16
Dose Response Relationships

Dose and Dosage

Dose-response relationship is the association between the dose administered and the response/s that is/are exhibited. Response/s and dose are causally related (Eaton and Klaassen, 1996). Establishing a cause–response relationship is very important in the analysis/assessment of a risk (Christensen *et al*., 2003). Though the terms 'dose' and 'dosage' refer to more or less a same thing, there is a difference between these two terms. Dose refers to a stated quantity or concentration of a substance to which an organism is exposed and is expressed as the amount of test substance per unit weight of test animal (example, mg/kg body weight), whereas dosage is a general term comprising the dose, its frequency and the duration of dosing. Dosages often involve the dimension of time (example, mg/kg body weight/day) (Hayes, 1991).

Margin of Exposure, NOAEL, NOEL

Determining the presence or absence of a dose-response relationship is one of the primary criteria of a risk assessment (IPCS, 2009). In drug development, assessment of dose-response should be an integral part in the study design. The studies should be designed to assess dose-response an inherent part of establishing the safety and effectiveness of the drug (EMEA, 2006). Once a dose-response relationship is established for a test substance, the margin of exposure is determined. The margin of exposure lies between a defined point on the dose-response relationship and the human exposure level. In animal experiments, NOAEL (No-observed-adverse-effect-level) and NOEL (No-observed-effect-level) on the dose-response curve are usually considered as this defined point. Though in reality, both NOAEL and NOEL have similar meaning, JECFA (Joint

FAO/WHO Expert Committee on Food Additives) differentiated between the terms NOEL and NOAEL in risk assessments with the following definitions (WHO, 2007):

NOEL: Greatest concentration or amount of a substance, found by experiment or observation, that causes no alteration of morphology, functional capacity, growth, development, or lifespan of the target organism distinguishable from those observed in normal (control) organisms of the same species and strain under the same defined conditions of exposure.

NOAEL: Greatest concentration or amount of a substance, found by experiment or observation, which causes no detectable adverse alteration of morphology, functional capacity, growth, development, or lifespan of the target organism under defined conditions of exposure.

An adverse response is defined as 'change in morphology, physiology, growth, development or life-span of an organism which results in impairment of the functional capacity or impairment of the capacity to compensate for additional stress or increase in susceptibility to the harmful effects of other environmental influences'. Decisions on whether or not any effect is adverse requires expert judgment (WHO, 1994). This definition shows that the environmental standard setting in general is adjusted to subtle effects which represent early steps in biological effect chains or can be interpreted as first signs of a pathological process (Neus and Boikat, 2000). An alternative approach is to classify dose-related effects in to physiological, toxic and pharmacological responses (OECD, 2000a). Physiological responses are not considered as adverse responses. For example, changes in pulse rate or respiration rate as long as it occurs within the normal functioning of the animal. Changes in physiological function as a result of interaction of a test substance with a cellular receptor site are considered as pharmacological responses. Pharmacological responses are reversible and of short duration, and can be adverse if they cause harm to the animals. Toxic responses are adverse and they can be reversible or irreversible. A chemical which causes a physiological or pharmacological effect may produce a toxic response if the exposure is prolonged and/or if the dose is increased beyond a certain level.

But, there is no consistent standard definition of NOAEL (Dorato and Engelhardt, 2005). In an FDA document (FDA, 2005) NOAEL is defined as the highest dose level that does not produce a significant increase in adverse effects in comparison to the control group. Any biologically

significant effect is considered as an adverse effect, which may or may not statistically significant. NOEL refers to any effect, which may or may not be an adverse one. The definition of the NOAEL, in contrast to that of the NOEL, reflects the view that some effects observed in the animal may be acceptable pharmacodynamic actions of the therapeutic and may not raise a safety concern (FDA, 2005). Some other terminologies related to dose-response relationship are LOEL (Lowest-Observed-Effect Level), LOAEL (Lowest-Observed-Adverse-Effect Level) and threshold dose. LOEL is the lowest dose of a test substance which causes effects distinguishable from those observed in control animals and LOAEL is the lowest dose of a test substance which causes adverse changes distinguishable from those observed in control animals. Threshold dose is the minimum dose required to elicit a response. NOAEL has lot of importance in the clinical development of a drug. For example, the calculation of the first dose in man is based on NOAEL (EMEA, 2007). We may briefly explain some of the practical issues in determining NOEL/NOAEL.

Determining NOEL and NOAEL

One of the main objectives of conducting repeated-dose toxicity studies is to arrive at NOEL or NOAEL. Most of the regulatory guidelines prescribe that the repeated-dose toxicity studies with rodents should be conducted with a minimum of three treatment doses (low, mid and high doses) and a control group (OECD, 1995). The low dose level is carefully selected so that the animals exposed to this dose level will not show any effect of the treatment compared to the control dose. But, most of the repeated-dose toxicity studies show some effect of the treatment in few parameters of the low dose group. In such cases considering the low dose as an NOEL/NOAEL may be questionable. Kobayashi *et al*. (2010) investigated 109 numbers of 28-day repeated dose administration studies in rats and examined the measurable items (functional observational battery, urinalysis, hematology, blood chemistry and absolute and relative organ weights) of the low dose group. Their investigation revealed that, 205/12167 (1.6%) measurable items showed a significant difference ($P<0.05$) in the low dose groups compared to the respective controls. The authors concluded from the investigation that the low dose may be considered to be NOEL, if the significant difference of the measurable items showed by this dose group is about 2% (maximum <5%), compared to the control. However, due consideration may be given to the clinical relevance of the items that showed a significant difference.

It is not uncommon to encounter situations in repeated-dose toxicity studies where mid dose group alone shows an insignificant difference compared to control, whereas low and high dose groups show a significant difference. The guidelines do not mention how to determine the mid dose, except an indication that this dose is required to examine dose dependency. According to Gupta (2007), the mid dose selection should consider threshold in toxic response and mechanism of toxicity. Determining the mid dose is as important as determining the high and low doses in repeated-dose toxicity studies, since mid dose plays a determining role in establishing the dose dependency. For determining dose-related trend in repeated-dose toxicity studies, Williams' test is generally carried out (Bretz, 2006). The disadvantage of Williams' test is that it uses an estimated value for the mean rather than the original mean value for the analysis. Hence, it is likely that Williams' test may indicate a dose-related trend, when it actually does not exist (Williams' test is covered in detail in Chapter 11). Therefore, to analyse such data the use of Dunnett's multiple comparison test for comparing each dose group with the control, followed by Jonckheere's trend test for examining dose-related trend is recommended.

Benchmark Dose

NOAEL is based on a single data point and it does not consider the shape of the dose-response curve, the number of animals in the group, or the statistical variation in the response and its measurement (EPA, 1998). An alternative approach to NOAEL is the Benchmark dose approach (Kimmel and Gaylor, 1988). The Benchmark dose is defined as the dose of a chemical that is required to achieve a predetermined response of a toxicological effect (Sand *et al.*, 2006). The Benchmark dose method uses the full dose response data for the statistical analysis, hence the result obtained from the analysis is considered to be more reliable than the single data point based NOAEL. Unlike the NOAEL approach, the Benchmark dose method includes the determination of the response at a given dose, the magnitude of the dose at a given response and their confidence limits. According to EPA SAB (1998): "The [categorical regression] process makes use of every bit of data available. The underlying premise of the approach is that the severity of the effect, not the specific measurement or outcome incidence, is the information needed for assessing exposure-response relationships for non-cancer endpoints…. All the available data is plotted on a single chart and one can immediately see a rough picture of the

level of the concentration multiplied by time values that can be expected to cause adverse effects of varying severity." The U.S. EPA's CatReg Program (Strickland, 2000) utilizes categorical regression to establish the relationship between concentration, time, and severity of the resulting effect. Response variability and uncertainty are addressed by confidence limits bounding the derived relationship curves. Three statistical models (Logit, Probit and Complementary Log-Log) are available in the CatReg program.

Probit Analysis

Probit analysis was originally published in Science by Bliss (Bliss, 1934). He was an entomologist and was involved in research to find a pesticide to control insects that fed on grape leaves (Greenberg, 1980). Bliss transformed the percentage mortality into a "probability units" (or "Probits") and plotted the 'Probits' against concentrations. But, he did not have a statistical tool to compare the effects among various pesticides. In 1952, Finney of the University of Edinburgh wrote a book, 'Probit Analysis' (Finney, 1952). Probit analysis, a preferred method for analyzing dose-response relationship even today described elaborately in Finney's book, is based on the idea developed by Bliss. One of the assumptions of Probit analysis is that the response *vs* dose data are normally distributed, if not, Finney suggested using the logit over the Probit transformation (Finney, 1952). Both Logit analysis (Muhammad *et al*., 1990) and Probit analysis (Finney, 1978) are used in biological assays.

Performing Probit analysis manually is tedious. An example is provided below to show the steps involved in this statistical analysis. Most of the commercially available statistical software can perform Probit analysis.

Groups of rats (10 rats/group) were given a drug at different dose levels. The response shown by the number of animals at each dose level is given in Table 16.1.

Let us plot a graph with dose on X axis and percent response on Y axis (Figure 16.1).

The very purpose of carrying out the Probit analysis is to find out that dose which causes the response in 50% of the animals. If the response that we are looking at is mortality, the dose that causes mortality in 50% of animals is called as LD_{50}. Since the inception of the LD_{50} test by Trevan (1927), the test has gained wide acceptance as a measure of acute toxicity of all types of substances (DePass, 1989).

