Chapter 10
General Toxicology, Safety Pharmacology, Reproductive Toxicology, and Juvenile Toxicology Studies

Steven A. Bailey, Dingzhou Li and David M. Potter

Abstract This chapter provides a survey of key nonclinical safety assays. For each study type, we discuss the typical study designs employed, including a summary of the type of endpoints collected. We then provide an overview of common statistical approaches in each setting. There are some general themes that are common across the study types (e.g., trend testing). At the same time, the different study types may have features that require special consideration (e.g., cross-over designs for safety pharmacology studies, intra-litter correlation in reproductive toxicology studies). While some of the design aspects of these studies are to some extent "fixed" by precedent across the industry, we do address sample size and power considerations, as this information can be valuable to understanding how statistical results can contribute to the overall interpretation of these studies. Finally, for any discussion of statistical approaches, there are likely to be multiple reasonable approaches. We've attempted to cover some of the more common approaches in detail, but we recognize that our treatment is not exhaustive. Where possible, we have provided references for further reading.

Keywords Toxicology • Juvenile toxicology • Reproductive toxicology • Safety pharmacology • Preclinical safety • Trend testing

10.1 Introduction

As was discussed in Chap. 9, a range of safety studies must be conducted prior to clinical trials. General toxicology, genetic toxicology and safety pharmacology studies, for example, must be conducted prior to Phase I clinical trials. The need for and timing of other safety studies, such as reproductive toxicology and carcino-

S.A. Bailey
Drug Safety R&D Statistics, Pfizer Inc, Andover, MA, USA

D. Li • D.M. Potter (✉)
Drug Safety R&D Statistics, Pfizer Inc, Groton, CT, USA
e-mail: david.m.potter@pfizer.com

L. Zhang (ed.), *Nonclinical Statistics for Pharmaceutical and Biotechnology Industries*, Statistics for Biology and Health, DOI 10.1007/978-3-319-23558-5_10

genicity studies, will depend on the type of drug (e.g., small molecule, biologic) and the intended patient population. These safety studies are termed “regulatory” studies because of requirements and guidance from regulatory agencies on the scope, duration, and timing of these studies. The three main regulatory bodies, the US Food and Drug Administration (FDA), the European Medicines Agency (EMA), and the Japanese Pharmaceutical and Medical Devices Agency (PMDA), differ in some instances in their recommendations regarding the safety studies needed to support various stages of clinical development. There have, however, been efforts to standardize and harmonize recommendations across the three regions (International Conference on Harmonization 2010). This chapter addresses general toxicology, safety pharmacology, reproductive toxicology, and juvenile toxicology studies.

10.2 General Toxicology

10.2.1 Overview

The overall objectives of general toxicology studies include identifying target organ toxicity (i.e., which organs are potentially affected), characterizing the dose–response or exposure-response relationship, assessing the potential reversibility of test article effects, and identifying possible endpoints to use to monitor adverse events in clinical trials. A wide range of endpoints are collected during these studies. Quantitative endpoints include body weight, food consumption, organ weights, and clinical pathology measurements. Clinical pathology can include hematology, serum chemistry, and urinalysis endpoints. In addition, more qualitative endpoints include clinical signs, gross pathology, and histopathology. Toxicokinetics (measurement of exposure to the test article) are included in these designs to monitor drug exposure levels, to aid with interpretation of findings in the study, and aid in selecting starting clinical doses.

An expectation of these studies is that they establish a dose–response relationship, from no effect to adverse effect. To do this, most studies include a control and at least three doses of the drug being evaluated, with doses chosen to exceed anticipated clinical doses, and the high dose chosen to induce toxicity in order to help identify target organ toxicity. The highest dose selected is typically the “maximum tolerated dose” (MTD) meaning that animals show evidence of toxicity (e.g., decreased body weight, changes in clinical signs) but do not experience mortality or morbidity. Dose selection and its impact on the effective and efficient use of animals is a critical consideration in these studies. If doses are chosen too high (potentially causing mortality), or too low to cause any toxicity, studies may need to be repeated.

The studies are typically conducted in a rodent species (e.g., rat) and a non-rodent species (e.g., dog, non-human primate (NHP)). For rodents, there are typically 10 animals per sex per treatment group. For non-rodents, 3–4 animals per sex per group is common. In many cases, exploratory (i.e., non-regulatory) studies are conducted in advance of the regulatory studies in order to inform the design of the regulatory studies. These studies are typically smaller (e.g., 5 rodents per sex per group, 1 non-rodent per sex per group).

The duration of a general toxicology study is chosen based on the clinical studies it supports. The first set of general toxicology studies are in support of Phase I clinical trials in humans. Initial "dose-range finding" (DRF) studies are non-regulatory studies that seek to identify the maximum tolerated dose (MTD) to inform dose selection in subsequent regulatory studies. Their duration is usually 2 weeks or less, but may be much longer depending on the half-life of the test article. Regulatory studies, typically up to 1 month in duration, are then conducted. To support Phase II studies, longer (typically up to 6 months) regulatory studies are required, again, in both rodents and non-rodents. To support Phase III studies and post-approval use in patients, regulatory studies up to 9 months in duration are conducted, usually in non-rodents. In addition, carcinogenicity assessments for lifetime exposure are conducted (typically 2 years in rats and 6 months in transgenic mice). Some recent research on the design and analysis of carcinogenicity studies is presented in Chap. 12.

For the most part, these studies are "multi-dose" or "repeat-dose" studies, meaning that the drug is administered regularly (e.g., daily) during the course of the study. "Acute" studies, by contrast, are characterized by either a one-time administration of the drug or possibly repeated doses in a short time-frame (e.g., 1 day).

The results of these studies, in conjunction with other nonclinical safety assessments, contribute to the identification of No Observed Adverse Effect Level (NOAEL). Definitions of the NOAEL vary somewhat, but in general it is taken to be the highest experimental dose without an adverse effect. Note that confidence in the determination of the NOAEL will depend on the study design. For an excellent review and discussion of the NOAEL and its limitations, as well as alternative approaches, see Dorato and Engelhardt (2005). Included in their article is a discussion of the Benchmark Dose (BMD) Method (BMD), introduced by Crump (1984). The idea is to fit a dose–response model to the study data and select through calibration (i.e., inverse prediction) the dose level that corresponds to a prespecified adverse response (e.g., a certain percentage increase over the control group response). Then, the lower bound on a confidence interval for the dose level is used as the identified dose. For a review of the BMD approach, see Filipsson et al. (2003).

10.2.2 Statistical Analysis Methods

10.2.2.1 Comparing Dose Groups to Control

Since these are parallel group designs, typically with a control and three dose groups, the analysis options are relatively straightforward. Analysis of variance

(ANOVA) methods can be used, with pairwise comparisons of each dose group back to control. Because the primary comparisons of interest are typically relative to a single control group, Dunnett's Test (Dunnett 1955) is often used. Trend testing methods are also used, taking advantage of the natural ordering of the dose groups in most studies. This approach allows for an overall assessment of a dose–response relationship, and with sequential testing variations (Tukey et al. 1985), also provides a way of estimating a "no statistical-significance of trend dose [NOSTASOT]", as Tukey described it. Assuming a control group and three dose groups (low, intermediate, and high), a sequential trend test could be conducted as shown in the flowchart in Fig. 10.1.

The advantage of trend testing methods is that for monotonic dose–response patterns, the methods are more sensitive than pairwise comparisons alone. Note that the NOSTASOT for a particular endpoint is being declared based on a *lack* of statistical significance, which could result due to lack of a true effect, or due to a lack of power to detect an effect.

There are several options for implementing trend testing. One common approach is to estimate linear contrasts in the context of an ANOVA. Consider the data shown in Table 10.1 and Fig. 10.2 for the liver enzyme alanine aminotransferase (ALT) from a hypothetical 1-month general toxicology study. Elevations in ALT often reflect changes in liver structure or function.

The data suggest an elevation in ALT levels with increasing dose. Note also that variation appears to increase with the magnitude of ALT. This is common for many clinical pathology parameters, and a log transformation is often appropriate. For this example, Levene's Test didn't suggest strong evidence of unequal variance ($p = 0.085$), and hence we analyze the data on the original scale. The overall F-test from an ANOVA indicates a significant difference among the groups ($F = 23$, $p < 0.001$, $df = 3.16$). The toxicologist is specifically interested in which dose groups differ from control, and so pairwise comparisons are conducted. Table 10.2 shows the results of pairwise comparisons using Dunnett's Test and a sequential trend test.

Table 10.3 shows the linear contrasts used for the trend test. Contrast 1 tests for an overall linear trend among all four dose groups. Contrast 2 tests for a linear trend among the control, low, and intermediate dose groups only, and it is only conducted if Contrast 1 is statistically significant. Similarly, Contrast 3 tests for a difference between the control and low dose group and is only conducted if Contrast 2 is statistically significant. For these data, both the high and intermediate doses would be declared significantly different from control, at the 5 % level. Note that using this sequential approach, testing at subsequent doses only occurs if the initial contrast is statistically significant. Hence, the overall (i.e., family-wise) error rate of the procedure is less than or equal to α.

