the context of observation may not be available because it is not provided by the pathologist or is unreliable because it is based upon assumptions about lethality which may be contentious.

67. A pathologist may be unwilling to make definitive statements about the COO of a tumour. Some pathologists have argued that it is difficult retrospectively to diagnose accurately if a tumour is the true cause of death of an animal. There may not be a single factor which solely determines the death; there may be multiple causes of death with the presence of a tumour being just one of them. Alternatively, more than one tumour may contribute to the cause of death. There may not be a single factor which solely determines the death. Haseman quoted by the FDA has pointed out, though, that 'incidental' and 'fatal' tumours are on a continuum and that it is an oversimplification to say that a tumour is either 100% 'fatal' or 100% 'incidental'.

68. Peto et al's (1980) approach to this problem was to suggest four categories of tumours - (1) probably fatal, (2) possibly fatal, (3) possibly incidental and (4) probably incidental - with the suggestion of combining the categories in different ways using different cut-offs to produce a binary endpoint: e.g. combining 1+2 versus 3+4; 1+2+3 versus 4 and 1 versus 2+3+4 and analyzing each combination as a form of sensitivity analysis. This has proved controversial because it is argued that it is a device to subsequently reduce the categories back to two: 'fatal' and 'incidental'.

69. Lee & Fry (Lee et al, 2002) point out that the 'fatal' definition is often misunderstood by pathologists. The key issue is that a 'fatal' tumour is wholly responsible for the death of the animal not that the tumour had the potential to kill the animal at some time in the future. Lee and Haseman (Lee et al 2002) also disagreed on how easy it was for the pathologist to make a distinction between the COOs. Lee quoted Peto's example of the high proportion of definitive diagnoses in the BIBRA nitrosamine study where 94% of over 4500 tumours could be classified as either "definitely incidental" or "definitely fatal," even though the pathologists had initially expressed reservations about whether such classifications could be made reliably. Haseman (Lee et al 2002) argued that there was instances where it was difficult to make the distinction and when made it was often incorrect. There was also a tendency for the over-designation of fatal tumours (Abu et al, 2001; Kodell et al, 1982). Haseman argued that pathologists should be allowed the freedom to make their own judgements. Soper and Kodell (Lee et al, 2002) argue for a more objective classification based upon the large historical data base available.

Time to tumour onset

70. The time until an event occurs such as death can be analyzed by the statistical technique of survival analysis. There is much widely available statistical software that can be used to carry out the standard analyses. Such methods are used to analyze the survival data from the LTRCB (see below).

71. Analysis of time to tumour data, however, is less straightforward. Ideally, the time when a tumour first occurred is needed so that these times can be compared between treatment groups. There are some tumours which can be observed during routine observation of an animal such as some skin and some palpable mammary tumours. In practice, the time, when a tumour is identified is an arbitrary but objective endpoint such as recognized when the tumour reaches a certain size. These tumours are called 'mortality independent' and can be analysed by standard life-table survival analysis methods.

72. However, these are the exception. Most tumours are internal (and, therefore, hidden or 'occult') and will usually be detected only during a post mortem examination of the animal. The specific time (the age of the animal or the tumour onset time) when the tumour initially arose is consequently unknown. It cannot simply be replaced by the time when the tumour was first identified as an 'occult' tumour (in other words the time of identification of an incidental tumour is not a surrogate for the time of onset). This lack of information complicates analysis. Statistical methods which take into account the time until an event occurs (such as survival analysis methods) need to make a number of assumptions for analyses of such 'occult' tumours.

73. Dinse (1994), for instance stated: "without direct observations of the tumor onset times, the desired survival adjustment usually is accomplished by making assumptions concerning tumor lethality, cause of death, multiple sacrifices or parametric models" (Dinse, 1994). Approaches to get around this problem are to model and impute the onset times, to include extra data (context of observation) or to make more assumptions (poly-k test)" (see later).

74. One approach to the problem is to increase the sample size and have planned interim kills/'sacrifices'. In practice, this is rarely done so there are no universal guidelines for analyses of such studies despite an appreciable number of papers relating to the approach.

Standard (simple) statistical analysis of qualitative data

75. Tests for pair-wise comparisons and trends are recommended for the statistical analysis of tumour data to test for treatment-related effects (EPA, 1996; FDA 2001). There are a set of standard statistical methods for comparing proportions between one or more groups. In these tests the basic information is the number of animals at risk as the denominator and the number of animals with the tumour (or pathology) as the numerator of interest. These tests make no assumptions or corrections to take into account 'COO' information or the time to tumour or death. They are described in many standard statistical textbooks and software for analysis is widely available.

76. The three main tests are:

1) A pair-wise test between the negative control and the treated group using the Fisher exact test (Fisher, 1950).

2) A chi-square test for heterogeneity of proportions between groups. This is an 'omnibus' test of differences between a series of groups (with no ordering to the groups).

3) A test for a linear trend such as the Cochran-Armitage trend test. (Snedecor & Cochran, 1980).

77. Pair wise comparisons are carried out using Fisher's exact test or its Chi-square approximation. Fisher's exact test is now preferred because of the availability of software for carrying it out.

78. The Cochran-Armitage trend test aims to detect a linear trend. The null hypothesis is that all animals are at equal risk of developing a tumour during the study. Problems arise if there are differences in mortality between the groups. The test is sensitive to increases in treatment related lethality and this leads to an incorrect level of the Type 1 error (the risk of falsely rejecting the null hypothesis).

79. The selection of the dose metric (the values representing the actual doses used) for use in the analysis is important. The doses could be the 'applied', logarithmic, some equally spaced rank or some measure of the effective dose at the target organ (based upon pharmacokinetics). A practical issues arises using a logarithmic scale with the need to choose a value to substitute for the zero dose level.

80. The FDA (2001) states that a trend test is the primary test. It is more powerful than the pair-wise test. A complication is that a trend test may fail to detect curvi-linear responses such as might arise from non-linear effects such as complications from saturation. In such case the pair-wise tests may give more appropriate results.

One- or two-sided tests

81. The choice of whether a one- or two-side test should be used should be made at the design rather than the analysis stage. A two-sided statistical test tests for a difference from the negative control (in a pair-wise comparison) in either direction. A one-sided comparison tests for a difference in only one pre-specified direction, but as a consequence has more power. In a LTRCB the expectation is often that the change will be an increase in tumours in the treated group so a one-sided test may be considered more appropriate, although this can be controversial. If the treatment could also be protective (i.e. reduce tumour incidence or delay it) then a two-sided comparison may be appropriate. Regulatory authorities may have specific opinions. For instance, EPA (1996) note that either "a two-tailed test or a one-tailed test may be used".

Tests of difference in survival

82. As discussed earlier, differences in survival can affect the conclusions drawn from the simple analyses. A number of methods can be used to test for differences in survival and for significant dose-response relationships or trends in studies. These include: the Cox test (Cox 1972; Thomas, Breslow, and Gart, 1977; Gart et al. 1986); the generalized Wilcoxon or Kruskal-Wallis test (Breslow 1970; Gehan 1965; Thomas, Breslow, and Gart 1977); and the Tarone trend tests (Cox 1959; Peto et al. 1980; Tarone 1975).

83. The number of animals surviving until the scheduled terminal kill can be compared by what is termed the product-limit or Kaplan-Meier method (Kaplan and Meier, 1958) and plotted graphically. This analysis compares length of survival in the groups (and does not involve any pathology findings.) Animals that are found dead other than from natural causes are usually treated as censored and excluded from the survival analysis. There is a pair-wise comparison method (Cox, 1972; Tarone & Ware, 1977) and a test for linear trend, Tarone's life-table test (Tarone, 1975).

84. The Kaplan-Meier analysis involves calculating the ratios of the surviving animals divided by the numbers of animals at risk of dying. Every time an animal dies the ratio is recalculated. These ratios can be plotted to show a curve which displays the probability of survival. When there are different groups a curve can be generated for each group. Formal statistical tests such as the log rank test can be used to test difference between groups. (Curves which are close together or cross usually indicate that any difference is not statistically significant.)

85. The Cox proportional hazard method is a regression method similar to logistic regression but also takes into account the time until the binary event of interest (death) occurs. The modelling allows the effect of an exposure/treatment to be investigated after adjusting for confounding effects.

Survival adjusted analyses

86. For the reasons discussed above survival adjusted methods are strongly advocated for comparisons of tumour incidences between groups.

87. The FDA (2001) describes some exact methods for use instead of the normal approximations used in the standard methods. The standard methods may underestimate P values when the numbers of tumours in the groups are small and exact permutation tests which are extensions of the Fisher exact test are suggested for use. The FDA provides detailed examples of use of the methods on hypothetical data.