Table 16.1. Response shown by rats following the administration of a drug

Dose (mg/kg b.w.)	Response shown by number of animals	Percent Response
5	0	0
6	0	0
8.6	1	10
10.3	1	10
12.4	4	40
14.9	6	60
17.9	7	70
21.5	9	90
25.8	10	100
31	10	100

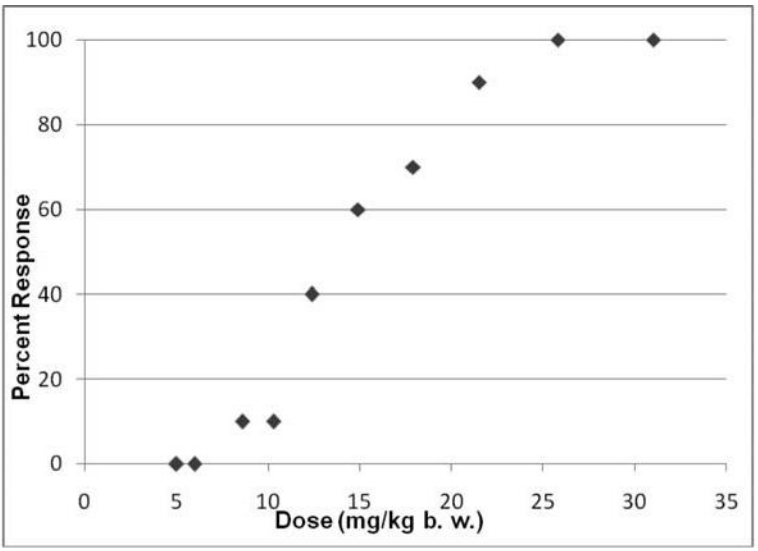

Figure 16.1. Dose *vs* response plot

We could have determined the dose which causes 50% response (for example, LD_{50}) straight away from the plot, had the plot been a straight line. In Finney's Probit analysis the dose response curve is converted to a straight line by transforming the doses to logarithmic values and percent mortality to Probit values (Finney, 1971). Let us try to understand what Probit values means. Percent response on Y axis can be converted to normal equivalent deviation (NED). What is an NED? We know that at one standard deviation below mean value (–1SD), 16% will show response and one standard deviation above mean value (+1SD) 84% will show

the response. Such a relationship can be established between standard deviation and response. The response converted to the corresponding standard deviation is termed as NED. NEDs of below 50 percent response are negative numbers and above 50 percent response are positive numbers. To make the subsequent calculation steps easier, the negative numbers can be converted to positive numbers by simply adding 5 to all NEDs. Now these NEDs are called as probability units or Probits. Finding the Probits for percent response using the above steps is cumbersome. Probit value of a percent response can be directly read from the 'Probit Table' given in several statistical books. Such a Table is given hereunder in an abridged form (Table 16.2).

Table 16.2. Transformation of percentage response to Probit values

Percentage Response	0	10	20	30	40	50	60	70	80	90	100
Probits	-	3.72	4.16	4.48	4.75	5.00	5.25	5.52	5.84	6.28	-

Lets us now plot a graph with log dose on X axis and Probit on Y (Figure 16.2.).

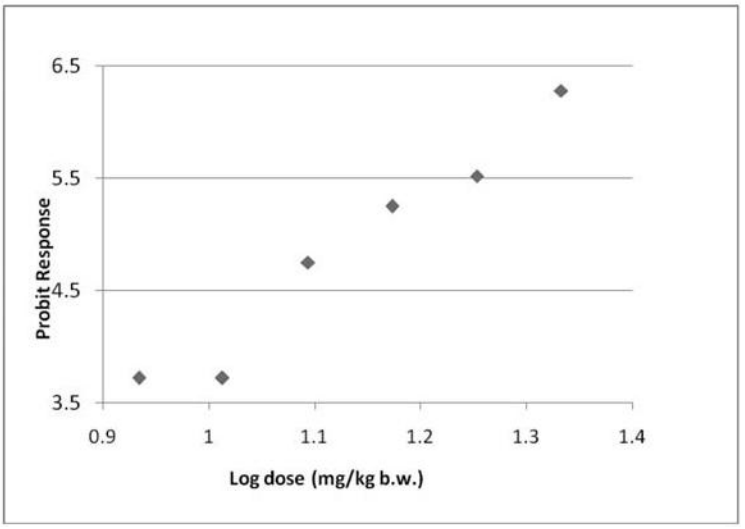

Figure 16.2. Log dose *vs* Probit response plot

You would have observed that Probit responses for 4 doses are missing in the Figure 16.2. The reason for this is that there are no Probit values for 0% and 100% responses. From the Figure one can find that the Probit values somewhat fall in a linear fashion. Let us closely observe the Probit values.

The middle region of the line (region of 50% response, *i.e*., the region of Probit 5) is linear, hence this region is somewhat reliable for making a prediction. The two ends of the line, where the data are controlled by few animals, are not so linear in fashion, hence these regions are seldom used for making a prediction. The variation in the middle region of the line is less, whereas it is on the higher side in the 2 ends. This variation can be minimised by using weighting coefficients. Once a best-fit line is drawn using a regression equation, a 'statistically reliable median response dose' can be estimated:

$\overline{Y} = a + b\overline{X}$, where

$\overline{Y}$ = 5 (Probit value corresponding 50% response)

$\overline{X}$ = Log dose

a = Intercept

b = Slope

Mentioning the term 'statistically reliable median response dose', is intentional as several reports have stated that 'median response dose', for example, LD_{50} is notoriously variable. Usefulness of LD_{50} test has been criticized, as the test only expresses mortality; the test requires large number of animals and the outcome of the LD_{50} test is influenced by several factors associated with the animal (for example, species, age, sex, etc.), animal house condition (for example, temperature, humidity, light intensity, etc.) and human error; many times the findings of the test cannot be extrapolated to man. On the contrary, supporters of the LD_{50} test are of the opinion that a properly conducted LD_{50} test can yield information on the cause and time of death, symptomatology, nonlethal acute effects; slope of the mortality curve can provide information on the mode of action and metabolic detoxification; the results can be used for the basis for designing subsequent subchronic studies; the test is first approximation of hazards to workers (Hodgson, 2010).

This method for calculating LD_{50}, requires a large number of animals, thus, not desirable. Interested readers may refer to the Up and Down Procedure, which requires less number of animals (OECD, 2000b).

IC_{50} and EC_{50} Determination

IC_{50} and EC_{50} determinations are performed for assessing pharmacological affinity of new pharmaceutical compounds. IC_{50} is the concentration of the

compound that provides 50% inhibition, whereas EC_{50} is the concentration that provides 50% of compound's maximal response. IC_{50} is determined for competition binding assays and functional antagonist assays, whereas EC_{50} is determined for agonist/stimulator assays. The procedure for determining IC_{50} and EC_{50} is similar.

For fitting an IC_{50}/EC_{50} curve, first convert the data into percentage inhibition/ percentage activity depending up on the assay performed. If the assay is carried out in replicates find the median percentage inhibition/percentage activity for each concentration. Plot a graph of log concentration *vs* percentage inhibition/percentage activity. The dose-response relationship can be derived using the Hill-slope model. It is also known as four parametric logistic model (4PL). The 4PL function is widely used in biological assays (Healy, 1972; Rodbard *et al*., 1978). The 4PL model equation is given below:

$$y = \text{Minimum Asymptote} + \frac{(\text{Maximum Asymptote} - \text{Minimum Asymptote})}{1 + (x / IC_{50} / EC_{50})^{\text{Hill Slope}}}$$

where y is the percentage activity/percentage inhibition and x is the corresponding concentration. The IC_{50}/EC_{50} given in the equation is not the absolute IC_{50}/EC_{50}, but, relative IC_{50}/EC_{50}. Relative IC_{50}/EC_{50} is the concentration giving a response half way between the fitted top and bottom of the curve. The relative IC_{50}/EC_{50} serves the purpose for most of the assays.

Bioassays with a quantitative response showing a sigmoid log-dose relationship can be analysed by fitting a non-linear dose-response model directly to the data (Vølund, 1978). If the quantitative response shows a non-normal distribution, a five-parameter logistic (5PL) function is more ideal to fit dose-response data. The 5PL can dramatically improve the accuracy of asymmetric assays (Gottschalk and Dunn, 2005).

Usually, several concentrations of the compound are employed for the determination of IC_{50}. Turner and Charlton (2005) proposed a method for determining IC_{50} using two concentrations. However, use of this method is not well accepted in drug discovery research.

Hormesis

All along we have been discussing about 'threshold dose-response curve'. It is widely believed that to initiate a biological effect some dose is required. This dose is called as the threshold dose. According to this belief a dose below the threshold dose level cannot initiate the effect. This concept has been disproven in recent years by introducing a hypothesis called 'hormesis'. The term hormesis was coined by Southam and Ehrlich

(1943). The hormesis hypothesis states that most of the chemical agents may stimulate or inhibit biological effects at doses lower than a threshold, while they are toxic at doses higher than the threshold. This hypothesis falls in line with Arndt-Schulz Law, which states that 'a weak stimulus increases physiologic activity, a moderate stimulus inhibits activity and a very strong stimulus abolish the activity (Schulz,1887). However, Arndt-Schulz Law is not widely known among the toxicologists and pharmacologists. One of the reasons for this is it was heavily criticised by earlier pharmacologists and toxicologists, hence did not find place in most books on toxicology and pharmacology. Alfred Clark, the renowned pharmacologist, in his book entitled 'The Mode of Action of Drugs on Cells' published in 1933 stated: "In 1885 Rudolf Arndt put forward the suggestion that if a weak stimulus excites an organism, then any drug in sufficiently weak dose ought to do this also. This suggestion was developed by Schulz, who had leanings to homeopathy" (Clark, 1933). Clark was well known among the statisticians like Fisher and Bliss, who contributed significantly to the threshold dose-response relationship. Another book by Clark, 'Handbook of Experimental Pharmacology" (Clark, 1937), which was very critical of the Arndt-Schulz Law, was published in seven editions, in 1970s, more than 30 years after his death. Holmstedt and Lijestrand in their book, 'Readings in Pharmacology, published in 1981 stated that Homoeopathic theories like the Ardnt-Schulz law and Weber-Fechner law were based on loose ideas around surface tension of the cell membranes but there was little physic-chemical basis to these ideas (Holmstedt and Lijestrand, 1981).

Brain-Cousens (1989) proposed a modified four-parameter logistic model in situations where hormesis is present. Several publications indicated that the hormetic dose-response is far more common and fundamental than the threshold dose-response models used in toxicology (Calabrese, 2005). According to Calabrese (2010), the hormetic dose-response model makes far more accurate predictions of responses in low dose zones than either the threshold or linear at low dose models.

References

Bliss, C.I. (1934): The method of Probits. Science, 79(2037), 38–39.