In contrast, Dunnett's Test indicates a difference between the high dose group and control only, at the 5 % level. In general, in settings where monotonic dose–response patterns are expected, then trend testing methods will be more powerful than Dunnett's Test. Simulations can be used to assess the extent of the advantage. For example, assume that the true mean levels of ALT are (21,24,26,29) in the

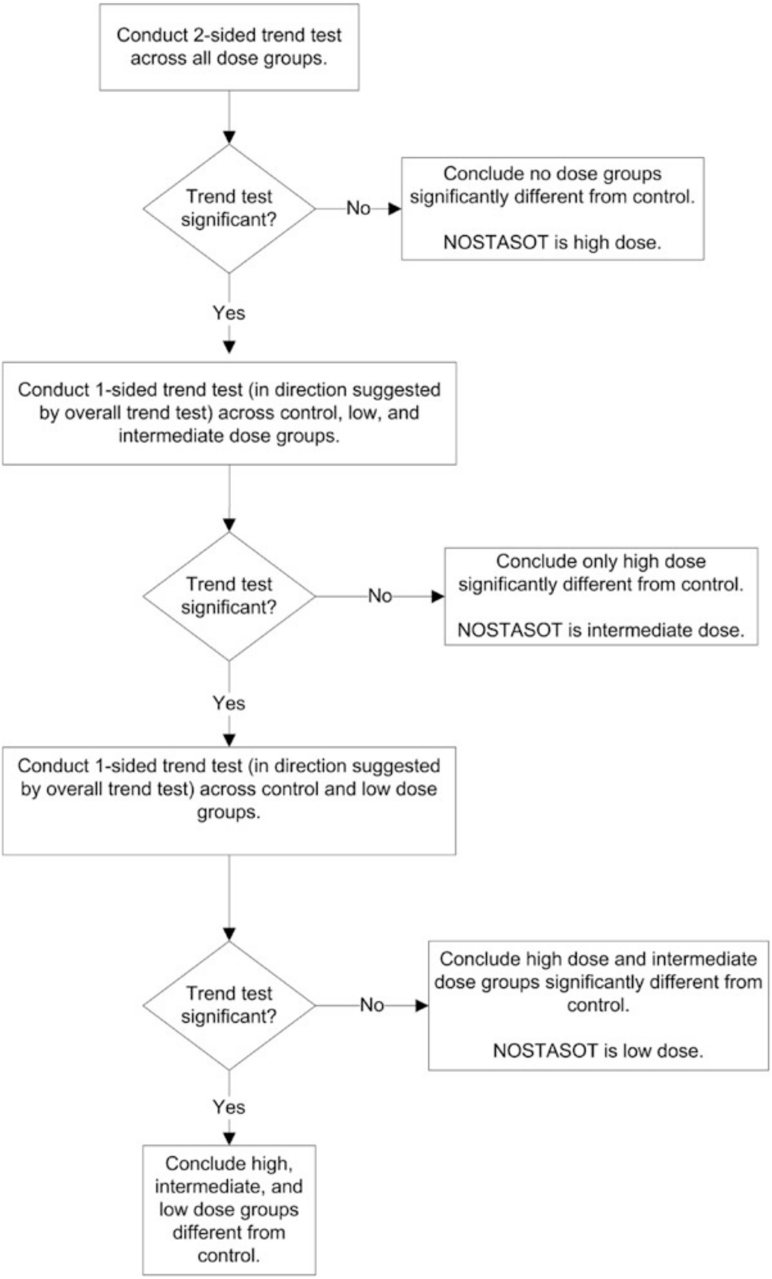
Conduct 2-sided trend test across all dose groups.
Trend test significant?
No
Conclude no dose groups significantly different from control.
NOSTASOT is high dose.
Yes
Conduct 1-sided trend test (in direction suggested by overall trend test) across control, low, and intermediate dose groups.
Trend test significant?
No
Conclude only high dose significantly different from control.
NOSTASOT is intermediate dose.
Yes
Conduct 1-sided trend test (in direction suggested by overall trend test) across control and low dose groups.
Trend test significant?
No
Conclude high dose and intermediate dose groups significantly different from control.
NOSTASOT is low dose.
Yes
Conclude high, intermediate, and low dose groups different from control.

Fig. 10.1 Flow chart for sequential trend testing

Table 10.1 Example hypothetical ALT (U/L) data from a 1-month general toxicology study in rats

Control	Low dose	Intermediate dose	High dose
21.3	17.3	33.6	40.3
16.9	20.2	22.8	48.5
24.4	23.1	28.3	34.5
19.2	27.6	25.2	38.6
21.3	19.2	26.0	51.0

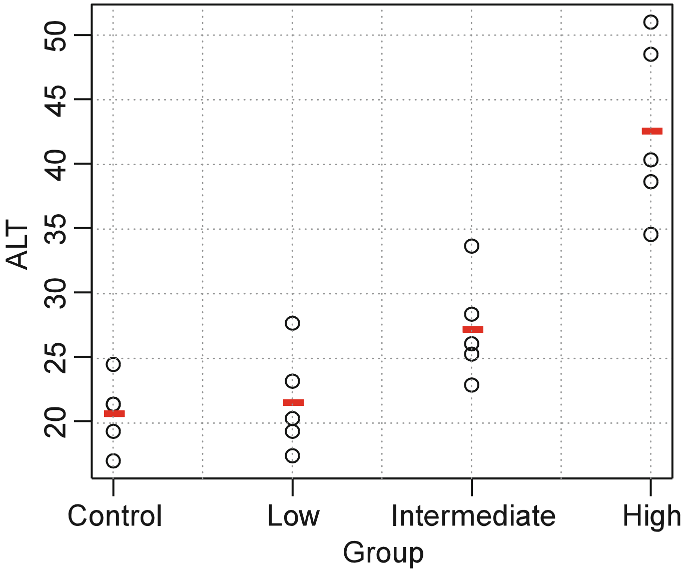

Fig. 10.2 Scatterplot of example ALT data

Table 10.2 Summary of pairwise comparisons for ALT

				P-value	
Treatment	N	Mean	SD	Dunnett	Trend
Control	5	20.6	2.8	–	–
Low	5	21.5	4.0	0.98	0.39
Intermediate	5	27.2	4.1	0.10	0.02
High	5	42.6	6.9	<0.01	<0.01

Table 10.3 Linear contrasts for trend testing with four treatment groups

	Contrast 1	Contrast 2	Contrast 3
Control	−3	−1	−1
Low	−1	0	1
Intermediate	1	1	0
High	3	0	0

Table 10.4 Power to detect differences between high dose group and control using sequential trend test and Dunnett's test

N per group	Power – Trend test (%)	Power – Dunnett's test (%)
3	32.0	17.4
4	42.7	25.9
5	52.9	33.8
6	61.2	42.3
7	70.0	51.6
8	75.7	58.5
9	80.4	63.9
10	84.4	69.2
15	96.1	89.5
20	98.9	96.2

Simulated group means = (21.24,26,29). Standard deviation = 6. 10,000 runs

control, low dose, intermediate dose, and high dose groups respectively, and that our estimate (from historical control data) of the standard deviation is 6. For a range of sample sizes, Table 10.4 shows the proportion of times the high dose group was significantly different from the control group at the 5 % level.

In addition to the linear contrasts approach, there are other methods for testing for trends. For example, some methods assume only a monotonic response. The null and alternative hypotheses in this setting are:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_\mathrm{A} : \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \, (\text{or } \mu_1 \geq \mu_2 \geq \mu_3 \geq \mu_4)$$

with at least one strict inequality. One method due to Williams (1971, 1972) is basically a series of pairwise t-tests of each dose group versus control, based on *amalgamated* means. Because of this similarity to the traditional *t* test, it was termed the $\bar{t}$ test. The amalgamation procedure enforces a non-decreasing (or non-increasing) ordering of the sample means, using the *pooled adjacent violators* (*PAV*) *algorithm* Consider an example where the dose group means are (2,4,6,5) as shown in Fig. 10.3. The PAV algorithm moves left to right, checking for non-monotonicity among adjacent pairs. In this case, the intermediate and high dose groups violate monotonicity, so their means are averaged. The final group means are thus (2,4,5.5,5.5). In a second more extreme example, assume that the dose group means are (2,7,7,1), as shown in Fig. 10.4. In this case, the intermediate and high dose group means violate monotonicity, and so their means are pooled, resulting in group mean equal to (2,7,4,4). Because the low dose group mean now violates monotonicity when compared to the pooled intermediate and high dose group means, those three group means are pooled, resulting in (2,5,5,5) for the final amalgamated group means. In addition to the pairwise t-tests based on amalgamated