88. Three different types of statistical procedures have been developed depending upon the type of tumour.

1) The prevalence method for non-lethal (incidental) tumours (Hoel & Walburg, 1972; Peto et al, 1980)
2) The death rate method for lethal (fatal) tumours (Tarone, 1975; Peto et al 1980)
3) The onset–rate for mortality independent (observable) tumours (Peto et al 1980)

The prevalence method

89. The prevalence method, also called the incidental method, is, effectively the Hoel-Walburg procedure for nonlethal tumours which makes no assumption about the context of observation of the tumour (Hoel & Walburg, 1972).

90. The procedure involves carrying out a life-table analysis, a method for analysing censored observations that have been grouped into intervals, for tumours which were found incidental (incidental context) to the death of the animal. The experimental period is split into a set of intervals (including any interim or terminal sacrifices).

91. It has been argued that the choice of partitions is not critical. Peto et al (1980) say the partitions should not be so short that the prevalence of incidental tumours is unstable nor so large that the prevalence could differ markedly from one half of an interval to the other and suggest a partition based upon 'ad hoc' runs. However, there have been concerns about constructions of ad hoc intervals in the Peto analysis and attempts made to standardize them. The FDA (2001) suggested partitions of 0 – 50, 51 – 80, 81 – 104 weeks, plus the terminal and any interim kills. The NTP has suggested 0 – 52, 53 – 78, 79 – 92, 93 – 104 plus the terminal and any interim kills.

92. In the analysis the denominator is the number of animals dying within the specific partition and the numerator is the number of animals dying with an incidental tumour. The analyses for each individual partition are combined using the Mantel-Haenszel method. This compares two groups for proportions with an adjustment for control variables. A series of k 2 x 2 contingency tables are produced with k being the number of strata of the different control variables such as age or sex. The stratification increases the power of the design to detect an association. The test statistic is a chi-square. Full details of the equations used to conduct it are found in the FDA guidelines (FDA, 2001).

93. The methodology uses normal approximations for the tests but the approximation may be unreliable if the number of tumours in a group is small and the FDA suggests the use of a permutation test involving the hypergeometric distribution. The FDA (2001) provides an example.

The death rate method (for comparing rapidly fatal tumours)

94. This test is also referred to as the log rank test assuming that all the tumours cause death (Peto, 1974), It is used when tumours are observed in a fatal context. In the analysis the stratification is based upon a partition into intervals (often a week) where one or more animals died. In each stratum the number of animals entering the partition or strata who are tumour free is the denominator and the number of animals dying during the partition with a fatal tumour is the numerator. The analyses for each individual partition are combined using the Mantel-Haenszel methods. Full details of the method and equations are found in the FDA (2001).

Mortality independent analysis (onset rate method)

95. In those tissues such as skin, and, possibly mammary tumours where the tumour can be observed in the live animal the tumours are described as mortality independent. The onset rate method (the same basic statistical method as used for the death rate or fatal analysis) is used for these mortality independent tumours. The endpoint in this case is the occurrence of a tumour based upon it reaching some predefined size rather than the death of the animal. Once a tumour has been identified using this criterion the animal makes no further contribution to the analysis. It is no longer 'at risk' of developing a tumour because it now has a tumour even though it may live on for some time. The calculations used to produce the statistics for the onset rate methods are comparable to those produced by the death rate method.

Peto / IARC analysis

96. The Peto analysis, as described in an IARC monograph (Peto et al, 1980), is a combination of the prevalence and death rate based upon the context of observation of a tumour as either being called nonlethal ('incidental') or the cause of the animal's death ('fatal'). It is a joint test of age-adjusted tumor lethality and age-adjusted tumor prevalence. The assumptions underlying the method are that the control and treated animals are equally likely to be killed at any particular stage in tumour's development, the animals dying of other causes are representative of all animals surviving in that interval and that the pathologist is 'content' to make the 'fatal'/'incidental' classification.

97. The analysis combines the two separate approaches. Animals whose tumours are termed 'fatal' are analysed by Peto's death rate method while those called 'incidental' are analysed using the prevalence method. The two analyses are then combined using the Mantel-Haenszel method to provide a test for trend. A problem is that the analysis will be biased if the assumption on the nature of the tumour being either called 'fatal' or 'incidental' is inaccurate (Dinse, 1994). Although the Peto test is described as robust, there is some debate as to how big a problem misclassification is.

98. Peto et al (1980) provide an illustration of the implications of wrongly defining the context of observation. For example, in the BIBRA study of pituitary tumours in animals treated with N – nitrosodimethylamine (NDMA) different conclusions are drawn depending upon the context of observation. If all the tumours were considered fatal then NDMA was wrongly considered carcinogenic, if they were considered incidental then it was wrongly considered protective. Taking the context of observation into account produced the correct result that there was no carcinogenic effect in the pituitary.

99. A comparison of 4 different procedures showed that the Peto procedure was more robust in the analysis of a set of simulated tumour data sets than the chi-square, Hoel-Walburg or log-rank procedures (Graves, 1994). It gave the most accurate nominal error rates and statistical power.

Logistic regression

100. Logistic regression is a regression method used where the outcome is binary such as whether an animal has a tumour or not. The effect of an exposure on this binary outcome can be adjusted for confounding factors in a study. Dinse and Haseman (1986) suggested it as an approach for the analysis of incidental-tumour data from the LTRCB.

Poly-k test

101. More recent approaches have been the development of the poly-k tests (Bailer & Portier, 1988*;* Dinse, 1994). The poly-k test does not need arbitrary partitions of time periods or COO information. The test is based upon the assumption that the time to tumour onset can be modelled based upon the tumour onset times raised to the power k. Initially, the test was proposed without identifying how to derive k but now it is suggested that k should be 3 because of observations that tumours can be modelled by a polynomial of order 3 from an analysis of NTP historical control data for F344 rat and B6C3F1 mice (Portier et al, 1986.) The poly-3 test is then a special case of the poly-k test. The power of 6 (or k=6) can be used when the tumour onset times are close to a

polynomial of order 6. The value of k need not be critical as poly-k tests are reported to give valid results if the true value lies between 1 and 5 (Bailer and Portier, 1988).

102. The tests are, in effect, modified Cochran-Armitage tests which adjust for differences in mortality in the treated groups by a modification of the number of animals in the denominator to reflect the less than whole animal contributions because of reduced survival. The approach gives a value (w) from 0 to 1 for each animal based upon a weight which relates to the time of death or the time of the final sacrifice so that w relates to the fraction of the length of time the animal survived on the study over the total length of the study to the power k. The value w is <1 if the animal died early without developing a tumour and w = 1 if the animal died with a tumour or survived until the study was completed. The number of animals at risk is replaced by a new estimate in the Cochran Armitage test. The method tests for a dose-related trend in the mortality-adjusted lifetime tumour incidence rate.

103. The Bieler-Williams variance (1993) is used in the test which is sometimes referred to the Bieler-Williams method/test where the poly-3 test is modified using the delta method and weighted least squares techniques in order to adjust the variance estimation of the test statistic.

Comparison between Peto and Poly-k methods

104. Rahman & Lin (2008) compared the false positive rates of the Peto and poly-k tests using a simulation study. Kodell in Lee et al (2002) compared the properties of the Peto and poly-k tests and concluded that both are valid for adjusting for differential mortality. The problem remains that the comparison is ideally on the tumour onset data but because the tumours are occult, the tumour onset cannot be actually observed but there is also a need to adjust for any differences in inter-current mortality in the analysis.

105. Debate over the methods for the analysis of COO data using the Peto analysis generated controversy. The Society of Toxicologic Pathology (STP) set up the STP Peto Analysis Working Group and produced draft recommendations for the classification of rodent tumours for the Peto analysis. These were critical of the Peto analysis and felt the poly-3 methodology was more appropriate (STP 2001).

106. They concluded:

1) Pathologists cannot determine the time of onset accurately from post mortems.

2) If the Peto analysis uses death as a surrogate for time of onset then the method seems inappropriate

3) Better to use other methods which do not require COO

107. Lee & Fry (Lee et al, 2002) responded to the STP Peto Analysis Working Group recommendations and their comments together with those of other statisticians were published as in a collection of comments in Toxicologic Pathology (Lee et al, 2002).

108. Kodell (Lee et al, 2002) concluded that both the Peto and poly-3 tests are valid for adjusting for differential mortality. Both are fairly robust to deviations from their assumptions although both could be improved by modifications and generalizations.

109. Based upon these comments particularly relating to issues around the use of the Peto method and the onset of fatal tumours the STP withdrew its criticism of the Peto approach while recognising that the poly-3 test was appropriate in certain circumstances (STP, 2002).

110. These led to new STP recommendation that:

- The Peto test should be performed whenever study pathologist and peer review pathologist can consistently classify neoplasms as Fatal or Incidental

- If Fatal and Incidental classifications are not applied, the Poly-3 or another alternative to the Peto test should be employed

Approaches used by various regulatory authorities

US National Toxicology Program (NTP)

111. The NTP used routinely to carry out two trend tests. One assumed that all tumours in dead or moribund animals were 'fatal'; the other assumed all the tumours were non-fatal ('incidental'). The current approach is that life-table tests or prevalence tests are no longer used. Instead, the poly-3 test with Bieler-Williams variance with a trend test and pair-wise tests with controls is used. Sometimes this test is used with k=1.5 and/or k=6.