Brain, P. and Cousens, R. (1989): An equation to describe dose responses where there is stimulation of growth at low dose. Weed Res., 29, 93–96.

Bretz, F. (2006): An extension of the Williams trend test to general unbalanced linear models. Comp. Stat. Data Anal., 50(7), 1735–1748.

Calabrese, E.J. (2005): Toxicological awakenings: the rebirth of hormesis as a central pillar of toxicology. Toxicol. Appl. Pharmacol., 206(3):365–366.

Calabrese, E.J. (2010): Hormesis is central to toxicology, pharmacology and risk assessment. Hum. Exp. Tox., 29(4), 249–261.

Christensen, F.M., Andersen, O., Duijm, N.J. and Harremoës, P. (2003): Risk terminology—a platform for common understanding and better communication. J. Hazardous Materials, A103, 81–203.

Clark, A.J. (1933): The Mode of Action of Drugs on Cells. The Williams & Wilkins Company, Baltimore, USA.

Clark, A.J. (1937): Handbook of Experimental Pharmacology. Springer, Berlin, Germany.

DePass, L.R. (1989): Alternative approaches in median lethality (LD_{50}) and acute toxicity testing. Toxicol. Lett., 49(2-3), 159–170.

Dorato, M.A. and Engelhardt, J.A. (2005): The no-observed-adverse-effect-level in drug safety evaluations: Use, issues, and definition(s). Reg. Toxicol. Pharmacol., 42(3), 265–274.

Eaton, D.L. and Klaassen, D. (1996): Principles of Toxicology. In: Casarett and Doull's Toxicology; The Basic Science of Poisons, 5th Edition, McGraw-Hill, New York, USA.

EMEA (2006): European Medicines Agency. Note for Guidance on Dose Response Information to Support Drug Registration (CPMP/ICH/378/95). ICH Topic E4 Dose Response Information to Support Drug Registration. EMEA, London, UK.

EMEA (2007): European Medicines Agency. Guideline on Requirements for First-in-man Clinical Trials for Potential High-risk Medicinal Products. EMEA/CHMP/SWP/28367/2007. Committee for Medicinal Products for Human Use, EMEA, London, UK.

EPA (1998): United States Environmental Protection Agency. Methods for Exposure-Response Analysis for Acute Inhalation Exposure to Chemicals: Development of the Acute Reference Exposure (ARE). EPA/600/R-98/051. External Review Draft. April 1998. USEPA, Washington, DC., USA.

EPA SAB (1998):United States Environmental Protection Agency Science Advisory Board. A SAB Report: Development of the Acute Reference Exposure: Review of the Draft Document Methods for Exposure-Response Analysis for Acute Inhalation Exposure to Chemicals: Development of the Acute Reference Exposure (EPA/600/R-98/051) by the Environmental Health Committee of the Science Advisory Board (SAB), EPA-SAB-EHC-99-005. US EPA SAB. November 1998. USEPA, Washington, DC., USA.

FDA (2005): Food and Drug Administration. Guidance for industry—Estimating the Maximum Safe Starting Dose in Initial Clinical Trials for Therapeutics in Adult Healthy Volunteers. Centre for Drug Evaluation and Research, Food and Drug Administration, USFDA, Rockville, USA.

Finney, D.J. (1952): Probit Analysis. Cambridge University Press, Cambridge, UK.

Finney, D.J. (1971): Probit Analysis. 3rd Edition. Cambridge, London, UK.

Finney, D.J. (1978): Statistical Method in Biological Assay. 3rd Edition. Charles Griffin & Co., London, UK.

Gottschalk, P.G. and Dunn, J.R. (2005): The five-parameter logistic: A characterization and comparison with the four-parameter logistic. Anal. Biochem., 343, 54–65.

Greenberg, B.G. (1980): Chester I. Bliss, 1899–1979. International Statistical Review/ Revue Internationale de Statistique, 8(1), 135–136.

Gupta, R.C. (2007): Veterinary Toxicology—Basic and Clinical Principles. Academic Press, New York, USA.

Hayes, W.J. (1991): Dosage and other factors influencing toxicology. In: Hayes, W.J. & Laws, E.R. (Editors). Handbook of Toxicology, Vol. 1, General Principles. Academic Press, San Diego, USA.

Healy, M.J.R. (1972): Statistical analysis of radioimmunoassay data. Biochem. J., 130, 107–210.

Hodgson, E. (2010): A Text Book of Modern Toxicology. John Wiley & Sons Inc., New Jersey, USA.

Holmstedt, B. and Lijestrand, G. (1981): Readings in Pharmacology. Raven Press, New York, USA.

IPCS, (2009): International Programme for Chemical Safety. Principles and Methods for the Risk Assessment of Chemicals in Food, Chapter 5, Dose-response Assessment and Derivation of Health-based Guidance Values. Environmental Health Criteria 240, IPCS, Geneva, Switzerland.

Kobayashi, K., Pillai, K.S., Michael, M., Cherian, K.M. and Ohnishi, M. (2010): Determining NOEL/NOAEL in repeat-dose toxicity studies, when the low dose group shows significant difference in quantitative data. Lab. Animal Res., 26(2), 133-137.

Kimmel, C.A. and Gaylor, D.W. (1988): Issues in qualitative and quantitative risk analysis for development toxicology. Risk Anal., 8, 15–20.

Muhammad, F., Khan, A. and Ahmad, A. (1990): Logistic regression analyses in dose response studies. J. Islamic Acad. Sci., 3:2, 103–106.

Neus, H. and Boikat, U. (2000): Evaluation of traffic noise-related cardiovascular risk. Noise & Health, 2(7), 65–77.

OECD (1995): Organization for Economic Cooperation and Development. OECD Guidelines for Testing of Chemicals. Repeated Dose 28-Day Oral Toxicity Study in Rodents, No. 407. OECD, Paris, France.

OECD (2000a): Organization for Economic Cooperation and Development. Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies. OECD Environment, Health and Safety Publications Series on Pesticides, No. 10. OECD, Paris, France.

OECD (2000b). Organization for Economic Development and Co-operation. Acute Oral Toxicity: Up-and-Down Procedure. OECD Guideline for the Testing of Chemicals, Revised Draft Guideline, No. 425. OECD, Paris, France.

Rodbard, D., Munson, P.J. and DeLean, A. (1978): Improved curve fitting, parallelism testing, characterization of sensitivity and specificity, validation, and optimization for radioimmunoassays 1977. Radioimmunoassay and Related Procedures in Medicine 1, Vienna, Italy. Int. Atomic Energy Agency (1978) 469–504.

Sand, S., von Rosen, D., Victorin, K. and Filipsson, A.F. (2006): Identification of a critical dose level for risk assessment: Developments in Benchmark dose analysis of continuous endpoints. Tox. Sci., 90(1), 244–251.

Schulz, H. (1887): Zur lehre von der arzneiwirdung. Virchows Archiv fur Pathol. Anatom. und Physiol. fur Klinische Medizin,108, 423–445.

Strickland, J.A. (2000): CatReg Software User Manual. US Environmental Protection Agency. EPA 600/R-98/052.

Southam, C.M. and Ehrlich, J. (1943): Effects of extracts of western red-cedar heartwood on certain wood-decaying fungi in culture. Phytopath., 33, 517–524.

Trevan, J. (1927): The error of determination of toxicity. Proc. R. Soc., 101B, 483−514.

Turner, R.J and Charlton, S.J. (2005): Assessing the minimum number of data points required for accurate IC_{50} determination. Assay Drug Devp. Technol., 3(5), 525–531.

Vølund, A. (1978): Application of the four-parameter logistic model to bioassay: comparison with slope ratio and parallel line models. Biometrics, 34(3), 357–365.

WHO (1994): World Health Organisation. Assessing Human Health Risks of Chemicals: Derivation of Guidance Values for Health Based Exposure Limits. Environmental Health Criteria 170, WHO, Geneva, Switzerland.

WHO (2007): World Health Organisation. Evaluation of Certain Food Additives and Contaminants. Sixty-eighth Report of the Joint FAO/WHO Expert Committee on Food Additives. WHO Technical Report Series No. 947, WHO, Geneva, Switzerland.

17

Analysis of Pathology Data

Pathology in Toxicology

Pathology occupies a pivotal role in animal experiments. The toxicity of a compound can be assessed by linking compound-related changes in biochemical, haematological or urinalysis parameters with organ weight, gross pathology and/or histopathological changes (Tyson and Sawhney, 1985; Krinke *et al*., 1991). All regulatory guidelines on animal experiments have given special emphasis to pathology. For example, in the long-term repeated dose administration studies, it is a regulatory requirement that all data relating to moribund or dead animals as well as the results of postmortem examinations is scrutinized and the analysis of the cause of individual deaths is done (OECD, 2000).

Pathologists usually make a biological judgment based on their experience, which differs from one pathologist to the other (Glaister, 1986). In a repeated dose administration study involving a large number of animals, the observation of tissue section slides may be completed over a substantial length of time. Thus it is not possible to maintain the consistency of grading the lesions, causing a 'diagnostic drift'. It has been stated that even the nomenclature used to describe pathology findings in toxicology studies suffers from the lack of uniformity. Use of different nomenclature for describing the lesions causes difficulties while interpreting the observations (Haseman *et al*., l984). Statistically and logically, blinding the slides is the best way to avoid the bias. But, several veterinary pathologists do not favor this, because they fear that blinded reading of slides of animal tissues/organs may result in loss of information critical to interpretation, such as the ability to relate

observations in different tissues (Iatropoulos, 1984; Newberne and de la Lglesia, 1985; Prasse *et al*., 1986; Goodman, 1988; House *et al*., 1992; FDA, 2001). Mistakes can be easily made when assigning, opening codes, and recording results in blinded reading (Iatropoulos, 1988).

Microscopical data obtained from toxicity studies is usually classified into several grades. The grades of the control group is usually shown by minus (–) and those of the treated groups by (+1), (+2), (+3), so on. For statistical analysis, the difference of the grades between the control group and treatment groups is examined by Fisher's probability test or cumulative chi-square test. By these methods, only the presence or absence of a difference among several groups or between two groups can be ascertained and the degree of pathology lesions remains uncertain.