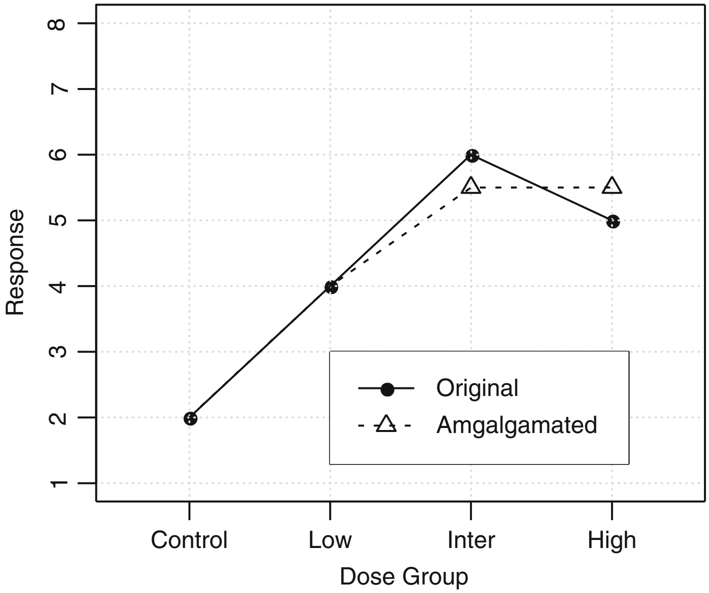

Fig. 10.3 Example of amalgamation procedure

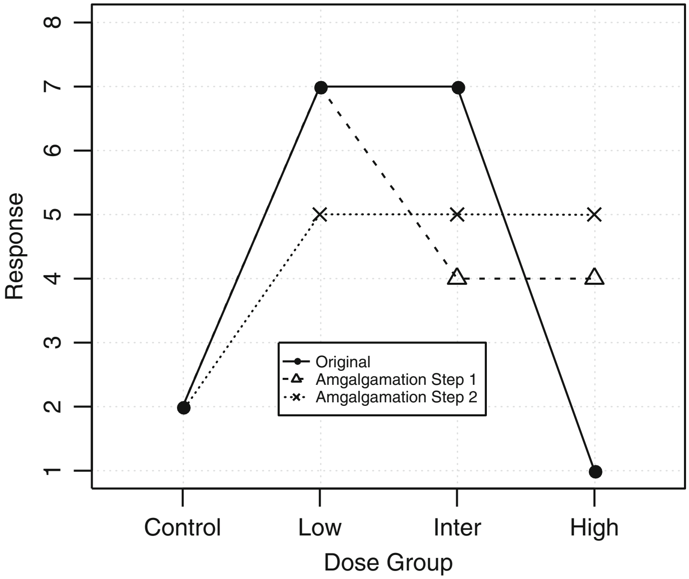

Fig. 10.4 Example of multiple-step amalgamation procedure

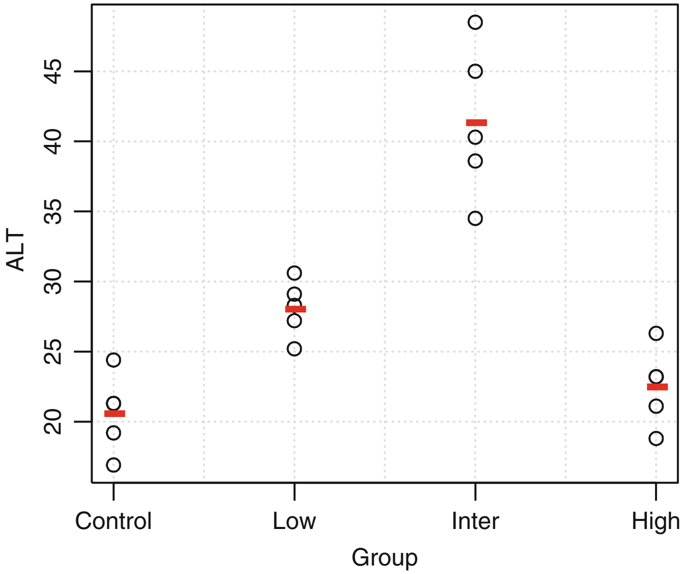

Fig. 10.5 Example of non-monotonic relationship with dose group

means due to Williams, there is also a trend test based on amalgamation means. This test, called the $\overline{E}^2$ test (Barlow et al. 1972).is essentially an overall ANOVA F-test based on amalgamated means. For more details on the $\bar{t}$ and $\overline{E}^2$ tests, including critical values and calculation of p-values, see the original papers, as well as Bailey (1998).

Occasionally in general toxicology studies, the observed pattern in group means with increasing dose is not monotonic. See, for example, data shown in Fig. 10.5. In these cases, sequential trend testing methods lead to one of two conclusions. In one case, the initial trend test will not be significant, so the sequential testing stops with no dose groups being declared different from control. In a second case, the initial trend test across all dose groups will be significant (due to the influence of the low and intermediate dose groups), leading to the high dose group being declared significant. In the case shown in Fig. 10.3, the initial trend test (using linear contrasts) is significant ($p = 0.016$), leading to the conclusion that the high dose group differs from control. In either of the two cases, interpretation can be challenging. One way of addressing this issue is to include a check for monotonicity upfront. One approach to testing for non-monotonicity is based on comparing the group means with the amalgamated means, using an F-test. See Healey (1999) for more details. If there is evidence of non-monotonicity, then the results from pairwise comparisons (say, using Dunnett's Test) could be provided instead. Alternatively, the approach taken by Bretz and Hothorn (2003) could be considered, in which multiple contrast tests (MCT's) are used to identify a potential trend only up to a given dose level.

10.2.2.2 Parametric vs. Nonparametric Methods

In a typical statistical comparison of groups (e.g., using an ANOVA model), the researcher checks (often visually) distributional assumptions such as normality of the residuals and equal variability across treatment groups. Depending on the assessment, it might make sense to transform the dependent variable (e.g., log) or use a nonparametric approach. However, in the analysis of general toxicology studies, the statistical methods are often implemented as part of automated systems; the same model will be run for multiple endpoints, possibly across multiple timepoints. Hence, it may not be feasible to check data distributions and other assumptions for each model at the time of analyses. There are at least three approaches to handling this issue.

One approach is to automate the assessment of the distributional or other assumptions prior to analysis. This approach is often represented as a "decision tree" (not to be confused with classification and regression trees used in predictive modeling) or flow chart. For example, depending on an initial test of normality, the data may or may not be rank transformed, or undergo some other transformation, prior to analysis. Note that in some systems we have encountered, data are analyzed nonparametrically if an initial test suggests a departure from the assumption of constant variability across treatment groups. Many nonparametric approaches (e.g., Wilcoxon, Kruskal–Wallis) still rely on the homogeneous variance assumption, however. A second approach is to evaluate historical control data to evaluate the distribution of each endpoint. Those variables that deviate appreciably from normality could be routinely analyzed using a log or rank transformation; these choices for each endpoint would be prespecified in the automated system. A third approach is to use a rank transformation for all parameters. This would guard against cases where an extreme value may mask a potential effect, and in general won't result in an appreciable loss of power. For example, using the setup in the previous section, with mean ALT equal to (21, 24, 26, 29), we can compare the power of a rank-based approach to the original, using the sequential trend test based on linear contrasts. The results are shown in Table 10.5 and suggest some loss of power with the rank-based approach, but typically only a few percentage points in this scenario.

In addition to applying contrast-based trend tests to the ranks of the data, there other nonparametric implementations of trend tests, including those due to Shirley (1977) and Jonckheere (1954).

10.2.2.3 Sex Effects

General toxicology studies are typically conducted in both sexes. Traditionally, statistical analyses have been conducted separately for each sex. There may be a gain in sensitivity by conducting a combined analysis including a model term for sex (i.e., as a block) in the ANOVA. The challenge arises if there is a statistical interaction between sex and dose for one or more parameters in a given study. It is not uncommon to observe sex-related differences in exposure due to differential

Table 10.5 Power comparison of parametric and rank-based sequential trend tests

	Based on original data values			Based on ranks		
N per group	Low dose (%)	Intermediate dose (%)	High dose (%)	Low dose (%)	Intermediate dose (%)	High dose (%)
3	15.6	24.4	32.0	15.9	24.9	31.7
4	17.7	30.0	42.7	17.9	30.0	41.3
5	19.3	35.4	52.9	18.7	34.4	50.4
6	21.6	39.6	61.2	21.2	39.2	58.9
7	23.0	45.4	70.0	22.9	44.7	67.6
8	26.4	49.6	75.7	26.0	48.7	73.0
9	27.5	53.6	80.4	26.7	53.1	78.0
10	29.4	57.0	84.4	28.5	56.4	82.7
15	38.2	73.0	96.1	37.4	71.6	95.2
20	47.6	83.5	98.9	46.5	82.3	98.6

True group means = [21,24,26,29]. Sequential linear contrast approach

metabolism or hormonal influence. In this case, the toxicologist will want to evaluate the impact of treatment separately for each gender. Again, since these analyses are often automated, the conventional approach has been to assume the potential for an interaction, and analyze by gender. This is an area for continued development.

10.2.2.4 Time Effects

Some endpoints, such as organ weights, can be collected only once in a general toxicology study. Others, such as body weights and food intake, are captured more frequently (e.g., weekly). Clinical chemistry, hematology, and urinalysis parameters are typically collected at the end of study, but may also be collected multiple times (e.g., monthly in a 3-month study). In some cases, and especially for large animal studies (e.g., NHP, dog), a baseline measurement (i.e., prior to any dosing) may be taken for each animal. In these studies, it is typical for the toxicologist and/or clinical pathologist to focus on changes from baseline values, rather than comparing control and test article-administered animals at each time point, especially when the sample sizes are small (e.g. 3/sex/group).

Incorporating a baseline adjustment (i.e., as a covariate) into an ANOVA model may, in some cases, improve the power of these analyses. However, it is important to evaluate the extent of correlation between baseline and follow-up measurements for each endpoint. A recent internal Pfizer study of clinical pathology control data from 20 GLP general toxicology studies in NHP’s showed a substantial range in correlation (approximately 0.15–0.95) across about 45 endpoints. Some endpoints (e.g., ALT) had within-animal correlations above 0.90. Others (e.g., glucose) were in the range of 0.35. Overall, more than 1/3 of the endpoints had within-animal correlations below 0.5; for these endpoints, including a baseline covariate may actually reduce the sensitivity of statistical tests.