US FDA Draft Guidance (2001)

112. The Peto prevalence method should be used for tests of trend in the analysis of incidental tumours and Peto death-rate method should be used for tests of trend in the incidence of fatal tumours when COO information is available. The poly-3 test, with Bieler-Williams variance, should be used when COO is not available or is not accurate enough to perform a meaningful statistical analysis. Permutation tests are used when there are small numbers of tumours.

113. Specifically, 'statistical tests for positive trends in tumor rates are usually of greatest interest but …, in some situations, pairwise comparisons are considered to be more indicative of drug effects that trend tests'.

Assumptions for statistical analysis

114. In the context of the theme of this document, it should be reiterated that the design of the experiment is fundamental to the choice of statistical methods. The FDA (2001) notes:

"Of particular interest to statisticians are the methods used to allocate animals to treatment groups and caging rotation, the determination of sample size, and the duration of the study."

115. An important consideration is the need to balance the considerable experience build up in the use of these methods over many years with the appreciable empirical knowledge that exists on the implication of violating the assumptions associated with the statistical methods.

Randomization

116. An assumption underlying the design and subsequent statistical analysis is that the animals have been assigned at random to the treatment groups. Each animal entering the study should have the same chance as any other of being allocated to one of the experimental groups (including the control groups). Randomization can be carried out by a number of different methods. The primary objectives are to prevent bias and to ensure that uncontrolled covariates do not affect the results of the analysis as well as being one of the assumptions underlying standard statistical methods such as the Anova.

117. Stratified randomization with groupings based upon body weights may be used to reduce bias and ensure compatibility of the various treatment groups with respect to uncontrolled variables. Ideally the animals from all groups should be placed into the study at the same time. If this is not practical, strata or blocks that can be included as factors in the subsequent statistical analyses can be created by starting subgroups from the control and each of the treatment groups over several days.

118. An assumption in the statistical analysis is that randomization occurs at all points in a study so that biases are not introduced. If, for instance, the animals are initially randomized into groups but subsequent procedures are carried out in a pre-defined systematic order there is a risk that biases may be introduced.

Independent experimental units

119. The animal is often both the experimental and observational unit. If a cage is assigned to a treatment it becomes the experimental unit. However, the common statistical methods used in the analyses make the assumption that the experimental units are independent. In some cases this is clearly not so. In practice, although the assumption of independence of experimental units is an important one (more important than normality and homogeneity of variances; van Belle, 2009) it is not always taken fully into account in toxicological studies.

120. Individual housing is preferable to meet statistical assumptions but may have implication for the welfare and the representativeness of the animal. Again the trade off is between a theoretical optimum design, the avoidance of cage/confounding effects and practical husbandry issues, the 'pathology' of single housing and the possibly increased variability of isolated animals. Litter and caging effects should where possible be taken into account in the statistical analysis. If this is not possible this should be noted with an explanation of why this was not possible and the potential implications for the study.

121. Lack of independence may also arise from contamination of doses within and between cages. In the case of airborne contamination the random assignment of cages will mean that some of the lower dose and control groups will be exposed to some or higher concentrations than is implicit in the experimental design. There is clearly a trade off between the potential limitations resulting from contaminations arising from randomization in a room (but which also provides

some protection from this uncontrolled variable) and the potential confounding effect of separate handling and housing of dosed groups necessary to prevent cross treatment group contamination.

Equal information per unit

122. There are implications for the assumptions underlying the statistical analysis in the non-random reading of histopathology slides. In some studies more effort has been put into reading slides from the control and top dose groups with less emphasis on the examinations at the intermediate doses. This was particularly relevant when qualitative hazard identification was the objective. Although such studies have been accepted by some regulatory authorities this can create problems in the statistical analysis of dose-response trends and cannot be recommended if dose-response characterization is an objective of a study. Similar considerations arise if, because of uncertainty in a diagnosis, more slides are read for some individuals than others.

Blind or unblind reading of slides

123. Blind reading of slides is a controversial topic. Many statisticians expect blinding to treatment group to be included, as in clinical trials as a protection against biases. Some express concern that the current practice of open or unblended assessment of histopathology in the LTRCB can introduce bias. Temple et al (1988) have argued that blinding should be used to guard against biases the might arise (inadvertently) from the pathologist knowing the treatment group of the animal from which the tissue was derived. Some veterinary pathologists argue, however, that to use blinding can lead to the loss of information which is important for the interpretation, such as, for example, being able to link observations which are in different tissues. Others have argued that blinding and randomization can increase the risk of mistakes being made when using coded samples (Iatropolous, 1984 1988; Prasse et al. 1986). Haseman, in the STP discussions (Lee et al, 2002), has argued from a statistical point of view that blinding is not necessary given the long experience of using the assay. (This is another example of experience gathered from using a mature test over many years being used to provide a perspective on deviations from the assumption underlying the perfect experimental design). Blinding is, however, often used when re-reading of slides is carried out for disputed cases or where the results are close.

Confounding variables

124. One of the purposes of randomized and blinded studies is to minimize the effects of uncontrolled covariates and to prevent the introduction of biases by preventing confounding factors from distorting the results. A confounding variable is one which is so closely related to another factor in the design that the individual contributions to an effect cannot be separated. Haseman (1984) discussed a number of confounding effects such as cage location and litter mates although the latter should be controlled by randomization. Butterworth et al (2004) discuss another form of confounding where rather than the target chemical a contaminant may be responsible for the carcinogenicity detected.

125. In the case of the LTRCB, early differences in body weight between control and treated animals which persist through the study create a potential confounding effect between body weight, life span and tumour incidence. Ad libitum overfeeding is the most important uncontrolled variable which affects the results of a LTRCB. Keenan et al (1996) reported, for instance, that there is a highly significant correlation between food consumption, body weight /

obesity and shortened life-span in rodents. Kodell (2000) noted that the reduced survival of Sprague Dawley rat questioned the continued use of this stock in the LTRCB.

126. If the tumour profile of lean and obese rodents is different then there is a possibility that apparent treatment related differences in tumour incidence may, in fact, be wholly or partially caused by the body weight differences. Confounding may then make it difficult to identify if an effect was a direct result of the treatment or an indirect effect through the treatment affecting food intake and consequently body weight. Ibrahim et al, (1998) noted that there is usually not enough information to provide a regression based adjustment for differences in body weight.

Interpretation of statistical analyses

127. Interpreting the results of a LTRCB study is complex. A critical issue is the practical problem of the low power of the design when the tumour incidence is rare together with the multiple comparisons issues arising from the investigation of 20 or more tissues from both sexes of two species. So there is a risk of both Type I (false positives) and Type II (false negative) errors. It also remains to integrate the results of the full battery of statistical tests, significant or otherwise, with the importance of a series of biological issues in the assessment of the result. Factors which add importance to findings include: uncommon tumours, multiple sites, positive findings using more than one route of administration, effects in multiple species/ strains/ sexes, effects at the same site in both sexes and/or species, progression, increased preneoplastic lesions, reduced latency, metastases, unusually large responses, dose-related and a high proportion of malignant tumours together with a consideration of the observed tumour incidence in the study in comparison with the historical control rates.

128. Various approaches are taken to control the false positive rate. The FDA (2001) recognizes trend tests as more powerful and considers them the primary tests in evaluation. Pair-wise comparisons can be more relevant in those cases where there is appreciable non-linearity as a consequence of toxicity, metabolic factors.

129. The FDA developed rules for the interpretation of results from their experience from the use of CD mice and rats (the mouse and rat strains used in the pharmaceutical industry). They suggested their approach had an overall false positive rate of about 10% for both the standard design and the ICH one species (2 sexes) studies. Another approach uses Haseman's (1983) decision rule. This was based upon comparisons between the top dose and the negative control using data from the standard NTP experimental design studies using F344 rats and B6C3F1 mice. The rule uses a criteria of a statistically significant difference at $P<0.01$ for common and <0.05 for rare tumours (Lin & Ali, 1994). The definition of a rare tumour is an incidence of <1%. Above 1% the tumours are considered common. The EPA has used a one-sided trend test using Haseman's rule when there are 12 or fewer animals with tumours when combined over all groups.

130. It has been argued that Haseman's rule produces too high a false positive rate because all treatment groups rather than just the top dose are, in fact, used in the comparisons. The FDA (2001) suggests that in this case the false positive rate is actually about double the Haseman's rate if the criteria are used with trend tests. The FDA (2001) has proposed a revised decision rules of $P <0.025$ and $P <0.005$ for rare and common tumours respectively which result in an overall false

positive rate of 10% which is considered appropriate in a pharmaceutical industry regulatory environment (Lin & Rahman, 1998).