There is not enough specific statistical guidance available for the pathologists. Wade (2005) stated that most of the published statistical literature is not directly applicable to research in the field of pathology. In the toxicology studies with three or more groups, the relationship between the findings and the dose dependency should be examined. Dose dependency is often examined by the Cochran-Armitage trend test after Fisher's probability test or chi-square test. Kobayashi and Pillai (2003) proposed a method to examine both the degree of pathology lesions and the dose dependency. In this method, the pathology findings are scored in grades and analyzed by the rank sum test. For comparison between two groups, Mann-Whitney's or Wilcoxon's test, and for comparison among several groups, Dunnett's, Tukey's, Duncan's, Scheffe's, Wilcoxon's or Williams-Wilcoxon's non-parametric tests are proposed. However, the number of animals necessary to detect a significant difference between the low dose group and the control group greatly varies with these tests. Dunnett's multiple comparison test can detect a significant difference even with four animals per group when the dose dependency is very high. The authors suggested Jonckheere's trend test and Spearman's correlation coefficient (r) for examination of dose-dependency.

Analysis of Pathology Data of Carcinogenicity Studies

The objectives to be achieved as per the guidelines of OECD (2009) for rodent carcinogenicity studies are hazard characterization, describing the dose-response relationship and the derivation of an estimate of a point of departure such as the Benchmark dose or a no observed adverse effect level. Normally, carcinogenicity studies are conducted in rodents with

a control group and 2 or 3 treatment groups, each group containing a minimum of 50 animals of each gender. Mice are normally exposed to the test compound for 18–24 months, whereas rats are exposed for 24–30 months. Animals are sacrificed at intervals or at the end of the experiment. The major observations carried out in a carcinogenicity study are the survival time and status (presence/absence) of specific tumour types.

National Toxicology Programme (NTP) and U.S. Food and Drug Administration (US FDA), reported that there were issues in the application of statistical methods to carcinogenicity studies (Gad and Rousseaux, 2002). Tumour incidence (tumour incidence is defined as the rate of tumour onset among the tumour-free population) is considered the most appropriate measure of tumourigenesis (Malani and Van Ryzin, 1988; Dinse, 1994). Tumours can be classified as 'incidental,' 'fatal,' and 'mortality-independent (or observable)' according to the contexts of observation described by Peto *et al*. (1980). Tumours that are not directly or indirectly responsible for the animal's death, but are merely seen at the autopsy of the animal after it has died of an unrelated cause, are said to have been observed in an incidental context. Tumours that kill the animal, either directly or indirectly, are said to have been observed in a fatal context. Tumours, such as skin tumours, whose detection occurs at times other than when the animal dies are said to have been observed in a mortality-independent context (Lin, 2000). Benign and malignant tumours should be analysed separately (Mc Connell *et al.,* 1986; EPA, 2005), if it is considered scientifically defensible, further statistical analysis may be performed on the combined benign and malignant tumours of the same histogenic origin, even when those tumours are in different tissues.

Peto test

While most pharmaceutical companies use the Peto test (Peto *et al*., 1980), some do not categorize neoplasms as fatal or incidental. Generally, this test is considered to be useful for the groups with different survival rates. Before analysing, pathological findings should be examined (whether malignant of benign) and conclude whether the drug caused the death or not. Some categorize neoplasms as fatal or incidental based solely on the type of neoplasm rather than on an animal-by-animal basis. Others categorize neoplasms as fatal or incidental based on the gross and microscopic findings for each animal. Some controversies exist when relying on the Peto test for information on 'cause of death' (STP, 2002).

According to Lee *et al*. (2002), the 'fatal' definition is often misunderstood by the pathologists and there is a tendency for the over-designation of fatal tumours (Kodell *et al*., 1982; Ahn *et al*., 2000).

The US FDA recommends that both trend test and pair-wise comparison test be performed routinely for each study and that the results of both tests should be presented to regulatory officials (FDA, 2001). However, the Peto test is required for product registration in Europe. Based on current regulatory requirements, the STP recommends that the Peto test should be performed whenever the study pathologist and the peer review pathologist can consistently classify neoplasms as fatal or incidental (Morton *et al*., 2002).

Decision rules

A distinguished characteristic of the Peto test is that it involves dosages in the calculation procedure. The power of the Peto test is very high, when the significance level is set at 5% probability level. However, the use of significance set at 5% and 1% probability levels in tests for positive trend in incidence rates of rare tumours and common tumours, respectively, will result in an overall false positive rate around 10% in a study in which only one 2-year rodent bioassay (plus the shorter rodent study) is conducted (Lin, 1998; Lin and Rahman, 1998). The power to detect a significant difference is greater with the trend tests than with the pair-wise comparisons in an animal experiment with a control group and more than two treatment groups. There are situations in which pair-wise comparisons between control and individual treated groups may be more appropriate than trend tests. However, both trend and pair-wise comparison tests are likely to cause false positive results. In order to control overall positive rates associated with trend tests and pair-wise comparisons certain statistical decision rules were developed (Haseman,1983). The decision rules were developed based on historical control data of Crl: CDÒ BR rats and Crl: CD-1Ò (ICR) BR mice to achieve an overall false positive rate of around 10% for the standard *in vivo* carcinogenicity studies in rodents. The decision rule tests the significance difference in tumour incidences between the control and the treatment groups at 5% probability level for rare tumours (tumours with background rate of 1% or less) and at 1% probability level for common tumours (frequent tumours). However, the decision rule described by Haseman (1983) to analyse the trend tests would lead to an excessive overall false positive error rate about twice as large as that associated with control-high dose pair-wise comparison

tests. Statistical decision rules for controlling the overall false positive rates associated with tests for positive trend or with control *vs* high dose pair-wise comparison in tumour incidences in carcinogenicity studies were reported by FDA (2001). These decision rules test positive trend in tumour incidence at 2.5% probability level for rare tumours and at 0.5% probability level for common tumours. Although the overall false positive rate resulting from the use of the decision rule may vary from study to study, it is estimated that it will be around 10%.

The decision rules for testing positive trend or differences between control and individual treatment groups in incidence rates of tumours for standard studies using two species and two sexes as well as studies following ICH guidance and using only one 2-year rodent bioassay are summarized in Table 17.1.

Table 17.1. Statistical decision rules for controlling the overall false positive rates associated with tests for positive trend or with control *vs* high dose pair-wise comparisons in tumour incidences to around 10 percent in carcinogenicity studies of pharmaceuticals (FDA, 2001).

Study	Tests for positive trend	Control *vs* high dose pair-wise comparison
Standard 2-year studies with 2 species and 2 sexes	Common and rare tumours are tested at 0.5% and 2.5% probability levels, respectively	Common and rare tumours are tested at 1% and 5% probability levels, respectively
Alternative ICH studies (one two-year study in one species and one short- or medium-term study, two sexes)	Common and rare tumours are tested at 1% and 5% probability levels, respectively	Under development and not yet available.

Note: The decision rules were developed assuming the use of two-species and two-sex (or one-species and two-sex) for the standard design of a two-year study with 50 animals in each of the four treatment/sex/group.

Poly-k Type test

An alternative to the Peto-type is Poly-*k* type test (Bailer and Portier, 1988; Portier and Bailer, 1989; Piegorsch and Bailer, 1997). One advantage of this test is that it does not require the controversial ‘cause of death’ in the calculation procedure. NTP uses the Poly-*k* test to assess neoplasm and non-neoplastic lesion prevalence.

Analysis of Tumour Incidence—Comparison with Historical Control Data

Tumour incidence between the treatment group and control group is normally compared using Fisher's probability test. By this test, no significant difference in tumour incidence is observed between the treatment group and control group, if the incidence of tumour is 0/50 (number of animals in the group having tumour/total number of animals in the group) in the control group and 4/50 in the treatment group. However, a tumour incidence of 4/50 is considered to be significant from a pathological viewpoint. Comparison of the incidence of tumour in the treatment group with that of the historical control data may be useful, especially to assess the occurrence of rare tumours and marginally increased tumour incidences. But, certain requirements must be met before the use of historical control data, since the historical control data may change in time (Greim *et al*., 2003). Several procedures have been proposed for incorporating historical control data into the analysis of data obtained from carcinogenicity studies (Sun, 1999). If the data of the treatment group is compared with the historical control data using *t*-test, it should be remembered that the number of animals used in these groups is different, being much larger in the historical control group, since the source of historical control data is several studies. Table 17.2 shows a comparison of incidence of tumour observed in 50 animals in the treatment group with several historical controls having differences in number of animals but with similar tumour incidence (%).

Table 17.2. Comparison of treatment group with historical control data using Kastenbaum and Bowman test (Kastenbaum and Bowman, 1966)

Incidence of tumour (Historical control data[a])	Incidence of tumour in 50 animals (Treatment group)			
	1 (2%)	2 (4%)	3 (6%)	4 (8%)
1/ 200 (0.5%)	NS	NS	NS	NS
2/ 500 (0.4%)	NS	NS	NS	*
3/ 700 (0.4%)	NS	NS	NS	**
4/1000 (0.4%)	NS	NS	*	**
5/1250 (0.4%)	NS	NS	*	**
7/1500 (0.5%)	NS	NS	*	**
7/1700 (0.4%)	NS	NS	*	**
8/2000 (0.4%)	NS	NS	*	**
10/2500 (0.4%)	NS	NS	*	**

[a]Number of animals in the historical controls showing tumour/total number of animals in the historical controls; NS-Not significance, *P<0.05, **P<0.01.

The incidence of tumour in 1 or 2 animals out of 50 animals in the treatment group is not significantly different compared with the historical control animals showing the tumour incidence in 1 animal out of 200 or 10 out of 2500 animals. However, the incidence of tumour seen in 3 animals out of 50 animals in the treatment group is significantly different from the historical control animals with incidences of tumour as 4/1000, 5/1250, 7/1500, 7/1700, 8/2000 and 10/2500 (number of animals showing incidence of tumour/total number of animals). The incidence of tumour 8% (4/50) in the treatment group is significantly different from the historical control data showing the incidence of tumour as 2/500, 3/700, 4/1000, 5/1250, 7/1500, 7/1700, 8/2000 and 10/2500. It is obvious from the Table 17.2 that the number of animals used in constructing the historical control data plays a crucial role in determining a significant difference between the historical control data and the treatment group.