Table 10.6 Reference ranges for select clinical chemistry parameters, based on Wistar rats

Analyte	Range	Units
Glucose	91–218	mg/dL
Potassium	3.3–5.0	mmol/L
Cholesterol	30–71	mg/dL
Alanine Aminotransferase (ALT)	15–66	U/L

10.2.2.5 Reference Ranges

In addition to comparisons between dose groups and concurrent controls in a general toxicology study, for many endpoints there are well-established reference ranges, based on historical control data, against which to compare individual data values. A reference range is defined as an interval in which some percentage (e.g. 95 %) of an endpoint's values would fall, assuming a healthy population of subjects. These intervals can serve as the basis for determining whether individual drug-treated animals are unusual in their response. There are both parametric and nonparametric approaches to constructing these intervals. For the former case, the data (possibly transformed) are assumed to be normally distributed, and quantiles (e.g., 2.5 % and 97.5 %) are derived based on normal theory. In the nonparametric case, the sample quantiles are computed directly from the data. Some example reference ranges for Wistar Han IGS rats based on recent historical control data at Pfizer are shown in Table 10.6. These reference ranges were calculated using the EP Evaluator software (EP Evaluator 2005), which uses a nonparametric approach (Clinical and Laboratory Standards Institute 2000). When constructing reference ranges, it's important to note that there are species, strain, and age differences (e.g., a Wistar Han IGS rat is not the same as a Sprague–Dawley rat), reference ranges drift over time, and may be specific to a facility or testing platform.

An important step in computing reference ranges is to ensure that the samples used are relatively homogeneous with respect to key attributes like age, sex, and species, to the extent that these factors affect the normal range of values.

With a well-constructed historical control database, meaningful investigations into other possible sources of variation (e.g. seasonal) can be performed. See Sect. 9.3.2 in Chap. 9 for further information on plotting historical control data.

10.2.3 Sample Size and Power Considerations

Assessing the statistical power of general toxicology studies poses several challenges. In a simple two-sample comparison (one treatment group and one control group) for a given variable, a typical sample size calculation relies on an estimate of variability and a required difference to detect between group means. Estimating biological variability is relatively straightforward, given an adequate set of historical control data. Elucidating a single agreed-upon difference to detect for a given endpoint is often more difficult, as it may depend on a particular toxicologist's experience as well as the particular compound being studied and the disease area. In addition, a change in a single endpoint is rarely interpreted on its own; instead, the change is interpreted in the context of changes in other endpoints, both quantitative and qualitative. In this sense, the statistical results are univariate in nature, but the interpretation by the toxicologist is multivariate. Even having agreed on a difference to detect each endpoint, the question remains as to how to assess the suitability of the sample size relative to all of the collected endpoints (food consumption, body weights, organ weights, and clinical pathology data).

In general, the sample sizes used in general toxicology studies appear to be driven primarily by regulatory guidance and historical precedent. For example, an excellent review article by Sparrow et al. (2011) states:

> In regulatory general toxicology studies the animal numbers used are not driven by statistical input. There are several reasons for this, such as the potential hazards of a substance being unknown in advance of the studies being conducted. Therefore, there is no specific change that the study can be statistically powered to detect. In addition, the frequency of the potential hazard is unknown in the initial toxicology studies and may turn out to be a frequently occurring or a low incidence change.
>
> Due to the multifactorial nature of toxic changes, assessment of toxicity in all species is made by examination of the data generated for each individual animal by integration and correlation of in-life and post mortem findings. Experience has demonstrated that the numbers used and illustrated in this manuscript are sufficient to identify the most potential hazards, a dose/exposure response for the hazards and to generate data sets that are sufficient to provide study sponsors and regulators with information that allows decisions to be made about clinical trials and marketing.

Even with fixed sample sizes, the machinery of power calculations can still be used. That is, given a sample size and an estimate of variability, we can compute the minimum detectable difference (MDD) for each endpoint, assuming 80 % power. This can lead to fruitful discussions about the types of changes that can be meaningfully detected with statistical analyses.

10.3 Safety Pharmacology Studies

10.3.1 Overview

The main objective of safety pharmacology studies is to understand potential undesirable pharmacodynamic effects of a test article or an intervention on physiological functions in relation to exposure in the therapeutic range and above. A more comprehensive definition of safety pharmacology studies is given in ICH S7A guideline (International Conference on Harmonization 2001). We focus on the three major types of in vivo safety pharmacology studies: pulmonary-respiratory, cardiovascular (CV), neurofunctional (NF) experiments. Each type focuses on an important aspect of the possible acute adverse effects (typically within 48 h after dosing) caused by the test article or the intervention.

10.3.2 Pulmonary-Respiratory Studies

The focus of pulmonary-respiratory studies is to de-risk targets and test articles that may possess respiratory issues. The typical study design is a parallel group. Following a period of acclimation, unrestrained animals are placed in a whole body plethysmograph chamber for approximately 6 h. Originating from the pressure change in the chamber, the respiratory signal is routed through an amplifier to a data acquisition system, and the respiratory parameters are logged by the computer. The main parameters in the pulmonary-respiratory study are the tidal volume (i.e., the normal volume of air displaced between normal inhalation and exhalation), respiratory rate, and minute volume (i.e., the volume of air inhaled or exhaled from the lung per minute). The raw data (usually a data point every 5 s) are then averaged into some sequential time interval bins, including one at the baseline (i.e., prior to dosing), for subsequent statistical analysis.

10.3.3 Cardiovascular Studies

The primary goal of the CV studies is to determine if there is a CV risk associated with a test article or an intervention, measured by blood pressure, heart rate, and electrocardiogram readings. Typically, these endpoints are collected from "telemeterized" animals that are free to move about during the course of the study. A clear advantage of this technology is that it has minimal interference with the animal's normal function. Multi-channel signals are transmitted wirelessly from surgically-placed electrodes to the receiver in the monitoring room, giving a comprehensive recording of the cardiovascular changes in real time. At an adequate sampling rate, such a data acquisition system provides a more accurate picture of the dynamics in the physiology and may capture minute responses that would not be possible using non-ambulatory (i.e., recumbent) approaches, which require the animal to be restrained.

10.3.4 Neurofunctional Studies

Effects of the test article or the intervention on the central nervous system (CNS) are typically evaluated using assessments of motor activity, coordination, sensory/motor reflex responses, behavioral changes, and body temperature. Two key study types are the functional observation battery (FOB) and locomotor activity (LMA). FOB is the mainstay observational assay designed to identify points of CNS concern for follow-up. It consists of a battery of endpoints (the number varying from company to company, usually 20 to 30) covering different aspects of CNS issues, such as activity and excitability, and autonomic, neuromuscular, and sensory/motor responses. Most of the endpoints are binary (Yes/No), with the rest having more than two levels (e.g., Normal, Mild, and Severe). Upon completion of the FOB and body temperature assessment, the animal will be immediately placed into its assigned locomotor activity chamber and LMA testing will begin. LMA will include assessments of both horizontal (XY ambulation) and vertical movements (rearing).

10.3.5 Statistical Analyses

10.3.5.1 Overview

Although these studies measure different physiological aspects of the animal, the statistical analysis methods are similar across study types, depending on the nature of the endpoint of interest. For continuous endpoints, the main tool is ANOVA or ANCOVA. If a baseline or pre-dose measurement exists, ANCOVA with the baseline as a covariate is recommended as in general, it will be a more powerful approach (Senn 2007).

The discrete endpoints are typically analyzed using pairwise Fisher's exact tests or the Chi-square test. Since directionality is presumed for adverse events (i.e., Abnormal is always worse than Normal), comparisons between dose groups are made using one-tailed tests. Additionally, the Cochran–Armitage test or the Jonckheere–Terpstra test may be performed when the intent is to characterize a dose response relationship.

10.3.5.2 Super-Intervals

In safety pharmacology studies, particularly in CV studies, the raw data coming out of the data acquisition systems usually have high resolutions (at the sampling rate of 200 Hz there is a data point every 0.005 s). Even if the software uses a procedure to construct "moving-average" summaries of the individual recordings into longer intervals such as 15-min time bins, the process still generates a huge amount of data over a period of 24–48 h. Therefore, further coarse-graining is needed to reduce the data quantity without significant loss of fidelity in representing the characteristic physiology. To that end, the so-called super-interval binning method (Sivarajah et al. 2010) has been proposed to take into account the particular pharmacokinetics profile of the test article in a given experiment, and has been shown to have reasonably good performance in pilot studies using many known positive controls. With this method, each animal typically ends up with four to six observations per day per treatment. An example of super-intervals is shown in Fig. 10.6.