131. A study where no treatment–related effects are found should be reviewed to ensure it is valid by checking that sufficient numbers of animals lived long enough to ensure that adequate exposure had been achieved and so were 'at risk' of developing 'late in life' tumours. A check should also be made that appropriate dose levels were given to 'challenge' the animals such as by assessing body weight data to ensure that an adequate high dose has been attained.

Use of control data and dual control groups

132. Studies with concurrent controls come in two types: firstly, or Category A, where there is a vehicle control and an untreated control group; secondly, or Category B, where there are two identical control groups (Haseman et al, 1986).

133. With the Category A control groups the objective is to see whether the vehicle which is used in the dosing of the treated groups has any effect on tumour incidence in the control animal (or on other aspects of the test animal such as food intake or body weight) compared with the untreated. The vehicle control group is the comparator of choice.

134. The argument for dual controls (Category B) is that it provides a way of identifying the degree of variability in the negative control animal so providing a better basis for addressing the biological importance of any increase found in the treated groups. They can be considered as 'contemporary' historical control data. Complications arise in the interpretation of the data when the tumour incidences of the two concurrent control groups differ.

135. Data from the two groups (C1 and C2) can usually be combined (Haseman et al, 1990) but if differences are found either in mortality or tumour incidence then three tests with the treatment group (T) should be carried out (C1 v. T) (C2 v. T) and ((C1+C2) v T). Problems arise in the interpretation of these comparisons when the results differ especially when groups C1 and C2 differ. Consideration is needed on how or whether to take into account the multiple comparisons made.

136. Some have argued that if the C1 and C2 groups differ then comparisons with the test groups are only considered positive when comparisons with each control group individually is significant on the grounds that findings should be reproducible. However, if it is accepted the LTRCB is underpowered, then the comparison would be considered positive if any of the comparisons of the treated group with the control groups were significant. The first approach risks more false negatives, the second more false positives.

137. The FDA (2001) did not suggest an adjustment of the significance levels to maintain a preferred 10% overall false positive rate. The results should be considered equivocal unless the three tests involving the control groups give consistent results (either all positive or all negative). In the event of equivocal results the control data should be considered in the context of the historical control data.

Historical control considerations

138. In any discussion about historical control data, it should be stressed that the concurrent control group is always the most important consideration in the testing for increased tumour rates. The historical control data can, though, be useful provided that the data chosen are from studies that are comparable with the study being investigated. It is widely recognized that large differences can result from differences in factors such as pathology nomenclature, strain, husbandry, pathologists.

139. It has been suggested that historical control data should only be used if the concurrent control data are appreciably 'out of line' with recent previous studies and that only historical data collected over the last 5 years should be used. Such historical control data can be helpful in evaluating how 'normal' the data from the concurrent control groups, for evaluating differing results from the dual control groups and as a form of quality control for carcinogenicity studies. Any concerns over the appropriateness of the control groups need to be evaluated and discussed.

140. Historical control data can help interpret results when the tumour incidence rates are low (uncommon or rare tumour types); when the significance levels obtained are borderline/marginal making it difficult to distinguish between a true negative result and a false positive; when there are rare tumours; when common tumours have very variable incidence rates and when the concurrent control incidences are smaller or larger than usually noted. Data in historical control databases can be analysed by both informal and formal methods.

141. An informal approach has been to assess whether the tumour rates in treated groups fall within the range of the historical control rates. It is then argued that if the formal test is at the margins of statistically significance then the result can be explained as a random occurrence of a low concurrent control rate. Similarly a non-significant increase for a rare tumour may be considered a true negative if it falls within the historical control range. Crucial to such an approach is that the historical control rate is 'reliable'. This is, though, not wholly satisfactory because the ranges can be very wide and the range may sometimes be considered as the difference between the maximum and minimum values without any consideration of the distribution of the values risking that one or a small group of studies can have an appreciable effect on the range. As the number of studies increases, the range is likely to increase so the range has limited use.

142. Elmore & Peddada (2009) discussed how to incorporate historical control data into the statistical analysis of the LTRCB. The mean and SD can be affected by a 'rogue' outlier while the median and interquartile range (IQR) is not. They argue outliers should be identified and not discounted but considered alongside other relevant data in the assessment of the results. They suggest the use of exploratory methods such as box and whisker plots (with their associated 5 number summaries) to give graphical presentations of historical control data (when there are more than 15 studies) together with the results of the treated groups and the concurrent negative control in a study. They illustrate the advantages of using the median and quartiles over the mean, SD and range especially to identify potentially misleading outlying results. They recommend using Bailer & Portier (1988) survival adjusted tumour incidence rates.

143. A number of methods has been described for incorporating historical control data into the formal statistical methods used in the analysis of a trend in the data. One suggestion has been to use the upper confidence limit on the binomial proportion to help in interpretation. The FDA (2001) suggests this approach should probably replace the historical range approach. Tarone (1982) developed a method using the beta-binomial distribution (a binomial distribution where the value of p is a random variable rather than a single fixed value to account for the variability between studies to model historical control data and to derive both exact and approximate tests. Including the historical control data increases the power of tests, especially for comparisons with rare tumours but the method does not take into account differential survival. Ibrahim & Ryan (1996) have developed a test which uses historical control data in survival-adjusted tests. The study period is split into intervals and in each of these the multinomial distribution is used to model the number of animals dying with tumours. The prior distribution for the historical control rate is based upon the Dirichelet distributions. The method, though, can only be applied to fatal tumours. Ibrahim et al (1998) developed methods for incorporating historical control data into age-adjusted tests. The approach, though, is limited because it makes strong assumptions about the tumour lethality.

144. Ibrahim and co-workers have also developed a method that assumes all the tumours are lethal (Ibrahim & Ryan, 1996) or all tumours are non-lethal (Ibrahim, Ryan and Chen 1998). The two tests, therefore, represent the extreme event and may not be accurate in practice. Fung et al (1996) developed methods for incorporating historical control data but this approach is also not an age-adjusted test.

145. A Bayesian approach has been suggested by Dempster et al (1983) making the assumption that the logits of the historical control rate are normally distributed. Another Bayesian approach has been developed by Dunson & Dinse (2001) which relaxes some of the assumption regarding the nature of the tumour. The prior probabilities for the parameter in the model, however, have to be chosen carefully and this requires a consensus between the pathologist and the toxicologist.

146. Peddada et al (2007) have proposed a non-parametric approach which provides separate comparisons of the dose groups with the concurrent and historical control data. A third comparison can be made between the two control data sets. The poly-3 correction is made to sample sizes to adjust for survival rate differences between the groups. Consequently, individual data rather than summary data for a group are needed. The three p-values obtained are compared using a ‘weight of evidence’ approach. Without the survival data there is a possibility of bias.

Dose-response modelling

147. Previously most of the emphasis on the LTRCB had been on hypothesis testing and the identification of a carcinogenic hazard. The tests have been about the identification of a significant effect or trend. An alternative approach is an estimation of the size of any dose-response relationship. However, the evolution of the LTRCB towards greater emphasis on quantitative dose response modelling has implication for the study design. The traditional design is limited for the secondary objectives described earlier.

148. The current experimental design is a trade-off between the optimum allocation of equal group sizes between a control and treated group to maximize the power to detect an effect (i.e. test a null hypotheses) and the need to describe in detail the shape of a dose response relationship. Studies need to be designed to identify which parts of the relationship are important and consequently may require a two-stage approach of identifying an area of interest in a preliminary range finding study and then moving to investigate this in more detail by concentrating resources in this region.

Extrapolation to low doses

149. In the current paradigm of risk assessment the objective is to identify the risks associated with human exposure. Risk assessment has traditionally been carried out differently for genotoxic and non-genotoxic carcinogens and non-carcinogens. A no-threshold (no safe dose) model has usually been assumed for genotoxic carcinogens while a threshold below which there are no toxic effects, linked to the concept of the no observed effect level (NOEL) has been assumed for some non-genotoxic carcinogens and non-carcinogenic endpoints.

150. Low dose extrapolation has been carried out by some regulatory agencies by fitting mathematical models to the observational data for carcinogens and then extrapolating the models to the low doses/exposures that might be expected to occur in the human population (EPA, 1986). The EPA used the Linearized Multistage LMS model as the default model for such extrapolations. These approaches aimed to identify, for instance, Virtually Safe Doses (VSDs) where it has been estimated that such life time exposure would lead to an upper bound increase of 1 extra lifetime cancer death in 1 million exposed individuals (or a 10^{-6} lifetime risk) or some similar low risk. The low dose extrapolations were, however, highly dependent upon the mathematical function assumed for the dose-response relationship and could give very different estimates of, for instance, the VSDs. Such low dose extrapolation now seems unrealistic and the approaches have been refined. Modelling is now confined to the observed experimental ranges and a point of departure (POD) identified such as, for example, an estimate of the Benchmark Dose.

151. Some authorities now propose linear extrapolation from the POD by drawing a straight line from the POD to the zero extra/additional risk and reading off the VSD associated with the 10^{-6} excess/additional risk. Others propose using the ratio of the POD to the human exposure to derive a Margin of Exposures (MoE).