The circumstances that prompted the use of historical control for the analysis of carcinogenicity data should be properly explained and justified. It must be remembered that the concurrent control group is the most relevant comparator for determining treatment-related effects in a study (FDA, 2001; EMEA, 2002; OECD, 2002). In evaluating the data from historical controls, statistically significant increases in tumours based on the concurrent control should not be discounted simply because incidence rates in the treatment groups are within the range of historical controls or because incidence rates in the concurrent controls are low (Keenan *et al*., 2009). OECD guidelines (OECD, 2002) emphazise the historical control data should be generated by the same laboratory in animals of contemporaneous studies in the same species and strain, maintained under similar conditions, at which the study being assessed was performed. Furthermore, the historical control data should come from studies conducted within five years prior to, or within two to three years from the conclusion of the study. The guidelines recommend parameters that could affect the occurrence of spontaneous tumours in historical control data are identified. In studies exhibiting the lowest incidence (less than a few percent) of tumours, the Kastenbaum and Bowman test appears to be more relevant, since it takes into account the sample size of both the historical control data base and each treatment group in the study. In studies where a wider range of tumour incidence is exhibited, a statistical method which employs a rejection limits based on the range of incidence in the historical data is recommended. When malignant tumours are evident in treatment groups, no matter how low the incidence, the tumour should be analyzed

statistically and compared with the incidence in the historical control data as well as those in the concurrent control group (Kobayashi and Inoue, 1994).

Analysis of Incidence of Tumour Using X^2 Test

Chi square test is an excellent tool to evaluate the significant difference in occurrence of tumours among the groups. An example is given in Table 17.3.

Table 17.3. Total number of occurrence of tumours in different organs in a two-year carcinogenicity study

Control	Low dose	Mid dose	High dose	Total
58	50	62	65	235

Note: Each group consists of 50 animals.

$$X^2 = \frac{58^2}{235 \times 0.25} + \frac{50^2}{235 \times 0.25} + \frac{62^2}{235 \times 0.25} + \frac{65^2}{235 \times 0.25} - 235 = 2.157$$

Note: 0.25=1/4: Assumed probability distribution (4=Number of groups).

The chi-squared Table value for 3 degrees of freedom is 7.82 at 5% probability level. The calculated value 2.157 is less than 7.82, which means that there is no significant differences in the occurrence of tumours among the groups. If a significant difference is observed, difference between control and each group is analyzed.

However, use of chi-square goodness-of-fit in multistage model to carcinogenicity has been questioned in recent years. According to Sielken (1988) "although the chi-square goodness-of-fit is a very widely used statistical test, it is also well documented (though not sufficiently widely known) that the test can have very little power to reject inaccurate models".

Comparison of Incidence of Tumours in Human, Rats, Mice and Dogs

Considerable debate about the need of conducting carcinogenicity studies in rats and mice has been taken place in recent years (Ennever and Lave, 2003; Billington *et al*., 2010; Storer *et al*., 2010). Most of the scientists are of the opinion that there is no need to conduct long-term rodent carcinogenicity studies in mice, since the use of the mice in carcinogenicity testing does not provide useful scientific information (Griffiths *et al*., 1994;

Carmichael *et al*., 1997; Meyer, 2003; Doe *et al*., 2006). However, some current regulatory programmes require carcinogenicity testing in rats and mice.

Kobayashi *et al*. (1999) made an interesting comparison of incidence of spontaneous malignant tumours in human, rats, mice and dogs. The prevalence of each carcinoma in rodents was calculated as the population ratio *P*, at a 95% confidence interval, and compared with that in humans. The primary carcinomas according to sex in Japanese people who died of cancer were cited from the report of investigations on the population dynamics and economy in 1992, "Malignant neoplasm" published by the Welfare Statistics Association, Japan (Ministers' Secretariat, 1994). Data on spontaneous incidence of tumours in rats, mice and dogs were obtained from Biosafety Research Centre—Foods, Drugs and Pesticides, Japan. The incidence of spontaneous malignant tumours of various organs in humans, rodents and dogs is shown in Table 17.4.

Table 17.4. Incidence (%) of spontaneous malignant tumours in dead humans, rodents and dogs

Organ	Male			Female			Male+Female
	Human	Rat	Mouse	Human	Rat	Mouse	Dog
No. of deaths with cancer	139674	105	120	92243	117	100	5845
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Esophagus	4.7	0	0	1.4	0	0	0.3
Stomach	21.8	1.0	0	19.0	0.9	1.0	0.3
Intestine[a]	4.4	0	0.89	4.3	0	0	1.0
Liver	14.0	0.	52.5	8.1	1.7	24.0	0.7
Pancreas	5.6	0	0	6.9	0.9	2.0	0.5
Lung, trachea, bronchi	20.9	2.9	5.0	11.9	0	1.0	0.6
Mammary gland	<0.1	0	0.	7.1	2.6	8.0	9.1
Uterus	-	-	-	5.1	10.3	11.0	0.3
Leukemia	2.4	53.3	20.8	2.69	59.8	31.0	4.3
Other	26.1	42.9	20.8	33.9	23.9	22.0	82.9

[a]Including colon and anus in humans, small intestine, duodenum, large intestine and colon in rodents, and colon in dogs.

The incidence of tumours in the organs of humans who died of cancer differed considerably from that of mice, rats and dogs. For example, very low or no incidence of tumour was seen in esophagus, stomach and intestine of rats and mice. The incidence of hepatocellular carcinoma in mice and leukemia in rats and mice were higher than those in humans,

while the incidence of malignant tumours in the lungs of rodents was lower than those in humans. The authors stated it is important to consider the spontaneous tumours and the probable target organ when selecting the appropriate species for a carcinogenicity study.

Analysis of Organ Weight Data

Organ weights (absolute and relative organ weights) are an important quantitative end point in the repeated dose administration studies. Many pathologists are of the opinion that it would be better to calculate organ weight relative to brain weight (organ-to-brain weight ratio) rather than to body weight (organ-to-body weight ratio). Animals are usually fasted before necropsy. The deprivation of food can affect the body weight of the animals, and also the physiological adaptability to fasting may vary significantly among the animals. When the body weight gain is affected, alterations of organ weight/body weight ratio may be due to the physiological response of the animal to decreased nutrient intake. Organ-to-body weight ratios are preferable for analysis of liver and thyroid weights, whereas organ-to-brain weight ratios are best for analysis of ovary and adrenal weights, and both organ-to-body weight ratios and organ-to-brain weight ratios do not accurately model brain, heart, kidney, pituitary, or testis weights (Bailey *et al*., 2004). Regardless of the study type or organs evaluated, organ weight changes must be evaluated within the context of the compound class, mechanism of action, and the entire data set for that study (Sellers *et al*., 2007).

Absolute weight of the mouse liver in a 13 week repeated dose administration study is given in Table 17.5.

Table 17.5. Absolute weight (g) of the mouse liver in a 13 week repeated dose administration study

Group	Control	Low Dose	Mid Dose	High Dose
Individual value	1.08, 1.09, 1.15, 1.09, 1.16, 1.00, 1.12, 1.01, 1.12, 1.02	1.09, 1.12, 1.15, 1.09, 1.04, 0.99, 1.24, 1.15, 0.99, 1.12	1.10, 1.20, 1.09, 1.02, 1.07, 1.12, 1.13, 1.06, 1.11, 1.20	1.16, 1.15, 1.24, 1.16, 1.22, 1.10, 1.18, 1.07, 1.18, 1.09
n	10	10	10	10
Mean ± SD	1.08 ± 0.06	1.10 ± 0.08	1.11± 0.06	1.16 ± 0.05
In % Control	100	102	103	107

Since there are four groups, the data is analysed using one-way ANOVA, which shows a non-significant *F* value, indicating that there is no significant difference in the absolute weight of the liver among the groups. Close examination of the mean value of the groups indicates that there is a dose-dependent increase in the absolute weight of the liver. When the data is analysed using Dunnett's multiple comparison test, absolute weight of the liver of the high dose group is found to be significantly different from the control group. It may be worth mentioning in this context that Dunnett (1964) did not recommend ANOVA prior to multiple comparison tests. Several authors are of the opinion that the error of second kind can be prevented by carrying out direct multiple comparison tests without subjecting the data to ANOVA (Hamada *et al*., 1998; Sakaki *et al*., 2000; Kobayashi *et al*., 2000).

Interpretation of Pathology Observations

Interpretations made from the organ weight data should be used with caution. Indicating a significant or non-significant difference in organ weight alone by statistical analysis, particularly in studies with small size, has little use in evaluating the organ weight changes (Sellers *et al*., 2007). According to Gad and Rousseaux (2002), treatment-related alterations in organ weight may not be statistically significant, similarly statistically significant alteration in organ weight may not be treatment related.

In the long-term toxicology studies, animals may show age-associated changes, which can have a significant effect on histopathology (Mohr *et al*., 1992, 1994, 1996). Spontaneous degenerative lesions, especially when misinterpreted as toxic effects can cause major difficulty in hazard evaluation. In these situations, the data can be compared with historical control data. It has been stated that historical control tumour data is useful in the interpretation of long-term rodent carcinogenicity bioassays, especially to assess the occurrence of rare tumours and marginally increased tumour incidences (Deschl *et al*., 2002). However, the advantage of a concurrent control as the comparator for treatment-related effects should not be overlooked, when historical control data are used as the comparator.

References

Ahn, H., Kodell, R.L. and Moon, H. (2000): Attribution of tumour lethality and estimation of time to onset of occult tumours in the absence of cause-of-death information. App. Stat., 49, 157–169.

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

Bailey, S.A., Zidell, R.H. and Perry, R.W. (2004): Relationships between organ weight and body/brain weight in the rat: what is the best analytic endpoint? Toxicol. Pathol., 32, 448–466.

Billington, R., Lewis, R., Mehta, J. and Dewhurst, I. (2010): The mouse carcinogenicity study is no longer a scientifically justifiable core data requirement for the safety assessment of pesticides. Crit. Rev.Toxicol., 40, 35–49.