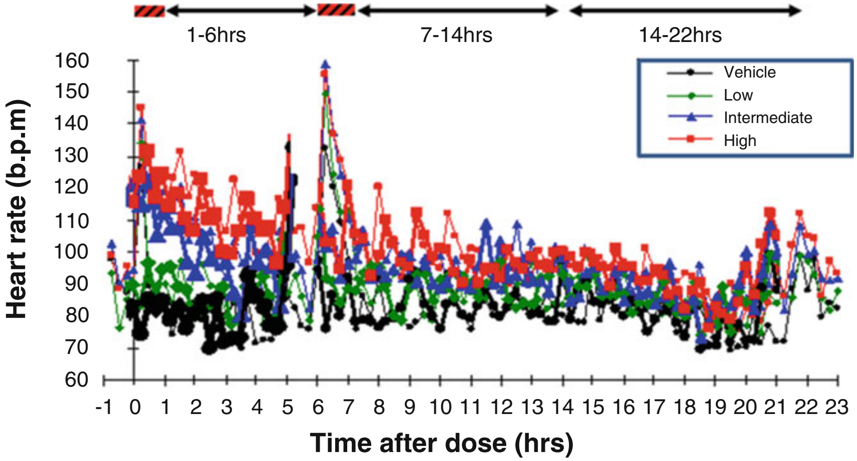

Fig. 10.6 Example of super-intervals in a cardiovascular study

10.3.5.3 Repeated Measures vs. the Cross-Sectional approach

The statistician has several options with the super-intervals. One approach is a repeated-measure ANOVA/ANCOVA, which takes into account the within-subject correlation across time intervals. Alternatively, ANOVA/ANCOVA models can be fit at each time interval (this approach is often called a “time-by-time” or “cross-sectional” approach). The rationale of either approach will be discussed in subsequent sections for each study type. An advantage of the time-by-time approach is that identifying and/or modeling the precise form of the correlation is not a concern. On the other hand, the approach may suffer from the loss of statistical power relative to the repeated measures approach (detailed discussion on this will be provided in Sect. 10.3.6).

The repeated-measure approach is fairly straightforward except when it comes to the choice of covariance structure. When the super-intervals have equal lengths and are evenly spaced, it is natural to assume an AR(1) structure between observations on the same animal. In contrast, when the super-intervals vary significantly in length and position, the AR(1) structure may not fit as well. One approach is to try several candidate structures and use criteria such as AIC or BIC to select the best model. However, this procedure may not be robust against data changes, meaning that missing data, replacement animals, or even the experimental day could change the overall structure. Furthermore, simulation studies have shown that using criteria such as AIC, one could still select an incorrect covariance structure. Furthermore, with small sample sizes in some safety pharmacology studies, we may not be able to afford to use an overly complex structure. With these considerations, it is common to start with the simplest structure: compound symmetry. Only when the estimate of the intra-class correlation is negative would we consider alternatives to compound symmetry. In fact, when negative intra-class correlation happens, it is usually due

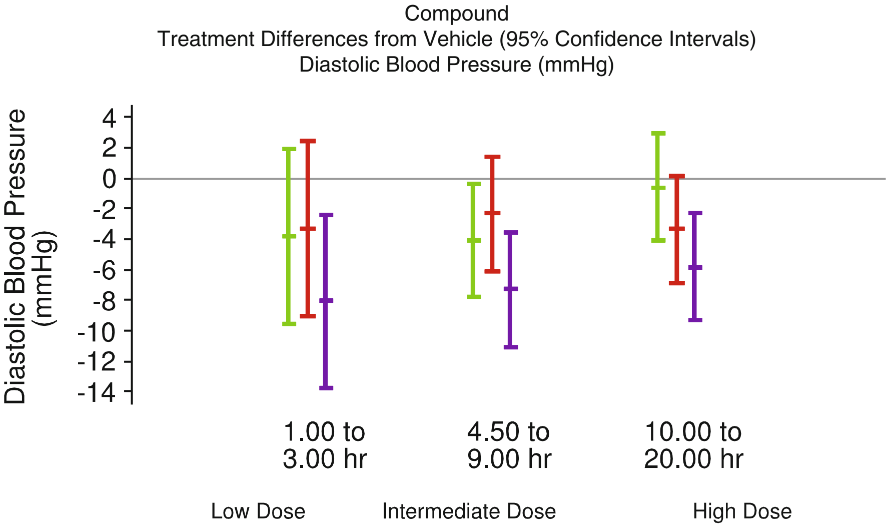

Fig. 10.7 Example confidence intervals from CV study

to one or two animals that behaved drastically differently from the rest. In this case, we may have to abandon the repeated-measure paradigm altogether and switch to the cross-sectional method.

10.3.5.4 The Use of Confidence Intervals

Although significance testing and p-values have been extensively used in assessing test article effects, confidence intervals are also recommended. Confidence intervals and significance testing are linked as they have the same components (e.g. difference in means and standard error of that difference); however, a confidence interval provides more information than a single p-value. An example of the 95 % confidence intervals for treatment effects in a hypothetical CV study is shown in Fig. 10.7. The statistical significance and magnitude of the treatment effect at each time interval as well as their evolution over time can be clearly identified from the figure.

Confidence intervals also serve as a useful bridge between statistical significance and biological relevance. In Fig. 10.8, each confidence interval corresponds to a hypothetical CV study. Consider a common CV endpoint, the QT interval. A drug-induced prolongation of the QT intervals can lead to serious adverse effects. Typically, a heart-rate adjusted version, the “corrected QT (QTc)” is used. Suppose a QTc change of 8 ms or higher is considered biologically important. Then our conclusions based on the confidence intervals are:

- Study 1 was both statistically significant and biologically relevant (indicating a QTc prolongation);

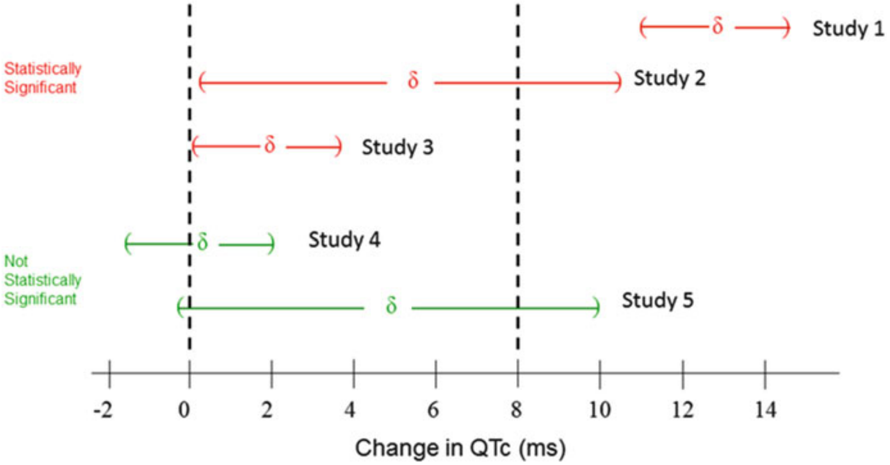

Fig. 10.8 Confidence intervals for QTc measurements from hypothetical CV studies

- Study 2 was statistically significant but its biological relevance was unknown;
- Study 3 was statistically significant but not biologically relevant;
- Study 4 was neither statistically significant nor biologically relevant;
- Study 5 was not statistically significant and its biological relevance was unknown.

Therefore, Studies 2 and 5 did not provide definitive information about the biological relevance of the QTc effect, and hence either further reduction in the measurement variability or a larger sample size is needed in those cases.

10.3.5.5 Trend and Monotonicity Testing

Many safety pharmacology endpoints are likely to demonstrate a monotonic dose response relationship. In these cases, greater statistical power can be achieved using trend testing methods. For that reason, a decision-tree type of analysis is sometimes adopted. Specifically, a monotonicity test (see some of the methods discussed in Sect. 10.2.2.1) is carried out upfront. If monotonicity cannot be rejected, then the significance of the treatment effect of each dose will be determined from a sequential trend test. On the other hand, if monotonicity is rejected, then the significance of the treatment effect will be based on pairwise comparisons (i.e., t-test) between each dose and the vehicle. This decision-tree method is able to detect treatment effects that would otherwise not be caught by either method (i.e., trend test or pairwise t-test).

Table 10.7 Latin squares design for crossover study

Animals	Treatment 1	Treatment 2	Treatment 3	Treatment 4
1	Vehicle	Low	Intermediate	High
2	Low	High	Vehicle	Intermediate
3	Intermediate	Vehicle	High	Low
4	High	Intermediate	Low	Vehicle

10.3.5.6 The Crossover Design

The crossover design is common in large animal safety pharmacology studies. In such design, each animal receives all the treatments in a randomized sequence, with a washout period between treatments. Such choice of design is mainly driven by the relatively large between-animal variability and smaller sample sizes. Nevertheless, the suitability of the crossover design is also determined by the pharmacokinetics of the test article. If the washout period is not sufficient for the test article to be eliminated from the animal, or in other words, if significant carryover effects are present between treatments, the crossover design is not recommended. For that reason, such a design is not usually adopted for biologics, which tend to have longer half-lives than small molecules.

There are three factors in the crossover design for a typical CV study: treatment (vehicle vs. various dose levels of a test article), period (the day or week when the animal is treated), and animal. Therefore, a Latin square is employed to balance all the factors (Table 10.7). Since a typical study has four doses (vehicle, low, intermediate, and high dose), the 4 × 4 square has become the most commonly used design. It then follows that the treatments take place on four different days and the number of animals is a multiple of four.

An improved design is the Williams square in which the first-order carryover effects are also balanced (The Latin square in Table 10.7 is also a Williams square). In other words, in such a design every treatment immediately follows the other treatments once and only once. Note that there are 24 different 4 × 4 Williams squares, the choice of which needs to be clearly conveyed and agreed upon with the study director. A study that has an odd number of doses requires two different Williams squares to balance the first-order carryover effects.