NOEL, NOAEL, LOEL, LOAEL

152. A NOEL (no observable effect level) is obtained by finding the highest dose level where there is no significant increase in treatment related effects compared with the negative/vehicle control. The LOEL (lowest observable effect level) is the lowest dose where there is a significant effect. Note that these statistical comparisons are pair-wise comparisons and do not take into account information on the dose-response relationship or the more powerful trend test.

153. The NOEL is usually defined to mean the NOAEL (the no observable adverse effect level, sometimes designated NO(A)EL) to distinguish between changes which are adverse rather than any treatment related effect which may in some case not be adverse. The NOAEL is one example of a reference dose (RD) or Point of Departure (POD). Similarly the LOAEL (or LO(A)EL)) is the lowest observable adverse effect level. There are also NOEC, NOAEC, LOEC and LOAEC where the C refers to concentration rather than level.

154. In the case of those non-genotoxic carcinogens or non-cancer endpoints where there is believed to be a threshold dose below which no toxic effects occur then health-based guidance values such as Acceptable or Tolerable Daily Intakes (ADIs and TDIs) are derived by applying Safety or Uncertainty Factors (SFs or UFs) to a POD such as the NOAEL (No Observable Adverse Effect Level) derived from a study. It is then considered that there are no appreciable health risks below these health-based guidance values (WHO, 1999). The SF/UFs are used to account for inter- and intra-species variability. An extra SF is applied if no NOAEL can be found and the LOAEL is used to derive health-based guidance values.

Limitations of NOAEL approach

155. Although widely used the NOAEL approach has its critics.

The main limitations of the approach are that:

- The NOAEL is based upon the study design, sample sizes, dose spacing, background levels, the power of the design (and statistical test)
- It is based upon a hypothesis testing approach in that failure to reject the null hypothesis is taken as evidence of no difference. This differs from conventional view of the interpretation of the failure to reject a null hypothesis.
- The smaller the experiment or the more variable the endpoint the less chance of detecting a real effect
- The NOAELs tend to be higher for measures with a high control/background level because it is more difficult to demonstrate a statistically significant difference in standard designs than for measures with low control/background levels
- It is a single dose level and takes no account of the dose-response relationship in the data
- It is not an estimate of a dose so cannot be presented with a measure of precision such as a confidence interval (CI)
- It is a single experimental dose level so is not necessarily representative of the true 'threshold'
- The NOAEL cannot be considered a risk or response-free exposure level
- The NOAEL is not necessarily a dose where there is no effect (i.e. below a threshold).

156. There is now an appreciable body of thought that there should be a move away from the use of the NOAEL as the POD towards the use of the Benchmark dose methodology.

157. Other statistical approaches have been developed for identifying change points in a dose-response study as an alternative to the NOAEL for continuous data. The change point is defined

as the largest dose level which has the same response as in the negative control group. West & Kodell (2005) proposed a method that investigates the profile of the least squares criterion over each of the intervals between the dose points in an experiment. They carried out simulation studies to show that the 95% lower confidence interval of the estimate of the change point had better statistical properties than the NOAEL. West & Kodell suggest linking the approach to the BMD methodology but that the method will need to be developed for a range of relevant change point model but note that conventional toxicology studies may have too few doses to estimate the parameters that explain more complex dose-response relationships.

158. The NOSTATOT method (no statistical significance of trend test) identifies the maximal dose which is not significantly different from the negative control group (Tukey et al, 1985). In general the NOSTATOT dose is higher than the no-effect dose.

Benchmark dose approach

159. Criticism and the limitations of the NO(A)EL are well known (see above). An alternative approach, the Benchmark dose (BMD) approach, was first proposed by Crump (1984) as an alternative for the identification of estimates of dose levels helpful for risk assessment. It had a slow uptake despite early work to generalize the concept but has gradually become more widely accepted. EFSA (EFSA, 2005) and JECFA (JECFA, 2005) have proposed the use of the Benchmark Dose (BMD) as the reference dose (RD) for the calculation of the MoE of genotoxic carcinogens.

160. Key points made by proponents is that, through the use of confidence limits, it provides an estimate of uncertainty, that the uncertainty is reduced in large studies with better designs usually leading to higher PODs. It makes more use of all the data in a dose-response experiment. The POD does not have to be one of the experimental dose levels and a POD can be calculated even if there is no NOAEL derived from a study.. It is also argued that the response at the NOAEL is not assessed in terms of the Benchmark Response (BMR) and can vary from case to case but that the BMDL is a consistent and explicit response level (Sand et al, 2008).

161. Acceptance is not universal. Travis et al (2005), for instance, argue against its routine use and that there are issues about how to use it in cases where there is either no LOAEL identified (i.e. no obvious dose-response) or no NOAEL is identified (significant effect at all dose levels). Travis et al argue that the NOAEL is best for routine use in toxicology studies but that the BMD may have a role in the interpretation of the most influential/critical studies in a regulatory package.

162. The Benchmark Dose (BMD) is the dose associated with a pre-specified biological difference. This response, the Benchmark Response (BMR) is a change in the endpoint of interest above the negative control response. In its original form the response was based upon an increase over the control incidence of a quantal measure (such as tumour incidence). Values such as a 5% or 10% additional or extra risk have conventionally been used to define the BMD for quantal data.

Mathematical modelling for the BMD

163. Those endpoints which show 'visual' trends are analysed further to identify if the dose-response relationship is suitable for further analysis by fitting dose-response models to the experimental data. After suitable models have been identified, the BMD and the BMDL are determined for each suitable model. The specific BMD value is determined either by interpolation within the experimental data or by extrapolation just beyond the experimental data. The BMDL is the lower one-sided 95% confidence limit or bound on the BMD value. The BMDL can be interpreted as meaning that there is 95% confidence that the true effect at this dose would be less than the effect associated with the BMDL.

164. Values are obtained using software (BMDS or PROAST) specifically developed for the purposes or using statistical packages such as SAS. (It is important that the various assumptions and defaults underlying these approaches and, in some cases, incorporated in the software are appreciated.) One assumption in the modelling approach is the distribution of the data. In the case of quantal data the assumption is a binomial distribution, in the case of quantitative data, either a normal or a lognormal distribution.

165. The modelling approaches use algorithms to identify the optimal values of the parameters which specify the mathematical model. These values are derived by minimizing the difference between the fitted values and the observed values. One approach to this is the maximum likelihood method where likelihood is a measure of how likely the parameters have these specific values given the observed data; the criterion is the combination of parameter values that maximize this measure.

166. The BMDL is usually estimated using a likelihood ratio test, a method also used for comparing the fits of different models. The term, (minus) twice the difference between the log likelihood of two different (comparable) models, follows a chi-square distribution with degrees of freedom (dfs) equal to the difference in number of parameters in each model. A chi-square value significant at $P<0.05$ is taken by some as evidence that the two models are considered significantly different.

167. The range of models potentially suitable for quantal data includes the 'standard' tolerance functions: probit, logistic and Weibull and log based versions. Others include the multistage and gamma multi-hit model. More complex models are provided in the software to take into account the more complex multilevel data from developmental/teratology data where there are, for instance, intra-litter correlations.

168. A large number of models can be used to describe a quantitative dose-response relationship. Examples include the polynomial, power, exponential and Hill function models (Slob, 2002). The Hill and exponential models have the properties of being sigmoid (S-shaped) and bounded (levelling of at maximum and minimum response values such as 0 and 100%) which is compatible with biological data with parameters that can be easily related to the shape of the D-R relationship. (On a log scale the Hill model is symmetrical but not on the normal dose scale.). Models can also be specified to take into account different variances within the dose groups.

169. In the case of studies where there is a sigmoid/S-shaped dose-response, modelling requires that the dose range must be covered in each region (low, medium and high). The sigmoid curves

models (Hill and exponential) have four parameters so at least five dose levels are needed to avoid over-parameterization (perfect fit) and provide an opportunity to test the model fit. There may also be a problem in identifying the dynamic range if data are not available for the high dose. A family of a specific model, e.g. polynomial, Hill and exponential, are often used in the BMD for continuous data. Testing within the families can be carried out with a test for whether extra parameters improve the fit. If they do not, then by parsimony, they are left out.

170. There is complexity in comparing models across classes. The EPA has suggested the use of Akaike's Information Criteria (AIC). The AIC is a measure of the fit of the model weighted by the number of parameters fitted with the model with the lowest AIC selected. A complication is that when models are similar the relative ranking in terms of AIC may be somewhat arbitrary.

171. In practice, different models with the same number of parameters can often be found to give a satisfactory fit to the same dataset. The approach then is to calculate the BMDs and BMDLs from these various mathematical models and compare the range of values for all acceptable models for their similarity and consistency. The BMDL which is considered most appropriate is identified as the Reference Point (RP) which is used by the different approaches to defining a guidance level for risk assessment (ADI, TDI, MOE etc). The choice of which model is used for subsequent calculations may be based upon criteria such as which gives the lowest or most conservative BMDL. Such choices, therefore, have potentially appreciable input from the modeller.