Carmichael, N.G., Enzmann, H., Pate, I. and Waechter, F. (1997): The significance of mouse liver tumour formation for carcinogenic risk assessment: Results and conclusions from a survey of ten years of testing by the agrochemical industry. Environ. Health Perspect., 105, 1196–1203.

Deschl, U., Kittel, B., Rittinghausen, S., Morawietz, G., Kohler, M., Mohr, U. and Keenan, C. (2002): The value of historical control data—Scientific advantages for pathologists, industry and agencies. Toxicol. Pathol., 30, 80–87.

Dinse, G.E. (1994): A comparison of tumour incidence analyses applicable in single-sacrifice animal experiments. Stat. Med., 13, 689–708.

Doe, J.E., Boobis, A.R., Blacker, A., Dellarco, V., Doerrer, N.G., Franklin, C., Goodman, J.I., Kronenberg, J.M., Lewis, R., Mcconnell, E.E., Mercier, T., Moretto, A., Nolan, C., Padilla, S., Phang, W., Solecki, R., Tilbury, L., van Ravenzwaay, B. and Wolf, D.C. (2006): A tiered approach to systemic toxicity testing for agricultural chemical safety assessment. Cri. Rev. Toxicol., 36, 37–68.

Dunnett, C.W. (1964): New tables for multiple comparisons with a control. Biometrics, 20(3), 482–491.

EMEA (2002): European Medicines Agency. CPMP, Note for guidance on carcinogenic potential, EMEA, CPMP/SWP/2877/00, London, 25 July 2002. http://www.emea.europa.eu/pdfs/human/swp/287700en.pdf.

Ennever, F.K. and Lave, L.B. (2003): Implications of the lack of accuracy of the lifetime rodent bioassay for predicting human carcinogenicity. Reg. Tox. Pharm., 38, 52–57.

EPA (2005): United States Environmental Protection Agency. Guidelines for Carcinogen Risk Assessment. U.S. Environmental Protection Agency (USEPA), Washington DC, USA.

FDA (2001): Food and Drug Administration. Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals. Draft Guidance, US FDA, Rockville, MD, USA.

Gad, S.C. and Rousseaux, C.G. (2002): Use and misuse of statistics in the design and interpretation of studies. In: Handbook of Toxicologic Pathology, 2nd Edition. Editors, Haschek, W.M., Rousseaux, C.G. and Wallig, M.A. Academic Press, San Diego, USA.

Glaister, J.R. (1986): Principles of Toxicological Pathology. Taylor & Francis, Philadelphia, USA.

Goodman, D.G. (1988): Factors Affecting Histopathologic Interpretation of Toxicity-Carcinogenicity Studies. Carcinogenicity: The Design, Analysis, and Interpretation of Long-Term Animal Studies. ILSI Monographs, Springer-Verlag, New York, USA.

Greim, H., Gelbke, H.P., Reuter, U., Thielmann, H.W. and Elder, L. (2003): Evaluation of historical control data in carcinogenicity studies. Hum. Exp. Toxicol., 22(10), 541–549.

Griffiths, S.A., Parkinson, C., McAuslane, J.A.N. and Lumley, C.E. (1994): The utility of the second rodent species in the carcinogenicity testing of pharmaceuticals. Toxicologist, 14(1), 214.

Hamada, C., Yoshino, K., Matsumoto, K., Nomura, M. and Yoshimura, I. (1998): Three-type algorithm for statistical analysis in chronic toxicity studies. J. Toxicol. Sci., 23 (3), 173–181.

Haseman, J.K. (1983): A reexamination of false-positive rates for carcinogenesis studies. Fund. Appl. Toxicol., 3, 334–339.

Haseman, J.K., Huff, J. and Boorman, G.A. (1984): Use of historical control data in carcinogenicity studies in rodents. Toxicol. Pathol., 12, 126–135.

House, D.E., Berman, E., Seely, J.C. and Simmons, J.E. (1992): Comparison of open and blind histopathologic evaluation of hepatic lesions. Toxicol. Lett., 63, 127–133.

Iatropoulos, M.J. (1984): Editorial : Toxicol. Pathol., 12(4), 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 Long-Term Animal Studies, ILSI Monographs, Editros, Grice, H.C. and Ciminera, J.L., Spring-Verlag, New York, USA.

Kastenbaum, M.A. and Bowman, K.O. (1966): The minimum significant number of successes in a binominal sample. Oak Ridge National Laboratory (ORNL-3909), Oak, Tennessee, USA.

Keenan, C., Elmore, S., Francke-Carroll, S., Kemp, R., Kerlin, R., Peddada, S., Pletcher, J., Rinke, M., Schmidt, S.P., Taylor, I. and Wolf, D.C. (2009): Best practices for use of historical control data of proliferative rodent lesions. Toxicol. Pathol., 37, 679–693.

Kobayashi, K., Hagiwara, T., Miura, D., Ohori, K., Takeuchi, H., Kanamori, M. and Takasaki, K. (1999): A comparison of spontaneous malignant tumours in humans, rats, mice and dogs. J. Environ. Biol., 20(3), 189–193.

Kobayashi, K. and Inoue, H. (1994): Statistical analytical methods for comparing the incidence of tumours to the historical control data. J. Toxicol. Sci., 19(1), 1–6.

Kobayashi, K., Kanamori, M., Ohori, K., and Takeuchi, H. (2000): A new decision tree method for statistical analysis of quantitative data obtained in toxicity studies on rodents. San Ei Shi, 42, 125–129.

Kobayashi, K. and Pillai, K.S. (2003): Applied Statistics in Toxicology and Pharamacology, Science Publishers, Enfield, USA.

Kodell, R.L, Farmer, J.H., Gaylor, D.W. and Cameron, A.M. (1982): Influence of cause-of-death assignment on time-to-tumour analyses in animal carcinogenesis studies. J. Natl. Cancer Inst., 69, 659–664.

Krinke, G.J., Perrin, L.P.A. and Hess, R. (1991): Assessment of toxicopathological effects in ageing laboratory rodents. Arch. Toxicol. Suppl., 14, 43–49.

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

Lin, K.K. (1998): CDER/FDA formats for submission of animal carcinogenicity study data. Drug Information J., 32, 43–52.

Lin, K.K. (2000): Progress report on the guidance for industry for statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals. J. Biopharm. Stat., 10(4), 481–501.

Lin, K.K. and Rahman, M.A. (1998): False Positive Rates in Tests for Trend and Differences in Tumour incidence in Animal Carcinogenicity Studies of Pharmaceuticals under ICH Guidance S1B, Unpublished Report, Division of Biometrics 2, Center for Drug Evaluation and Research, Food and Drug Administration. USFDA, MD, USA.

Malani, H.M. and Van Ryzin, J. (1988): Comparison of two treatments in animal carcinogenicity experiments. J. Am. Stat. Assoc., 83, 1171–1177.

Mc Connell, E.E., Solleveld, H.A., Swenberg, J.A. and Boorman, G.A. (1986): Guidelines for combining neoplasms for evaluation of rodent carcinogenicity studies. J. Natl. Cancer Inst., 76, 283–289.

Meyer, O. (2003):Testing and assessment strategies, including alternative and new approaches. Toxicol. Lett., 140–141, 21–30.

Ministers' Secretariat (1994): The Report of Investigation on Population Dynamics and Social Economy in 1992 "Malignant Neoplasm", Welfare Statistics Association, Tokyo, Japan.

Mohr, U., Dungworth, D.L. and Capen, C.C. (1992): Pathobiology of the Aging Rat. Vol. 1. ILSI Press, Washington, DC, USA.

Mohr, U., Dungworth, D.L. and Capen, C.C. (1994): Pathobiology of the Aging Rat. Vol 2. ILSI Press, Washington, DC, USA.

Mohr, U., Dungworth, D.L., Ward, J., Capen, C.C., Carlton, W. and Sundberg, J. (1996): Pathobiology of the Aging Mouse. Vols. 1 & 2. ILSI Press, Washington, DC, USA.

Morton, D., Elwell, M., Fairweather, W., Fouillet, X., Keenan, K., Lin, K., Long, G., Mixson, L., Morton, D., Peters, T., Rousseaux, C. and Tuomari, D. (2002): The Society of Toxicologic Pathology's recommendations on statistical analysis of rodent carcinogenicity studies. Toxicol. Pathol., 30(3): 415–418.

Murakami, M., Yamada, M. and Yokouchi, H. (2000): Statistical method appropriate for general toxicological studies in rats. J. Toxicol. Sci., 25, 71–98.

Newberne, P.M. and de la Lglesia, F.A. (1985): Editorial: Philosophy of blind slide reading. Toxicol. Pathol., 13(4), 255.

OECD (2000): Organisation for Economic Cooperation and Development. Environment Directorate Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology. Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies. OECD Series on Testing and Assessment, Number 32 and OECD Series on Pesticides, Number 10, ENV/JM/MONO(2000)18, Paris, France.

OECD (2002): Organisation for Economic Cooperation and Development. Environment, Health and Safety Publications. Series on Testing and Assessment No. 35 and Series

on Pesticides No. 14. Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies. ENV/JM/MONO 19, Paris, France.

OECD (2009): Organization for Economic Cooperation and Development. Guidelines for Testing of Chemicals. Carcinogenicity Studies. Test Guideline 451. OECD, Paris, France.

Peto, R., Pike, M., Day, N.E., Gray, R.G., Lee, P.N., Parish, S., Peto, J., Richards, S. and Wahrendorf, J. (1980): Guidelines for Simple, Sensitive Significance Tests for Carcinogenic Effects in Long-term Animal Experiments. In: IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Humans, Supplement of Long-term and Short-term Screening Assays for Carcinogens: A Critical Appraisal. International Agency for Research on Cancer, Lyon, France.

Piegorsch, W.W. and Bailer, A.J. (1997): Statistics for Environmental Biology and Toxicology, Chapman and Hall, London, UK.

Portier, C.J. and Bailer, A.J. (1989): Testing for increased carcinogenicity using a survival-adjusted quantal response test. Fundam. Appl. Toxicol.,12, 731–737.

Prasse, K., Hildebrandt, P. and Dodd, D. (1986): Letter to the Editor: Vet. Pathol., 23, 540–541.