10.3.6 Power of Safety Pharmacology Studies

It is crucial to understand the power of a particular assay or animal model before applying it to routine studies. An underpowered design would lead to futility, whereas an overpowered one wastes resources, including animals. Moreover, the discriminant power is an important factor when developing a new assay or model that may replace an existing one; comparing the sensitivity of both assays is an important criterion in the decision. We will first discuss the power analysis for continuous variables as it is relatively more straightforward, the basic statistical tool for which is the non-central t- or F-distribution assuming the underlying distribution is normal. Next we will turn to categorical variables, whose power analysis is more complicated.

Besides the assay itself, the statistical design and analysis are also important components of the power analysis and sample size calculation. Since the major safety pharmacology study designs are either crossover or parallel groups, it follows that the comparison between within- and between-animal variability is critical to the experimental design.

In practice, a cohort of recent studies is selected to estimate the biological variability (between and/or within) from each study. To do this, an ANOVA-based model is fit to either the entire set of data or only the control arm of the data. Then, a representative metric (e.g., the median) of all the variability estimates can be used for the power analysis or sample size calculation. Note that the variability estimates from the control data may demonstrate a more identifiable distribution, but their magnitudes could be less than the ones in the treated data, and hence underestimate the overall variability. A remedy for this would be to inflate the estimate by some amount, e.g., using the 75th percentile instead of the median from all the studies.

Whenever baseline data are present, we would recommend doing the power analysis both with and without baseline adjustment and then plotting the power curves of both cases on the same graph. This visual comparison will serve as a useful tool to help educate the scientists on the importance of baseline adjustment and thus the benefits of using ANCOVA. An example of the power plot is given in Fig. 10.9 comparing the power of ANOVA to that of ANCOVA.

As mentioned in Sect. 10.3.5.3, the repeated-measures analysis can be more powerful than the cross-sectional approach. Given that the experimental unit (e.g., animal) is not changed between the two approaches, this claim may seem counter-intuitive. The key factor is the correlation among the observations from the same experimental unit. The benefit of the repeated-measure analysis increases when the correlation becomes smaller. In other words, when the observations from the same animal become more like uncorrelated samples, the repeated-measures analysis, which includes all the observations from the same animal, effectually enlarges the sample size and hence yields more power than the cross-sectional analysis which only takes one sample at a time. More detailed discussion of the statistical power of repeated-measures and split-plot designs can be found in Rochon (1991) and Bradley and Russell (1998).

When discussing power calculations, it may be necessary to remind the scientist that statistical power (or the minimum detectable difference, for that matter) derived in the above manner is a prospective estimate rather than a descriptive metric of the set of studies selected for the power analysis. Therefore, since we do not know the true treatment effect of these studies, one should not use the power or sample size estimate to "verify" that they apply to those studies. In other words, the power is a conditional probability (i.e., the probability that, with a significance test, a study is found to have a treatment effect size greater than a certain value out of all the studies

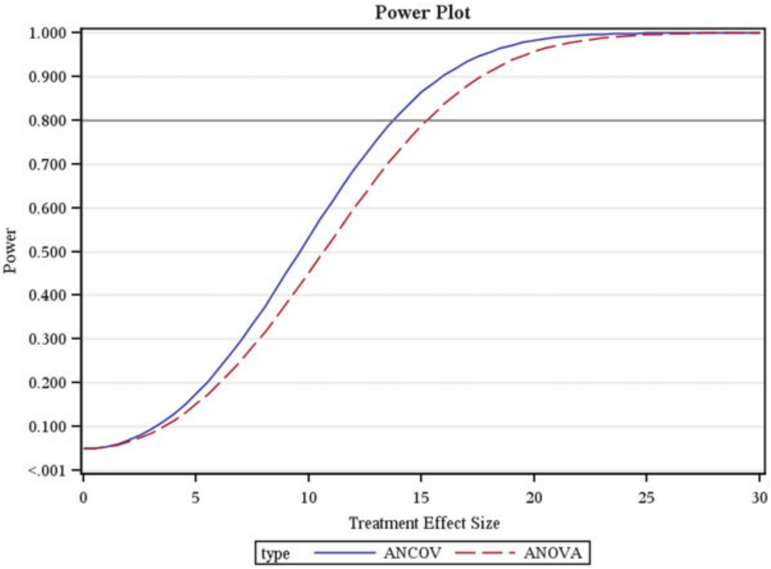

Fig. 10.9 Example of power comparison between ANOVA and ANCOVA models

whose true effect sizes are greater than that value) as opposed to an unconditional one (i.e., of all the studies conducted, regardless of their true treatment effect sizes). Based on our experience, misunderstandings sometimes occurred if such difference had not been clarified beforehand.

In contrast, the power analysis of categorical data from Safety Pharmacology studies can be more complex because: (1) the prior prevalence of some adverse events may not be known with sufficient precision, especially for rare events; and (2) it can be challenging to define a treatment effect “size” for categorical data. Power analysis of categorical data in many cases is based on the non-central distribution of the test statistic. A detailed discussion and examples are provided in Lachin (2011).

10.3.7 Other Issues in Safety Pharmacology Study Analysis

10.3.7.1 Missing Data

When a subset of the data is missing, the ordinary mean of the group differences may no longer be an unbiased estimate of the treatment effect. To address this, the use of fitted means (also known as least-squares means) is recommended for statistical reporting. The concept of fitted means can be challenging to communicate, and is often met with some resistance because the fitted means can differ from the ordinary means. However, the importance of using the fitted means can be

Table 10.8 Example study data from two-period parallel group study

	Period 1	Period 2
Vehicle	2, 5	9, 12, 15
Drug	6, 6, 7, 7	16

clearly explained using an example study with two factors and unequal number of observations at each combination of factors. The hypothetical endpoint in Table 10.8 comes from a single-dose parallel-group study with two treatment periods. The apparent means of the vehicle and drug groups are $(2 + 5 + 9 + 12 + 15)/5 = 8.6$ and $(6 + 6 + 7 + 7 + 16)/5 = 8.4$, respectively, whereas the fitted means of the vehicle and drug groups are $0.5 \times [(2 + 5)/2 + (9 + 12 + 15)/3] = 7.75$ and $0.5 \times [(6 + 6 + 7 + 7)/4 + 16] = 11.25$, respectively. This discrepancy is also known as the Simpson's paradox. The fitted mean takes into account the unequal number of animals in each combination of the factors, and appropriately weights each combination, providing unbiased estimates of the true group means. Missing data in a crossover design pose similar problems, and in some cases can be more difficult to address. This problem cannot be alleviated by simply increasing the sample size.

10.3.7.2 When the Sample Size Is Too Small

For some investigative or exploratory Safety Pharmacology studies, a study design with very small sample sizes may be chosen. This may lead to significantly underpowered studies. Moreover, even at a seemingly moderate sample size, the power of some study designs can be much lower than other designs. For example, a common GLP Safety Pharmacology cross-over design includes four animals, four dose groups, and four periods. Suppose that the underlying within-animal standard deviation of QTc is 4 ms. Then such a design can provide a statistical power of 66 % to detect QTc changes of 8 ms and higher. In contrast, with a 2 × 2 crossover design including four animals, two dose groups and two periods, the power to detect the same magnitude of change is reduced to 46 %.

Baseline adjustment can be also questionable when the sample size is too small. With an extremely small n, there is like to be complete confounding between animal (as a block effect) and baseline. It is also possible that the post-dose data are positively correlated with baseline when all the animals are examined, whereas such correlation becomes negative within each individual animal.

10.3.7.3 Use of Toxicokinetics (TK) Data

It is recommended that the statistician have a good working knowledge of the terminology and basic concepts in TK. Occasionally, TK parameters need to be included in a statistical model to account for variation in exposure levels; they might also be indirectly compared to the statistical results to better understand the correlation between the exposure and treatment effect. Furthermore, TK/PK parameters play an important role in cross-species and preclinical-to-clinical translation studies, as drug effects across different species are more appropriately compared at similar exposures.

10.4 Reproductive Toxicology Studies

10.4.1 Overview

The study of reproductive and developmental toxicology in pharmaceutical development owes its beginnings to the marketing of Thalidomide in the late 1950s (McBride 1961; Weaver and Brunden 1998a). Thalidomide was marketed for respiratory infections, insomnia, coughs, colds, and headaches. Thalidomide was also taken routinely by pregnant mothers to control morning sickness, however at that point in time drug testing during pregnancy was not routinely performed. While Thalidomide was initially considered safe, experts estimate that by the time it was removed from the market in November 1961, Thalidomide use resulted in the birth of more than 10,000 children with serious birth defects, many of whom subsequently died prior to their first birthday. As a result, the hurdles required for drug approval were modified, and the study of reproductive toxicology was added as a requirement for drug approval in many countries throughout the world.

Reproductive toxicology focuses on the study of toxicities associated with the reproductive process. This includes effects on male or female sexual function and mating behavior, female fertility, maternal care, and pup growth and development.

Developmental toxicology, also known as teratology, focuses on the study of toxicities associated with the development of the offspring. This includes effects on fetal development such as birth weight and developmental birth defects. Some consider developmental toxicology to be a subset of reproductive toxicology. The term “Teratology” refers to developmental toxicology, but is sometimes erroneously used to refer to reproductive toxicology as well.

Studies are designed to evaluate both reproductive and developmental toxicity simultaneously when possible. Studies are generally performed in either rats, mice, and/or rabbits. Periodically, studies are performed in non-human primates, but due to cost and sample size issues, these studies are rare and only performed when studies in the usual species will not suffice (e.g., no biological action).