172. In the case of the selection of the BMDL the case is made for the selection of the lowest (i.e. most conservative) BMDL from those models which fit the data satisfactorily. Model averaging has also been proposed using, for instance, Bayesian model averaging where the averaging is a weighting derived from the support for a particular model taking into account the data (Bailer et al, 2005).

173. The BMR may be expressed either as 'additional' or 'extra' risk. Extra risk is an adjusted rate which includes an adjustment for the background incidence rate and is based only on the fraction who are expected not to have a background incidence with BMR = P(BMD) – P(0) / 1 - P(0). Additional risk is an absolute rate: P(BMD) – P(0). The two terms become the same when P(0), the background frequency, is zero.

174. The extra or additional risk (BMR) used to derive default BMDL values has usually been 5 or 10% for quantal data based upon the similarity between the BMDL derived from them and the NOAEL derived in part from developmental toxicity studies. The BMDL for a 10% risk level initially seemed most similar to the NOAEL derived from the same studies. Using more complex models taking into account the intra-litter correlations suggested the BMDL for a 5% risk level were most similar to the NOAEL.

175. One limitation of modelling for LTRCB data is that it is generally done solely on qualitative data and does not make use of time-to-tumour data. Including these data should be both more informative and provide a more accurate assessment. However, modelling methods using them have not been developed and validated. This is particularly important when there is uncertainty over the cause of an animal's death. Other complexities for such methods are the consequence of early termination of studies or of some groups.

176. In the case of quantitative data the percentage size of effects that are considered biologically important needs to be defined. A number of different BMRs have been proposed.

177. One is related to the CES (critical effect size) Slob & Pieters (1988). The BMR could be based, for instance, upon a 'continuous benchmark response' a proportional, e.g. a 10% change, in the mean body weight over the control values for an adult animal or some fold change in an enzyme level of clinical chemistry value. There remains a debate over what the CES should be and the potential to use within animal variability to define it (Dekkers et al, 2001; 2006). A 5% CES has been suggested in the absence of other information (Woutersen et al, 2001) in part because such a response seemed close to the NOAEL found in some studies.

178. The BMD related to a change in response equal to one standard deviation above the negative control mean have also been suggested (Crump 1984, 1995; Kavlock et al, 1995). The US EPA had suggested using a change equivalent to one standard deviation (1 SD) in the endpoint. A 5% change in BMR has been suggested for foetal weight, 10% for brain cholinesterase and developmental neurotoxicology.

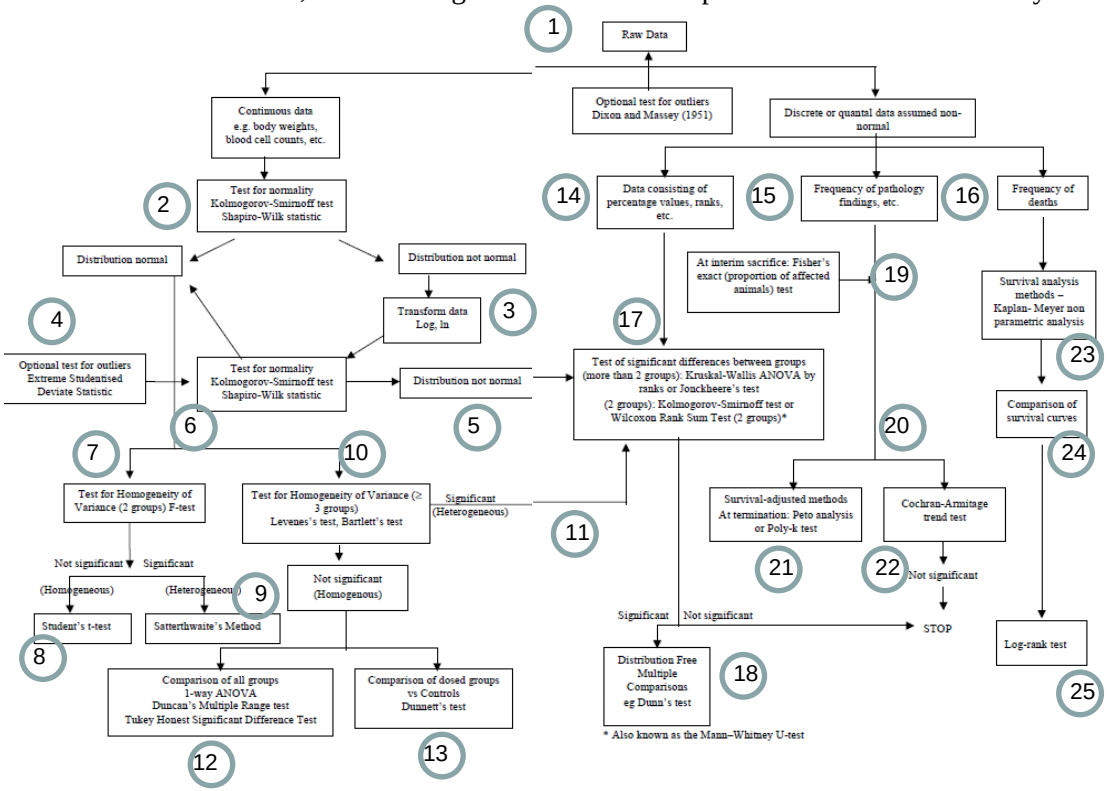
179. An approach based on change relative to the dynamic range (the maximum to minimum values which links to the Hill or exponential model parameters) has been suggested (Murrell et al, 1998). Change relative to the dynamic response may make comparison across endpoints possible. Other options include identifying doses when there are 'change points' on the dose response curve or the steepest point on the sigmoid curve.

180. One consideration is whether quantitative data should be converted into either binary or ordinal data so that it can be handled as if it were quantal data. Quantal data can in this 'model' be thought of as being on a continuous scale with various classes being defined as specific break points. Such dichotomization, however, results in a loss of information.

181. In conclusion, although modelling can appear a precise and a formal process there is the opportunity for an appreciable amount of expert but subjective input into, for instance, the choice of options and the inclusion or exclusion of outliers and anomalous curve fits.

Figure 1

A statistical decision tree, summarising common statistical procedures used for the analysis of data in long-term toxicology studies.



Appendix 1 Common statistical methods used in the analysis of data

The following glossary gives a brief description of the various tests in the flowchart (Figure 1) plus other tests that are commonly encountered in chronic studies. The definitions below are based upon a number of widely available glossaries and in particular Everitt (1995).

Tests for outliers

Dixon/Massey test: A test for outliers in a sample
Extreme Studentized Deviate (ESD) Statistic: A method used for identifying outliers; also known as Grubbs' test

Tests for non-normality
Chi-square test: A goodness of fit test that a set of data come from a hypothesized distribution such as the normal.
Kolmogorov–Smirnov one-sample test: A method that tests for goodness of fit of the data to a defined distribution
Shapiro–Wilk test: A method that tests that a set of random variable arise for a specified probability distribution used to test for departure from normality.

Tests for homogeneity of variance
Bartlett's test: A test for the equality (homogeneity) of the variances of a number of samples. The test is sensitive to departures from normality
Levene's test: A test for the equality (homogeneity) of the variances of a number of samples. The test is less sensitive to departures from normality
F test of variances: A test for a difference in the size of two variances.

Assumed normally distributed data
1. Overall tests

Analysis of variance (ANOVA): Statistical methodology which partition variability attributable to various causes. Family of modelling approaches simplest and commonest of which is the fixed effects one-way anova which compares means across a set of samples.
Analysis of covariance (ANCOVA): Extension of ANOVA which allows for possible effects of covariates on endpoint in to effects of treatments which may reduce error mean squares associated with analysis.
Pearson's correlation coefficient: A test of the association between two variables
Linear regression: A test of the relationship between the two variables: one the independent like the dose the other the dependent i.e. the response. Used to examine trends in dose effects and to test the significance of the regression slopes.

2. Pair-wise comparisons

Duncan's multiple range test: A modified version of the Newman-Keuls multiple comparison test used to test for multiple comparisons when the initial ANOVA between groups is significant.
Dunnett's t-test : A multiple comparison test which compares each of a number of treatments to a single control.
Scheffe's test: A multiple comparison test with less power than Newman–Keuls multiple range test.
Williams' t-test: A multiple comparisons method for comparing each of a number of treatments with a single control.
Student's t-test: A number of different tests but here the independent two sample t-test assuming equal variance in the two groups and testing for a difference between two means.

Satterthwaite test: An alternative to the pooled-variance t test, and is used when the assumption that the two populations have equal variances seems unreasonable. It provides a t statistic that asymptotically (that is, as the sample sizes become large) approaches a t distribution, allowing for an approximate t test to be calculated when the population variances are not equal. Also known as Welch's test,
Fisher's least significant difference (LSD) test: A pair-wise test equivalent to the independent two-sample t-test except that the estimate of error is based upon the within group error of an ANOVA.
Tukey's Honest Significant (HSD) Difference test: A single step multiple comparison method used after the initial ANOVA between groups is significant.