Sakaki, H., Igarashi, S., Ikeda, T., Imamizo, K., Omichi, T., Kadota, M., Kawaguchi, T., Takizawa, T., Tsukamoto, O., Terai, K., Tozuka, K., Hirata, J., Handa, J., Mizuma, H., Murakami, M., Yamada, M. and Yokouchi, H. (2000): Statistical method appropriate for general toxicological studies in rats. J. Toxicol. Sci., 25, 71–98.

Sellers, R.S., Mortan, D., Michael, B., Roome, N., Johnson, J.K., Yano, B.L., Perry, R. and Schafer, K. (2007): Society of Toxicologic Pathology Position Paper: Organ weight recommendations for toxicology studies. Toxicol. Pathol., 35(5), 751–755.

Sielken, R.L. (1988): A critical evaluation of dose-response assessment of TCDD. Food Chem. Toxicol., 26(1), 79–83.

Storer, R.D., Sistare, F.D., Reddy, M.V. and DeGeorge, J.J. (2010): An industry perspective on the utility of short-term carcinogenicity testing in transgenic mice in pharmaceutical development. Toxicol. Pathol., 38, 51–61.

STP (2002): STP Peto Analysis Working Group (2002). The Society of Toxicological Pathology's recommendations on rodent carcinogenicity studies. Toxicol. Pathol., 30, 415–418.

Sun, J. (1999): On the use of historical control data for trend test in carcinogenicity studies. Biometrics, 55, 1273–1276.

Tyson, C.A. and Sawhney, D.S. (1985): Organ Function Tests in Toxicology Evaluation. Noyes Publications, Park Ridge, New Jersey, USA.

Wade, A. (2005): Fear or favour? Statistics in pathology. J. Clin. Pathol., 53,16–18.

18

Designing An Animal Experiment in Pharmacology and Toxicology—Randomization, Determining Sample Size

Designing Animal Experiments

The use of animals raises scientific and ethical challenges (Workman *et al*., 2010). Therefore, an animal experiment should be designed with due consideration to ethics on a solid scientific platform. Animal experiment should have high precision, but should not waste resources or animals (Festing, 1997). It is important to select an appropriate study design to provide scientific evaluation of the research findings without bias (Lim and Hoffmann, 2007). Replication, randomization and blinding are the key components of the design of the animal experiment. But, these are less often used in animal research (Kilkenny *et al*., 2009). Hess (2011) reviewed statistical design given in 100 articles on animal experiments published in Cancer Research in 2010. In 14 of the 100 articles, the number of animals used per group was not reported. In none of the 100 articles was the method employed to determine the number of animals used per group reported. Among the 74 articles in which randomization seemed feasible, only 21 reported that they had randomly allocated animals to various groups. None of these articles described how the randomization was carried out.

In animal experiments, bias could arise from lack of randomization, not blinding the groups, failure to report excluded animals, small sample sizes or use of statistical tools with low power (Dirnagl and Macleod, 2009). If there is a large difference between the treatment group and control of a well designed study, an experienced analyser can draw a conclusion without

carrying out a statistical analysis of the data. But, if the difference is marginal, a mistaken or a biased conclusion could be avoided by subjecting the data to the statistical analysis (Lew, 2007).

It has been stated that several reports on animal experiments were biased or did not correctly model human disease and therefore were of little utility (Festing, 2003; Perel *et al*., 2007). Though the findings of most of the animals studies cannot be directly extrapolated to man, a properly designed study may provide vital information on efficacy and toxicity of the test substance. Acclimation and randomization procedures of animals, and rationale for fixing the number of animals in a group should be explained in the study plan. There are additional issues such as rationale for selection of species, animal house conditions, bedding material, diet, drinking water, *etc.*, which need to be considered in the study plan, but beyond the scope of this book.

Acclimation

It should be ensured that the animals are not stressed at the start of the experiment. One way to ensure this is by acclimating the animals to the laboratory conditions. The acclimation period can be used for health-related quarantine and monitoring, and for behavioral conditioning. This period may include habituation to, desensitization to, and training for procedures that will be involved in experimental use (Bloomsmith *et al*., 2006). Well-acclimated animals are able to deal appropriately with the challenges of the experimental environment. This ability is typically manifested in a transient divergence from equilibrium in response to a manipulation, followed by a gradual return to homeostatic balance (Schapiro and Everitt, 2006). Animals appearing to be behaviorally acclimated to a procedure may not necessarily physiologically acclimated to that procedure (Capitanio *et al*., 2006). For example, acclimated animals may sometimes show change in metabolic profiles. Changes in nuclear magnetic resonance spectroscopic-based urinary metabolite profiles were observed in germ-free rats acclimated in standard laboratory animal facility conditions (Nicholls *et al*. 2003).

Randomization

Appropriate randomization and statistical procedures in the design of animal experimentation provide confidence that statistically significant results are

not due to chance (EPA, 2005). Selection of an appropriate statistical tool is heavily depended on randomization, which is a fundamental element of good statistical design that acts to reduce potential bias during treatment allocation (Festing and Altman, 2002).

Infact the concept of randomization originated as early as 1935 (Fisher, 1935). Randomization transforms systematic errors into random errors and confirms comparability among experimental groups (Hamada and Ono, 2000). Though randomization is an important aspect in designing animal experiments, little consideration is given to it in most cases. This is evident from different terminologies that are used for randomization, like "animals were divided into four groups"; "animals were randomly divided"; "animals were sorted into groups"; "animals were randomly assigned"; and, "half of the animals were placed into one group and the other half in a second group" (Kozinetz, 2011). The key deficiencies that are seen in animal experiments are failure to randomly allocate animals to treatments and failure to blind observers to treatment assignment during outcome assessments (Hess, 2011). Failure of NXY-059, a neuroprotective agent for stroke patients, of Astra Zeneca, in Phase III has been attributed to improper randomization and bias in preclinical studies (Savitz, 2007). When comparing two treatments, analyser-related bias may occur. This bias can be avoided by blinding (Aguilar-Nascimento, 2005). In a clinical trial, blinding can take place at three levels: study units, researcher and data (Lim and Hoffmann, 2007). The same method can be applied to animal experiments also. In a blinded study, the researcher does not know which group of animals receives what treatment. According to Bebarta *et al*. (2003), "animal experiments where randomization and blind testing are not reported are five times more likely to report positive results". Therefore, effects of randomization have to be considered in planning and performing experiments as well as in the interpretation of experimental results (Vogt and Kloting, 1990).

In toxicological experiments, especially in repeated dose administration studies, young adult animals of an inbred strain are used. Though the animals of inbred strain are supposed to be isogenic, in reality it is not so. There could be some genetic variation between the individuals from one litter and the other. Let us work out an example. Body weight of rats from 3 litters is given in Table 18.1.

Let us randomly distribute the animals of litters 1, 2 and 3 into three groups. An unbiased randomization should distribute the variation of

Table 18.1. Body weight (g) of rats from 3 litters

Statistic	Litter 1	Litter 2	Litter 3
	180^1	195^2	210^3
	185^1	205^2	193^3
	189^1	215^2	190^3
	198^1	213^2	208^3
	203^1	211^2	201^3
Mean	191.00	207.80	200.40
CV (%)	4.93	3.89	4.42

Note: Superscripts indicate litter number.

animals of litter 1 more or less equally to the animals of litters 2 and 3. Similarly, an unbiased randomization should distribute the variation of animals of litter 2 more or less equally to the animals of litters 1 and 3 and so on. This can be achieved if the randomization results in an equal representation of animals from all the three litters to each group.

Just for academic interest the data (Table 18.1) was analysed using one-way ANOVA, and found that there is a significant difference in body weight among the groups.

Assign an arbitrary identification number to each animal and with the help of a random number table randomize the animals into three groups (Table 18.2).

One-way ANOVA of the above data (Table 18.2) resulted in a non-significant F value, indicating that the body weight of the rats did not

Table 18.2. Body weight (g) of rats after randomization

Statistic	Group 1	Group 2	Group 3
	198^1	213^2	185^1
	205^2	189^1	193^3
	210^3	215^2	195^2
	201^3	208^3	190^3
	203^1	211^2	180^1
Mean	203.40	207.20	188.60
CV (%)	2.22	5.07	3.24

Note: Superscripts indicate litter number.

differ among the groups. Strictly speaking, the randomization procedure is completed, but some researchers rearrange the animals among the groups, as explained below, to obtain a uniform mean value. On closely examining the mean values one should be satisfied with the mean values of Groups 1 and 2 since they are somewhat close to each other, but one should be concerned about the mean value of group 3, which deviates

considerably from the mean values of groups 1 and 2, particularly of group 2. This can be overcome by selecting one or two animals based on their body weight from each group and distributing them in other groups in such a way that the mean values of all the groups are more or less similar.

One way to reduce the mean value of group 2 and increase the mean value of group 3 is to take out the rat with the largest body weight from group 2 (215 g) and place it in group 3 and take out the rat with the smallest body weight from group 3 (180 g) and place it to group 2. Now the animals are distributed as given in Table 18.3.

Table 18.3. Body weight (g) of rats after rearranging the animals (first time)

Statistic	Group 1	Group 2	Group 3
	198^1	213^2	185^1
	205^2	189^1	193^3
	210^3	180^1	195^2
	201^3	208^3	190^3
	203^1	211^2	215^2
Mean	203.40	200.20	195.60
CV (%)	2.22	7.39	5.87

Note: Superscripts indicate litter number.

One-way ANOVA of the data given in Table 18.3 indicates that there is no significant different in body weight of rats among the groups. This is still not satisfactory for few researchers. The difference of the body weight between groups 1 and 3 is about 8 g. In order to bring the mean body weight of these two groups closer, one more adjustment is required. A rat of 210 g is taken from group 1 and placed in group 3. Then a rat of 185 g is taken from group 3 and placed in group 1. Now the animals are distributed as given in Table 18.4.

Table 18.4. Body weight (g) of rats after rearranging the animals (second time)

Statistic	Group 1	Group 2	Group 3
	198^1	213^2	210^3
	205^2	189^1	193^3
	185^1	180^1	195^2
	201^3	208^3	190^3
	203^1	211^2	215^2
Mean	198.40	200.20	200.60
CV (%)	3.99	7.39	5.56

Note: Superscripts indicate litter number.