Three standard study types are performed. These study types are outlined in the International Conference on Harmonization (ICH) guidelines, which were originally published in 1992 and most recently modified in 2005 (International Conference on Harmonization 2005). The three study types are defined as follows:

1. Fertility, reproductive performance, and early embryonic development

   This study focuses on the period from prior to mating through implantation of the embryo in the uterus. Both males and females are dosed prior to mating, although sometimes a pair of studies where a single sex is dosed in each is performed. Animals are allowed to mate, and females are sacrificed during gestation at some point after the mid-point of gestation. This study is conducted in a single species, most likely rats.

2. Embryo-Fetal Developmental Toxicity

   This study focuses on the ability of the mothers to successfully carry the pregnancy, on potential birth defects in the offspring, and on toxicity during the period of organ formation and development of the offspring. Females are dosed starting after the implantation of the embryo, and are sacrificed just prior to the end of gestation. This study is conducted in two species, most likely rats and rabbits.

3. Perinatal and postnatal toxicity

   This study focuses on the late stages of pregnancy and on the early stages of the pup's life. Females are dosed starting after implantation and continues through lactation until the pups are weaned. This study is conducted in a single species, most likely rats.

10.4.2 Study Design and Endpoints Collected

Studies typically have a control group and three or four dosed groups. The high dose level is chosen with the intent of inducing a toxic effect at the high dose level on the dosed animal (mother or father). The lowest dose is chosen to have no observable adverse effects, and the middle doses are typically chosen equally spaced between the low and high dosage levels on a log-scale. If a toxic effect can be induced on one of the non-reproductive parameters collected on the parent (e.g. body weight), without having any negative effect on the offspring, then the drug is 'safe' from a reproductive toxicity standpoint. The logic is that the offspring are protected from reproductive toxicities because the parents would never be dosed at a level that high. Therefore, the presence of a treatment effect on a parental, non-reproductive parameter is not a detrimental finding for the study.

The endpoints collected in each of the study types listed above are outlined in Table 10.9.

Note that Table 10.9 is an example of the standard set of parameters collected in each of the study types. Different testing facilities, and even separate studies within a testing facility may modify this list depending on the study design, the anticipated action of the test compound, and other information that may be available during protocol preparation.

10.4.3 Statistical Analysis

In general, the statistical issues discussed previously in the general toxicology section (Sect. 10.2) such as normality of the data, equality of variances and the use of trend testing for dose–response are applicable for reproductive toxicology studies as well.

Table 10.9 Summary of typical endpoints in reproductive toxicity studies

Endpoint	Fertility, reproductive performance, and early embryonic development study	Embryo-fetal developmental toxicity study	Perinatal/postnatal toxicity study	Type of data
Paternal body weight	✓			Continuous
Paternal food intake	✓			Continuous
Maternal body weight	✓	✓	✓	Continuous
Maternal food intake	✓	✓	✓	Continuous
Estrous Cycles	✓			Dichotomous and/or continuous
Mating Performance	✓			Continuous (time to mate) and dichotomous (mating, fertility indices)
Hysterotomy findings	✓	✓		Counts and/ or proportions
Fetal weight		✓		Continuous
Fetal Examination Findings (fetal malformations, variations, and retardations)		✓		Proportions, and/ or dichotomous
Gestation length			✓	Continuous
Pup findings at birth (live, dead)			✓	Counts and/ or proportions
Pup survival			✓	Proportions
Pups with milk in stomach			✓	Proportions
Pup weight			✓	Continuous

In addition, the issue of defining the appropriate experimental unit is critical for reproductive studies. The experimental unit is defined as the unit to which the treatment is applied (Chen 1998, 2000; Palmer 1974; Selwyn 1988; Staples and Haseman 1974; Weaver and Brunden 1998b). For reproductive studies, this is either the mother or the father. However, many of the parameters collected in these studies are collected on the offspring. It is well established that litter-mates are more similar than pups from different litters. Failure to account for the intra-litter correlation by treating the offspring data as independent experimental units will inflate the Type I error rate and result in an invalid statistical analysis. For some data types, the solution may be simple (e.g. a nested ANOVA design). However, for many of the data types collected in these studies (e.g. count data), the appropriate statistical approach may not be obvious.

From a statistical standpoint, another topic of interest is the varied data types that require analysis. In a general toxicity study, most parameters are continuous in nature. A reproductive toxicity study, on the other hand, is likely to contain continuous (e.g. body weights), count (e.g. number corpora lutea per litter), proportions (e.g. resorptions per litter), and dichotomous (e.g. mating index) data. Some standard statistical approaches generally utilized for each of these data types are outlined below.

10.4.3.1 Statistical Analysis of Continuous Endpoints

Most maternal and paternal endpoints (e.g., body weight, food intake, time to mate, gestation length, estrous cycle length) are continuous in nature. These parameters can be analyzed using the methods outlined previously in the general toxicology discussion.

Some parameters measured on the offspring (fetal weight, pup weight) are continuous in nature. However, because these parameters are not collected on independent animals, the analysis must account for this fact. This can be accomplished by the use of a nested ANOVA, using the litter as a factor in the model. Alternatively, sometimes a litter average is generated, and the standard general toxicology methods are applied to the litter averages.

In addition, for pup weight and fetal weight, the weight of the offspring is dependent on the size of the litter (Chen and Gaylor 1992; McCarthy 1967). In general, larger litters have smaller offspring. This can be addressed through the use of a covariate for the litter size in the model.

10.4.3.2 Statistical Analysis of Counts

Many of the parameters collected are counts of the number of occurrences per litter. Some examples of these are the number of corpora lutea per litter, number of implants, number of resorptions, number dead pups, and the number of live per litter.

It is important to note that for some of these parameters, there is a high percentage of zeroes among the litters in a study.

Count data can be analyzed using the generalized estimating equation approach (Liang and Zeger 1986; Zeger and Liang 1986). This approach is outlined in detail in Chen (1998). However, in some cases where there is a high proportion of 0 observations, the iterative fitting process required for this model may not converge to a solution. Because of this, count data are often analyzed using nonparametric rank-based methods such as a Kruskal–Wallis test or the Jonckheere trend test.

10.4.3.3 Statistical Analysis of Proportion Data

Many of the parameters collected are proportions of responders per litter. Whenever possible, the proportion per litter is a better endpoint for analysis than counts, since counts can also be influenced by changes in the litter size due to competing drug effects (Bailey 2008; Chen 1998). Some examples of these are proportion resorbed per litter, proportion live per litter, and proportion with malformations per litter (either individual malformations or some form of a combined category). In addition, similar to the count parameters, some parameters have a high percentage of zeroes among the litters in a study.

A common approach to the analysis of proportion data is to use a beta-binomial model (Williams 1975). Alternatively, this data can be analyzed using a generalized estimating equation approach (Chen 1998). Alternatively the data can be transformed using an arc-sine transformation and analyzed using standard ANOVA methods. Finally, a nonparametric rank-based approach such as a Kruskal–Wallis test or the Jonckheere trend test are often utilized for their robustness, simplicity and understandability.

10.4.3.4 Statistical Analysis of Dichotomous Endpoints

Parameters with dichotomous response data are generally maternal parameters, such as estrous cycles (cycling normally), and mating and fertility indices. Also, sometimes the presence of fetal examination findings in a litter (any fetus with a finding) is analyzed as dichotomous data.

Dichotomous data are expressed as a $2 \times C$ contingency table, where C is the number of groups in the study. The data are summarized as the percent responding per dosage group, and standard contingency table methods, such as the Chi-square test, Cochran–Armitage test, or the Fisher Exact Test are used for the analysis. Ciminera proposed a randomization test for dose–response trend that is an extension of the Fisher Exact test for histopathology data (Ciminera 1985) that would be applicable here as well.

10.4.3.5 Multivariate Approaches

Over the years there have been a number of publications proposing simultaneous analysis of multiple endpoints (Catalano et al. 1993; Catalano and Ryan 1992; Chen et al. 1991; Ryan 1992). One of the justifications for this type of approach is to reduce the number of analyses being performed on the study, thereby reducing the overall study-wide Type I error. This type of approach can also increase the power of detecting effects if the outcomes under consideration are due to a common biological mechanism.

These approaches have seen limited implementation in the pharmaceutical industry. The study scientists are reticent to identify parameters that they feel can be grouped for analysis without also considering the analysis results from each of the component analyses. Therefore, this would lead to more analyses being performed on a study, rather than less. Also, if a significant result is found, the study scientists want to know what components of the combination contributed to the significant result. While these approaches are nice in theory, they are too complex for many study scientists to understand and embrace.

10.5 Juvenile Toxicology Studies

Prior to 1994, drug testing was generally performed on adult animals in toxicity studies, and on adult subjects in clinical trials. If a drug was approved, it was then often prescribed to children, generally at lower doses, under the assumption that children would respond similarly to the compound as adults. However, due to differences in pharmacology, and differences in development (e.g. the human brain at birth is neurologically equivalent to a rat brain at 2 weeks of age), this assumption is often not true.