Non-parametric procedures (percentage values, ranks, etc.)
Kendall's coefficient of rank correlation: A non-parametric test of the association between two variables based upon ranks
Pearman's rank correlation: Another non-parametric test of the association between two variables based upon ranks
Mann–Whitney U-test: A non-parametric alternative to the independent two-sample t-test. Also called the Wilcoxon Rank Sum Test.
Kolmogorov–Smirnov two-sample test: A distribution-free method that tests for any difference between two population distributions.
Wilcoxon signed-rank test: A non-parametric alternative to the paired t-test for matched or paired data.
Kruskal–Wallis ANOVA test: A distribution-free method that is the analogue of the one-way analysis of variance. Which tests whether the groups to be compared have the same population mean.
Jonckheere-Terpstra test: A test for detecting departures form independence where both the rows and columns of a contingency table have a natural order.

Distribution-free multiple comparisons tests
Dunn's test: A multiple comparison test based upon the Bonferroni test
Shirley's test: A non-parametric equivalent of Williams' test.

Quantal data (mortalities, pathology findings, etc.)
Fisher's exact test: Test for independence of two variables forming a 2 x 2 contingency table, based upon the hypergeometric distribution.
R x C chi-square test: A measure of association between the row and column classification or a r x c contingency table of variables
Litchfield & Wilcoxon test: Graphical method of probit analysis for calculation ED50 and confidence intervals
Cochran-Armitage linear trend test: Chi-square test for linear trend in counts and proportions.

Multivariate methods
Hotellings T^2: A generalization of the t-test to multivariate data
MANOVA: A multivariate analysis of variance to test the equality of the means of more than 2 populations.

Survival-adjusted procedures for analysis of carcinogenicity data
Log-rank test: Compares the survival distributions of two or more samples, sometimes called the Mantel-Cox test.
Peto analysis: A test in IARC monograph combining a life table test for fatal tumours with a prevalence analysis for incidental tumours.
Life table test: A survival adjusted test for fatal cancers or cancers with observable onset times.
Hoel–Walberg procedure: A survival adjusted test for incidental tumours. Also called the prevalence method.
Logistic regression: A form of regression analysis used when the response is binary: i.e. tumour/no tumour.

Poly-k test: A survival-adjusted Cochran-Armitage test for testing for a dose-related trend and/or a pair-wise difference in the incidence of tumours.

References

Ahn, H., Kodell, R. L. & Moon, H. (2000). Attribution of tumor lethality and estimation of time to onset of occult tumors in the absence of cause-of-death information. Applied Statistics 49 157-169.

Bailer, A.J., Noble, R.B. & Wheeler, M.W. (2005) Model uncertainty and risk estimation for quantal responses. Risk Analysis 25 291-299.

Bailer, A.J. & Portier, C.J. (1988). Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. Biometrics 44 417-431.

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

Bieler, G.S. & Williams, R.L.(1993). Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity. Biometrics 49 793-801

Breslow, N. (1970) A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship, Biometrics 57 579–594.

Butterworth, B. E., Mathre, O.B, Ballinger, K. E. & Adalsteinsson O.(2004) Contamination is a Frequent Confounding Factor in Toxicology Studies with Anthraquinone and Related Compounds. Int J Toxicol. 23 335-344

Cox, D.R. (1959) The analysis of exponentially distributed life-times with two types of failures. Journal of Royal Statistical Society, Series B, 21 4121-421.

Cox, D.R. (1972) Regression models and life tables (with discussion). Journal of Royal Statistical Society, Series B, 34, 187-220.

Crump, K. S. (1984) A new method for determining allowable daily intakes. Fund. Appl. Toxicol. 4 854-871.

Crump, K. S. (1995) Calculation of benchmark doses from continuous data. Risk Analysis 15 79-89.

Dekkers, S., de Heer, C. & Rennen, M. (2001). Critical effect sizes in toxicological risk assessment: A comprehensive and critical evaluation. Environ. Toxicol. Pharmacol. 10 33–52.

Dekkers, S., Telman ,J., Rennen, M.A.J., Appel M.J. & and de Heer, C. (2006) Within-animal variation as an indication of the minimal magnitude of the critical effect size for continuous toxicological parameters applicable in the benchmark dose approach, Risk Analysis 26 867–880.

Dempster, A.P., Selwyn, M.R. &. Weeks B.J (1983) Combining historical and randomized controls for assessing trends in proportions. Journal of the American Statistical Association 78 221-227.

Dinse, G.E. (1994) A comparison of tumor incidence analyses applicable in single-sacrifice animal experiments. Statistics in Medicine 13 689-708.

Dinse, G.E. &. Haseman, J.K 1986) Logistic regression analysis of incidental-tumor data from animal carcinogenicity experiments. Fundamental and Applied Toxicology 6 751-770.

Dunson, D. B. & Dinse, G. E. (2001). Bayesian incidence analysis of animal tumorigenicity data. Appl Stat 50, 125–41.

Elmore, S. & Peddada, S.D. (2009) Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. Toxicologic Pathology 37 672-676.

EFSA (2005) Opinion of the Scientific Committee on a request from EFSA related to a harmonised approach for risk assessment of substances which are both genotoxic and carcinogenic. The EFSA Journal 282 1–31.

EPA (1986) Guidelines for Carcinogen Risk Assessment U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC.

EPA (1996) Proposed Guidelines for carcinogen risk assessment. U.S. Environmental Protection Agency (USEPA), Fed. Register 61 17960–18011

Everitt, B.S. (1995) The Cambridge Dictionary of Statistics in the Medical Sciences. Cambridge University Press.

FDA Redbook (2000) Food and Drug Administration/Center for Food Safety and Applied Nutrition Toxicological Principles for the Safety Assessment of Food Ingredients.

FDA (2001) Guidance for Industry: Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals. *Draft Guidance*. Silver Spring, MD: FDA/CDER.

Fisher R.A. (1932) Statistical methods for research workers. (3rd edition) Oliver and Boyd, Edinburgh.

Fisher, R.A. (1950) Statistical methods for research workers (11th edition). Oliver and Boyd, Edinburgh.

Fung, K.Y., Krewski, D. & Smythe, R.T. (1996) A comparison of tests for trend with historical controls in carcinogen bioassay. Can. J. Statist. 24 431-454

Gad, S.D. (2000) Statistics for Toxicologists. In Principles and Methods of Toxicology. Ed: A. Wallace Hayes. 4th edition. Taylor and Francis: Boston, MA, pp 369-451

Gad, S. (2006) Statistics and Experimental Design for Toxicologists and Pharmacologists. 4th Edition, Taylor and Francis, Boca Raton, FL.

Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. & Wahrendorf, J. (1986) Statistical Methods in Cancer Research, III The design and analysis of long-term animal experiments. IARC Scientific Publications 79. vol. III pp. 1–219.

Gehan, E.A. (1965)A generalized Wilcoxon test for comparing K samples subject to unequal patterns of censorship. Biometrika 52 203-223.

Graves, T.S. (1994) A comparison of test procedures for multidose animal carcinogenicity assays under competing risks. Journal of Biopharmaceutical Statistics. 4 289 - 320

Haseman, J.K. (1983) A reexamination of false-positive rates for carcinogenesis studies. Fundamental and Applied Toxicology 3 334-339.

Haseman, J.K. (1984) Statistical Issues in the design, analysis and interpretation of animal carcinogenicity studies. Environmental Health Perspective 58 385-392.

Haseman, J.K., Hajian, G. Crump, K.S. Selwyn, M.R.& Peace, K.E. (1990) Dual control groups in rodent carcinogenicity studies. In Statistical Issues in Drug Research and Development, Ed, Peace, K. E. Marcel Dekker, New York.

Haseman, J.K., Huff, J. & Boorman, G.A (1984) Use of historical control data in carcinogenicity studies in rodents. Toxicologic Pathology 12 126-135.

Haseman, J.K., Huff, J.E., Rao, G.N. & Eustis S.L.(1989) Sources of variability in rodent carcinogenicity studies. Fundam Appl Toxicol. 12 793-804.

Haseman, J.K., Winbush, J.S. & O'Donnell, M.W. (1986) Use of dual control groups to estimate false positive rates in laboratory animal carcinogenicity studies. Fundamental and Applied Toxicology 7 573-584.

Hoel D.G. & Walburg,. H.E. (1972) Statistical analysis of survival experiments. J Natl Cancer Inst 49 361-372.

Iatropoulos, M.J. (1984) Appropriateness of methods for slide evaluation in the practice of toxicologic pathology. Toxicol Pathol 12 305-306.

Iatropoulos, M.J. (1988) Society of Toxicologic Pathologists position paper: "Blinded" microscopic examination of tissues from toxicologic or oncogenic studies. in Carcinogenicity, The Design, Analysis, and Interpretation of Tong-Term Animal Studies, edited by H.C. Grice and J.L. Ciminera, ILSI Monographs, Spring-Verlag, New York

Ibrahim, J.G. & Ryan L.M. (1996) Use of historical controls in time-adjusted trend tests for carcinogenicity. Biometrics 52 1478-1485.