The mean values of the three groups are very close to each other, thus satisfactory. If you closely observe the individual values of the groups, you will realize that Group 3 represents animals from litters 2 and 3 and Groups 1 and 2 represent animals from all the three litters. Rearrangement increases variation within the groups, consequently, the animals respond to a treatment differently. This is evident from the Tables 18.2 and 18.4. The variations (CV%) of groups 1, 2 and 3 after randomization, but before rearrangement were 2.22, 5.07 and 3.24, respectively (Table 18.2). After the rearranging the animals a second time, the variations (CV%) of groups 1, 2 and 3 were 3.99, 7.39 and 5.56, respectively (Table 18.4). Such variations reduce the power of the experiment (Beynen *et al*., 2001). In the first randomization (Table 18.2), each group represented animals from all the litters and the variation (CV%) among the groups are less and somewhat close to each other. Therefore, rearrangements of observations after the randomization to obtain desired mean values should be avoided as far as possible.

Determining Sample Size

In regulatory toxicology, the guidelines clearly indicate the number of animals to be used in a group for a study (Hauschke, 1997). In the research and development of a pharmaceutical company, where a large number of new chemical entities (NCEs) are synthesized, often the scientists carry out experiments with 'inadequate number' of animals. Results from such studies may not be reproducible and may fail to provide the desired information on the effectiveness of the molecule.

Using too few animals in experiments will result in a low power to detect a biologically meaningful results. Similarly, the use of too many animals is not ethical and drain organization's resources unnecessarily. The right number of animals (not too few and not too many) required for obtaining a biologically meaningful result should be an important component of any animal experimental design. In an *in vivo* efficacy study, the number of animals required to obtain the desired result is determined based on certain specifications: the desired magnitude of treatment effect, the chance of obtaining Type I and Type II errors and the inter-individual variability.

An *in vivo* efficacy study is a comparison-oriented study. The comparison of the NCE-treated animals is usually done with the control animals, using an appropriate statistical analysis. The two errors which can occur in such comparisons are Type I error (α error) and Type II error (β error). Though much attention is given to α error, β error is often overlooked. β error is a very potential error in animal experiments and in certain situations more potential than α error. For example, in an *in vivo*

experiment you are confident that there is a treatment-related effect, but the statistical analysis does not show it because of random variation. This is a typical example of β error that commonly occurs in animal experiments. A large β error is a risk in detecting a genuine difference. Power of a study to detect a significant difference is explained by this risk:

Power = 1–β

In simple language, the power is the probability of obtaining a statistically significant result using a statistical test (Lenth, 2007). In other words, power of the test is the probability of correctly rejecting the null hypothesis, when false. A study with a high power is unlikely to fail in detecting a genuine significant difference, whereas a study with a weak power may fail in detecting a genuine significant difference. The power of the tests can be improved by increasing α, sample size, or limiting the statistical analysis to detection of large differences among samples (Hayes, 1987).

To design an experiment to investigate the effect of a hypoglycemic NCE in diabetic rats, the blood sugar in the individual diabetic rat is measured before and after the treatment with the NCE. Then the difference in blood sugar level of the individual rat is calculated. Another group of animals treated similarly, but with a placebo is also maintained. Let us work out number of animals required in each group to obtain the desired result. For that specifications of the study need to be defined:

1. The significance level (probability of α error). Usually it is set at 5% probability level.
2. Probability of β error is set at 10%. The statistical power (1–β) is 90%.
3. The desired treatment effect (difference between NCE treated group and placebo treated group. This is determined based on the factors like clinical, economical etc.)
4. Estimate of expected variation (variation between individual measurements with respect to difference of before and after treatments. This is estimated based on earlier experiments of similar nature or a pilot study)
5. Type of statistical analysis (since there are only two groups, the *t*-test would be better).

Number of animals in each group by two-sided test can be calculated using the formula,

$$n = 2\left[\frac{(Z_{a/2} - Z\pi)^2}{(\mu_1 - \mu_2/\sigma)^2}\right]$$

Number of animals in each group by one-sided test can be calculated using the formula,

$$n = 2\left[\frac{(Z_a - Z\pi)^2}{(\mu_1 - \mu_2/\sigma)^2}\right]$$

Let us work out an Example; $\alpha = 0.05$, $\pi = 0.9$, Desired effect = 25%; σ =15% (CV).

$Z_\alpha = 1.645$ (*vide* Appendix 3 for $Z_{0.05}$)

$Z_\pi = 1.282$ (*vide* Appendix 3 for $Z_{0.10}$)

$$n = 2\left[\frac{(1.645 + 1.282)^2}{(25/15)^2}\right] = 6.2;$$ Number of animals required in each group = 7

Animal Experimental Designs

Accuracy of an animal experiment depends on the design of the experiment. An animal experimental design should be unbiased, should have high precision, wide range of applicability and should be simple in design (Cox, 1958). An animal experiment can be designed in several ways, for example, completely randomized design, randomized block design, cross-over design, Latin square design etc. The commonly used design in pharmacology and toxicology is randomized design. Other designs may be adopted, especially for *in vivo* efficacy studies with NCEs, where more than one NCE at more than two dose levels, a control group, a group treated with a commercially available drug with known efficacy are involved. Perhaps the most important thing to remember while designing an animal experiment is the prior knowledge of all the factors that could affect the outcome of the experiment.

References

Aguilar-Nascimento, J.E. (2005): Fundamental steps in experimental design for animal studies. Acta Cir. Bras., 20(1), 1–8.

Bebarta, V., Luyte, D. and Heard, K. (2003): Emergency medicine research: Does use of randomization and blinding affect the results? Acad. Emerg. Med., 10, 684–687.

Beynen, A.C., Festing, M.F.W. and van Montfort, M.A.J. (2001): Design of Animal Experiments. In: Principles of Laboratory Animal Science, Editors, van Zutphen, L.F.M., Baumans, V. and Beynen, A.C. Elsevier Science B.V., The Netherlands.

Bloomsmith, M.A., Schapiro, S.J. and Strobert, E.A. (2006): Preparing chimpanzees for laboratory research. ILAR J., 47, 316–325.

Capitanio, J.P., Kyes, R.C. and Fairbanks, L.A. (2006): Considerations in the selection and conditioning of old world monkeys for laboratory research: Animals from domestic sources. ILAR J., 47, 294–306.

Cox, D.R. (1958): Planning Experiments. John Wiley & Sons, New York, USA.

Dirnagl, U. and Macleod, M.R. (2009): Stroke research at a road block: the streets from adversity should be paved with meta-analysis and good laboratory practice. Br. J. Pharmacol., 157(7), 1154–1156.

EPA. (2005): United States Environmental Protection Agency. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001B. http://www.epa.gov/iris/backgr-d.htm.

Festing, M.F.W. (1997): Teaching statistics can save animals, In: Animal Alternatives, Welfare and Ethics. Edited by van Zuphen, L.F.M. and Balls, M. Elsevier Science B.V., Amsterdam, The Netherlands.

Festing, M.F.W. (2003): Principles: the need for better experimental design. Trends Pharmacol. Sci., 24, 341–345.

Festing, M.F.W. and Altman, D.G. (2002): Guidelines for the design and statistical analysis for experiments using laboratory animals. ILAR J., 43, 244–258.

Fisher, R.A. (1935): The Design of Experiments. 8th Edition, 1966. Hafner Press, New York, USA.

Hamada, C. and Ono, H. (2000): The role of biostatistics in pharmacological studies (randomization and statistical evaluation). Nihon Yakurigaku Zasshi, 116(1), 4–11.

Hauschke, D. (1997): Statistical proof of safety in toxicological studies. Drug Inf. J., 31, 357–361.

Hayes, J.P. (1987): The positive approach to negative results in toxicology studies. Ecotox. Environ. Safety, 14(1), 73–77.

Hess, K.R. (2011): Statistical design considerations in animal studies published recently in cancer research. Cancer Res., 71, 625.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F.W., Cuthill, I.C., Fry, D., Jane Hutton, J. and Altman, D.J. (2009): Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS One, 4(11), 1–11.

Kozinetz, C.A. (2011): Application of epidemiologic principles for optimizing preclinical research study design. Int. J. Preclin. Res., 2(1), 63–65.

Lenth, R.V. (2007): Statistical power calculations. J. Anim. Sci. 2007. 85 (E. Suppl.), E24–E29.

Lew, M. (2007): Good statistical practice in pharmacology—Problem 1. Br. J. Pharmacol., 152(3), 295–298.

Lim, H.J. and Hoffmann, R.G. (2007): Study Design: The Basics. In: Topics in Biostatistics. Ambrosius, W.T. (Editor). Humana Press Inc., New Jersey, USA.

Nicholls, A.W., Mortishire-Smith, R.J. and Nicholson, J.K. (2003): NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats. Chem. Res. Toxicol., 16, 1395–1404.

Perel, P., Roberts, I., Sena, E., Wheble, P., Briscoe, C., Sandercock, P., Mcleod, M., Mignini, L.E., Jayaram, P. and Khan, K.S. (2007): Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ, 334, 197–200.

Savitz, S.I. (2007): A critical appraisal of the NXY-059 neuroprotection studies for acute stroke: a need for more rigorous testing of neuroprotective agents in animal models of stroke. Exper. Neurol., 205, 201–205.

Schapiro, S.J. and Everitt, J.I. (2006): Preparation of animals for use in the laboratory: Issues and challenges for the institutional animal care and use committee (IACUC). ILAR J., 47(1), 370–375.

Vogt, L. and Kloting, I. (1990): Effects of randomization masking diabetes relevant traits in animal experiments. Diabetes Res., 15(3), 131–135.

Workman, P., Aboagye, E.O., Balkwil, F., Balmain, A., Bruder, G., Chaplin, D.J., Double, J.A., Everitt, J., Farningham, D.A.H., Glennie, M.J., Kelland, L.R., Robinson, V., Stratford, I.J., Tozer, G.M., Watson, S., Wedge, S.R. and Eccles, S.A. (2010): Guidelines for the welfare and use of animals in cancer research. British J. Cancer, 102, 1555–1577.