In 1994, in order to address this gap, the FDA encouraged manufacturers to survey existing data of marketed compounds to determine if there was sufficient evidence to support pediatric use information in the drug's label. When this did not markedly increase the number of products with adequate pediatric labeling, the FDA created a regulation, approved in December 1998, requiring that pediatric safety and effectiveness be addressed either through studies or through a waiver in NDAs and BLAs. At that point in time, Juvenile toxicology studies to support pediatric clinical trials started to become more common. The regulations were invalidated in 2002 by the federal courts, but then resurrected by Congress as the Pediatric Research Equity Act (PREA) in December 2003. The PREA was then renewed by Congress in 2007. Conduct of the nonclinical aspects of a pediatric program are governed by FDA guidelines issued in 2006 (U.S. Food and Drug Administration 2006).

Juvenile toxicology studies are similar in scope to regulatory toxicity studies, with the difference in the age of the animals at the initiation of dosing. The age and duration of dosing is dependent on the anticipated age of the target human population, and should correlate with the corresponding stage of development in

Table 10.10 Litter-mate intraclass correlation for selected parameters

Parameter	Intraclass correlation (%)[a]
Pup weights during weaning (days 0–21)	69–72
Clinical Pathology at weaning	19–55

[a]Intraclass correlation is the proportion of the variance that can be attributed to which litter the pup is from

the test animals (Zoetis and Walls 2003). Therefore it is possible that dosing may start as early as 1 day of age. This causes problems because there is the temptation, to minimize animal usage for ethical reasons, to treat the animals as independent units where in reality they are not. Examples of litter-mate intraclass correlation for various parameters, based on historical data, are listed below in Table 10.10. Due to this intraclass correlation, the total number of animals required to maintain sufficient power in the study is greater than a study where adult animals are dosed (Bailey 2006). For example, for a parameter that has a 50 % litter-mate intraclass correlation, a study with 40 pups in 10 litters (4 pups/litter) has an ‘effective’ sample size of 16. This means that these 40 pups give the same amount of information as 16 pups if the pups came from 16 independent litters.

In addition to the usual parameters collected in a regulatory toxicology study, neurobehavioral parameters are sometimes collected if there is reason to believe the test compound might affect brain function or development. The same issues listed above apply for these parameters as well. An effective analysis of these parameters can be difficult, because of the small sample sizes and the large variability inherent in these parameters.

Parameters collected in these studies are continuous in nature. Therefore, the methods outlined in the Reproductive Toxicology section for continuous parameter are also applicable for the analysis of Juvenile Toxicology studies.

References

Bailey SA (1998) Subchronic toxicity studies. In: Chow S-C, Liu J-P (eds) Design and analysis of animal studies in pharmaceutical development. Marcel Dekker, New York, pp 135–195

Bailey S (2006) Design and analysis issues in juvenile animal toxicity studies for pharmaceutical development: a statistician’s perspective. Poster presented at the teratology society meetings, Tucson, AZ, 26 June 2006. Abstract: Birth Defects Research 76, p 383

Bailey S (2008) Relationships between litter size and resorptions, dead, and live fetuses: implications for statistical analysis and data interpretation. Poster presented at the teratology society meetings, Monterey, CA, 30 June 2008. Abstract: Birth Defects Research 82, p 352

Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD (1972) Statistical inference under order restrictions. Wiley, New York

Bradley DR, Russell RL (1998) Some cautions regarding statistical power in split-plot designs. Behav Res Methods Insturm Comput 30(3):462–477

Bretz F, Hothorn LA (2003) Statistical analysis of monotone or non-monotone dose–response data from in vitro toxicological assays. Altern Lab Anim 31:81–96

Catalano PJ, Ryan LM (1992) Bivariate latent variable models for clustered discrete and continuous outcomes. J Am Stat Assoc 87:651–658

Catalano PJ, Scharfstein DO, Ryan LM, Kimmel CA, Kimmel GL (1993) Statistical model for fetal death, fetal weight, and malformation in developmental toxicity studies. Teratology 47: 281–290

Chen J (1998) Analysis of reproductive and developmental studies. In: Chow S, Liu J (eds) Design and analysis of animal studies in pharmaceutical development. Marcel Dekker, New York, pp 309–355

Chen J (2000) Reproductive studies. In: Chow S (ed) Encyclopedia of biopharmaceutical statistics. Marcel Dekker, New York, pp 445–453

Chen J, Gaylor D (1992) Correlations of developmental end points observed after 2,4,5-trichlorophenoxyacetic acid exposure in mice. Teratology 45:241–246

Chen JJ, Kodell RL, Howe RB, Gaylor DW (1991) Analysis of trinomial responses from reproductive and developmental toxicity experiments. Biometrics 47:1049–1058

Ciminera J (1985) Some issues in the design, evaluation, and interpretation of tumorigenicity studies in animals. In: Proceedings of the symposium on long-term animal carcinogenicity studies: a statistical perspective. American Statistical Association, Washington, DC, pp 26–35

Clinical and Laboratory Standards Institute (2000) How to define and determine reference intervals in the clinical laboratory: approved guideline, vol 2. CLSI, CLSI document C28-A2, Wayne

Crump K (1984) A new method for determining allowable daily intakes. Fundam Appl Toxicol 4:854–871

Dorato MA, Engelhardt JA (2005) The no-observed-adverse-effect-level in drug safety evaluations: use, issues, and definition(s). Regul Toxcol Pharmacol 42:265–274

Dunnett CW (1955) A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 50:1096–1121

EP Evaluator (2005) Release 7 (EE7). Computer software for evaluating clinical laboratory methods. David G. Rhoads Associates, Kennett Square

Filipsson AF, Sand S, Nilsson J, Victorin K (2003) The benchmark dose method—review of available models, and recommendations for application in health risk assessment. Crit Rev Toxicol 33(5):505–542

Healey GF (1999) The F1 approximate parametric test for monotonicity. Internal Technical Information Document ST9725, Department of Statistics, Huntingdon Life Sciences, Huntingdon

International Conference on Harmonization (2001) S7A – Safety pharmacology studies for human pharmaceuticals. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm074959.pdf. Accessed 1 Sept 2014

International Conference on Harmonization (2005) S5(R2) – detection of toxicity to reproduction for medicinal products & toxicity to male fertility. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Safety/S5_R2/Step4/S5_R2__Guideline.pdf. Accessed 1 Oct 2014

International Conference on Harmonization (2010) M3(R2) – nonclinical safety studies for the conduct of human clinical trials and marketing authorization for pharmaceuticals. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm. Accessed 30 June 2014

Jonckheere AR (1954) A distribution-free k-sample test against ordered alternatives. Biometrika 41:133–145

Lachin J (2011) Power and sample size evaluation for the Cochran–Mantel–Haenszel mean score (Wilcoxon rank sum) test and the Cochran–Armitage test for trend. Stat Med 30:3057–3066

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

McBride WG (1961) Thalidomide and congenital abnormalities. Lancet 2:1138

McCarthy JC (1967) Effects of litter size and maternal weight on foetal and placental weight in mice. J Reprod Fertil 14:507–510

Palmer AK (1974) Statistical sampling and choice of sampling units. Teratology 10:301–302

Rochon J (1991) Sample size calculation for two-group repeated-measures experiments. Biometrics 47:1383–1398

Ryan L (1992) Quantitative risk assessment for developmental toxicity. Biometrics 48:163–174

Selwyn MR (1988) Preclinical safety development. In: Peace KE (ed) Biopharmaceutical statistics for drug development. Marcel Dekker, New York, pp 231–271

Shirley E (1977) A nonparametric equivalent of Williams' test for contrasting existing dose levels of a treatment. Biometrics 33:386–389

Senn S (2007) Statistical issues in drug development. Wiley, Chichester, pp 99–100

Sivarajah A1, Collins S, Sutton MR, Regan N, West H, Holbrook M, Edmunds N (2010) Cardiovascular safety assessments in the conscious telemetered dog: utilisation of super-intervals to enhance statistical power. J Pharmacol Toxicol Methods 62(1):12–19

Sparrow S, Robinson S, Bolan S, Bruce C, Danks A, Everett D, Fulcher S, Hill RE, Palmer H, Scott EW, Chapman KL (2011) Opportunities to minimize animal use in pharmaceutical regulatory general toxicology: a cross-company review. Regul Toxicol Pharmacol 61:222–229

Staples RE, Haseman JK (1974) Selection of appropriate experimental units in teratology. Teratology 9:259–260

Tukey JW, Ciminera JL, Heyse JF (1985) Testing the statistical certainty of a response to increasing doses of a drug. Biometrics 41(1):295–301

U.S. Food and Drug Administration (2006) Guidance for industry – nonclinical safety evaluation of pediatric drug products. U.S. Food and Drug Administration, Rockville

Weaver J, Brunden M (1998a) Design of developmental and reproductive toxicity studies. In: Chow S, Liu J (eds) Design and analysis of animal studies in pharmaceutical development. Marcel Dekker, New York, pp 291–308

Weaver J, Brunden M (1998b) The design of long term carcinogenicity studies. In: Chow S, Liu J (eds) Design and analysis of animal studies in pharmaceutical development. Marcel Dekker, New York, pp 227–258

Williams DA (1971) A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics 27:103–117

Williams DA (1972) The comparison of several dose levels with a zero dose control. Biometrics 28:519–531

Williams DA (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31:949–952

Zeger SL, Liang KY (1986) Longitudinal data for discrete and continuous outcomes. Biometrics 42:121–130

Zoetis T, Walls I (eds) (2003) Principles and practices for direct dosing of pre-weaning mammals in toxicity testing and research. A report of the ILSI risk science institute expert working group. ILSI, Washington DC