Ibrahim, J.G., Ryan, L.M. & Chen M.-H. (1998). Use of historical controls to adjust for covariates in trend tests for binary data. J. Amer. Statist. Assoc. 93 1282-1293.

ICH (1995). International Conference on Harmonization Guidance for Industry S1C Dose Selection for Carcinogenicity Studies of Pharmaceuticals.

ICH (1997). International Conference on Harmonization. Guidance for Industry S1B Testing for Carcinogenicity of Pharmaceuticals.

JECFA (2005) Joint FAO/WHO Expert Committee on Food Additives Sixty-fourth Meeting, Rome. Summary and Conclusions. FAO/WHO.

Kaplan, E.L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. 53 457-481

Kavlock, R. J., Allen, B. C., Faustman, E. M. & Kimmel, C. A. (1995). Dose–response assessments for developmental toxicity. IV. Benchmark doses for fetal weight changes. Fundam. Appl. Toxicol. 26 211–222.

Keenan, K.P., Laroque, P., Ballam, G.C., Soper, K.A., Dixit, R., Mattson, B.A., Adams, S. & Coleman, J.B.(1996) The effects of diet, ad libitum overfeeding, and moderate dietary restriction on the rodent bioassay: the uncontrolled variable in safety assessment. Toxicologic Pathology 24 757-768.

Kodell, R.L, Farmer, J.H., Gaylor, D.W. & Cameron, A.M. (1982) Influence of cause-of-death assignment on time-to-tumor analyses in animal carcinogenesis studies. J Natl Cancer Inst. 69 659-64

Kodell, R.L., Lin, K.K., Thorn, B.T. & Chen, J.J. (2000). Bioassays of shortened duration for drugs: Statistical Implications. Toxicological Sciences 55 415-432

Lee, P.N., Fry, J.S., Fairweather, W.R., Haseman, J.K., Kodell, R.L., Chen, J.J., Roth, A.J., Soper, K.,& Morton, D. (2002) Current issues: statistical methods for carcinogenicity studies. Toxicological Pathology 30 403–414.

Lin, K.K. & Ali, M.W. (1994) Statistical review and evaluation of animal tumorigenicity studies. In Statistics in the Pharmaceutical Industry. 2nd Ed., Ed. Buncher, C.R & Tsay J.Z. Marcel Dekker, New York.

Lin, K.K.& Rahman, M.A (1998) Overall false positive rates in tests for linear trend in tumor incidence in animal carcinogenicity studies of new drugs (with discussions). Journal of Pharmaceutical Statistics 81 1-22.

McConnell, E.E., Solleveld, H.A., Swenberg, J.A. & Boorman, G.A.(1986) Guidelines for combining neoplasms for evaluation of rodent carcinogenesis studies. J. Natl. Cancer Inst. 76 283-289.

Murrell, J.A., Portier C. J. & Morris R. W. (1998) Characterizing dose-response I: Critical assessment for the benchmark dose concept. Risk Anal. 18 1326.

OECD (2009) Organization for Economic Cooperation and Development Guidelines for testing of chemicals. Carcinogenicity studies. Test Guidelines TG 451 452 453. Paris, France.

OECD (2002) Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies, Series on Testing and Assessment No. 35 and Series on Pesticides No. 14, ENV/JM/MONO(2002)19, OECD, Paris.

Peddada, S.D., Dinse, G & Kissling, G (2007) Incorporating historical control data when comparing tumor incidence rates. J. Amer. Stat. Assoc. 102 1212-1220.

Peddada, S.D. & Kissling, G.E. (2006) A survival-adjusted quantal-response test for analysis of tumor incidence rates in animal carcinogenicity studies. Environmental Health Perspectives 114 537-541.

Peto R. (1974) Guidelines on the analysis of tumour rates and death rates in experimental animals. British J Cancer 29 101-105.

Peto, R., Pike, M.C., Day, N.E., Gray, R.G., Lee, P.N., Parish, S., Peto, J., Richards, S. & Wahrendorf, J. (1980) Guidelines for simple sensitive significance tests for carcinogenic effects in long-term animal

experiments. In: IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, supplement 2: Long-term and Short-term Screening Assays for Carcinogens: A Critical Appraisal. Lyon: International Agency for Research on Cancer 311-346.

Portier, C.J., Hedges, J.C. & Hoel, D.G. (1986) Age-specific models of mortality and tumor onset for historical control animals in the National Toxicology Program's carcinogenicity experiments. Cancer Res 46 4372−4378.

Portier, C. & Hoel, D.G. (1983) Optimal bioassay design under the Armitage-Doll multistage model. J Toxicol Environ Health 12 1−19.

Portier, C. & Hoel, D.G. (1984) Design of animal carcinogenicity studies for goodness-of-fit to multistage models. Fundam Appl Toxicol. 4 949−959.

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

Rahman, M.A. & Lin, K.K. (2008) A comparison of false positive rates of Peto and Poly-3 methods for long-term carcinogenicity data analysis using multiple comparison adjustment method suggested by Lin and Rahman. J Biopharm Statist 18 949-958.

Rhomberg, L.S. (2005) Seeking optimal design for animal bioassay studies. Toxicological Sciences 84 1-3

Rhomberg, L.R., Baetcke, K., Blancato, J., Bus, J., Cohen, S., Conolly, R., Dixit, R., Doe, J., Ekelman, K., Fenner-Crisp, P., Harvey, P., Hattis, D., Jacobs, A., Jacobson-Kram, D., Lewandowski, T., Liteplo, R., Pelkonen, O., Rice, J., Somers, D., Turturro, A,, West, W. & Olin, S. (2007) Issues in the design and interpretation of chronic toxicity and carcinogenicity studies in rodents: approaches to dose selection. Crit. Rev. Toxicol. 37 729-837.

Sand, S., Victorin, K. & Filipsson, A.F. (2008) The current state of knowledge on the use of the benchmark dose concept in risk. J Appl Toxicol. 28 405-21.

Schaffer, K.A., Sellers, R. & Barale-Thomas, E. (2008) Letter to the Editor Toxicologic Pathology 36 1018-1019.

Slob, W. (2002) Dose–response modeling of continuous endpoints. Toxicol. Sci. 66 298–312.

Slob, W. & Pieters, M.N. (1998). A probabilistic approach for deriving acceptable human intake limits and human health risks from toxicological studies: General framework. Risk Analysis 18 787–798.

Snedecor G,W. & Cochran W. G (1967) Statistical Methods (6th edition). The Iowa State University Press, Ames, Iowa.

Snedecor G.W. &, Cochran WG (1980) Statistical Methods, (7th edition). The Iowa State University Press, Ames, Iowa.

STP Working Group (2001) Draft recommendations on classification of rodent neoplasms for Peto analysis. Toxicologic Pathology 29 265-268.

STP Peto Analysis Working Group (2002). The Society of Toxicologic Pathology's Recommendations on Statistical Analysis of Rodent Carcinogenicity Studies. Toxicologic Pathology 30 415–418.

Tarone, R.E. (1975) Tests for trend in life table analysis. Biometrika 62 679-682.

Tarone, R.E. (1982) The use of historical control information in testing for a trend in proportions. Biometrics 38 215-220.

Tarone, R.E. & Ware, J.(1977) On distribution-free tests for equality of survival distributions. Biometrika 64 165-60.

Temple, R.T., Fairweather, W.R., Glocklin, V.C. & O'Neill, R.T. (1988) The case for blinded slide reading. Comments on Toxicology 2 99-109.

Thomas, D.G., Breslow, N. & Gart, J.J. (1977) Trend and homogeneity analyses of proportions and life table data. Computer and Biomedical Research 10 373-381.

Travis K., Pate, I. & Welsh. Z. (2005) The role of the benchmark dose in a regulatory context. Regulatory Toxicology and Pharmacology 43 280-291

Tukey, J.W., Ciminera, J.L. & Heyse, J.F. (1985) Testing the statistical certainty of a response to increasing doses of a drug. Biometrics 41 295-301.

van Belle, G. (2009) Statistical Rules of Thumb. (2nd Edition Wiley Hobeken

WHO (1999) Evaluation of Certain Food Additives and Contaminants. (Forty-Ninth Report of the Joint FAO/WHO Expert Committee on Food Additives). WHO Technical Report Series, No. 884.

West, R.W. and Kodell, R.L (2005) Changepoint alternatives to the NOAEL. Journal of Agricultural, Biological and Environmental Statistics 10 197-211.

Woutersen, R.A., Jonker, D., Stevenson, H., te Biesebeek, J. D. and Slob. W. (2001) The benchmark dose approach applied to a 28-day toxicity study with rhodorsil silane in rats: the impact of increasing the number of dose groups. Food Chem. Toxicol. 39 697-707.