Available online at www.sciencedirect.com
ScienceDirect
Neurotoxicology and Teratology 30 (2008) 326–348
NEUROTOXICOLOGY AND TERATOLOGY
www.elsevier.com/locate/neutera

Statistical issues and techniques appropriate for developmental neurotoxicity testing ☆
A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints

R. Robert Holson [a,*], Les Freshwater [b], Jacques P.J. Maurissen [c], Virginia C. Moser [d], Whang Phang [e]

[a] *Department of Psychology, New Mexico Tech, Socorro, NM, United States*
[b] *BioSTAT Consultants, Portage, MI, United States*
[c] *The Dow Chemical Company, Midland, MI, United States*
[d] *Neurotoxicology Division, NHEERL, US EPA, RTP, NC, United States*
[e] *OPPTS, US EPA, Washington, DC, United States*

Received 8 January 2007; received in revised form 29 May 2007; accepted 7 June 2007
Available online 15 June 2007

Abstract

The data from developmental neurotoxicity (DNT) guideline studies present a number of challenges for statistical design and analysis. The importance of specifying the planned statistical analyses *a priori* cannot be overestimated. A review of datasets submitted to the US Environmental Protection Agency revealed several inadequate approaches, including issues of Type I error control, power considerations, and ignoring gender, time, and litter allocation as factors in the analyses. Since DNT studies include numerous experimental procedures conducted on the dam and offspring at several ages, it is not unusual to have hundreds of significance tests if each was analyzed separately. Two general approaches to control experiment-wise Type I inflation are: 1) statistical/design considerations that reduce the number of *p*-values, including factorial designs, multivariate techniques, and repeated-measures analyses; and 2) adjustments to the α level, including newer approaches that are less conservative than, for example, Bonferroni corrections. The design of the DNT study includes testing of both sexes, and gender must be included in the statistical analysis for the determination of sex-related differences, and, indeed, including both sexes may increase power. The influence of litter must be taken into account in the allocation of test animals as well as the statistical analyses. This manuscript reviews many key considerations in the analysis of DNT studies with recommendations for statistical approaches and reporting of the data.

Keywords: Assumptions; Hypothesis generating; Hypothesis testing; Litter effect; Multiple comparison procedures; Power; Repeated measures; Sex effect; Type I and II errors

1. Introduction

Developmental neurotoxicity (DNT) studies typically include a number of evaluations of both the dam and the offspring. Appropriate statistical analyses of the behavioral data collected in the context of these DNT studies can be challenging. The experimental design often consists of multiple measures for each animal, as well as repeated testing across time. Other complications include the use of littermates for the same or different tests, and the need to account for the influence of such genetic and maternal factors. The type of data collected varies with the

☆ The information in this document has been funded in part by the U.S. Environmental Protection Agency. It has been reviewed by the National Health and Environmental Effects Research Laboratory and approved for publication. The approval does not signify that the contents necessarily reflect the views of the Agency nor does the mention of trade names or commercial products constitute endorsement or recommendation for use.
* Corresponding author. Tel./fax: +1 505 835 5862.
E-mail address: rholson@nmt.edu (R.R. Holson).

0892-0362/$ - see front matter © 2008 Published by Elsevier Inc.
doi:10.1016/j.ntt.2007.06.001

behavioral test, including continuous, ordinal, and binary data. We set out, in this paper, to examine the statistical approaches currently being used in DNT testing, which most often follows the US Environmental Protection Agency (US EPA) DNT Test Guideline [62]. We then provide a discussion of the statistical approaches to use in such studies, emphasizing the importance of key issues, especially those which are not currently included. The purpose of this paper is not to provide or suggest specific operational procedures for statistical analysis of each type of DNT test, but rather: 1) to emphasize key issues and concerns regarding the diverse (and sometimes incorrect) methods that are currently used as standard practice; 2) to describe in general terms some of the available methods that might be more appropriate for the data being analyzed (e.g., univariate ANOVA, MANOVA, mixed models, GEE); and 3) to discuss general or overall considerations that are important to consider when interpreting data (e.g., hypothesis generating vs. testing, multiple comparisons, sphericity).

2. Survey of current practice

To determine the current statistical practices, we collected multiple studies from six different testing laboratories. A few of these studies were published in the open literature [1,8,19,42,69,70,72] and the rest were submitted to the US EPA in support of pesticide registrations. It quickly became apparent that each of these laboratories adopted the same set of statistical procedures for all studies they conducted, but these procedures were different among laboratories. Four of these five laboratories began with a test for variance homogeneity, typically Bartlett's test, which determined whether parametric or nonparametric analyses would be used. One laboratory made use of transformations (e.g., log, square root) to achieve normality. Nonparametric methods were used, such as Kruskal–Wallis, Fisher's Exact Test, followed by Mann–Whitney *U*-tests. Parametric data were analyzed with ANOVAs in half the laboratories, the other half used some sort of linear trend test, followed by a variety of *post-hoc* tests (e.g., Williams, Dunnett's).

In almost all studies evaluated, males and females were analyzed separately. Some endpoints were not statistically considered, and a few laboratories stated that the data were visually inspected before deciding which to analyze. Most laboratories tried to allocate equally the offspring to the various behavioral tests, often choosing one male and one female from each litter for each test; however, the caveat was always "when possible." The number used for different tests varied by the test. Thus, it would not be clear how balanced the litter allocation actually was. For example, one report described "one male and/or one female from each litter (13–16/sex/dose, representing at least 18 litters per dose)." It is difficult, if not impossible, to know how many of the pups were actually siblings. Only one laboratory stated that a nested analysis of variance model was used. While the other laboratories did not mention this one way or another, there is evidence from reviews and other reports that at least one laboratory routinely did not use litter as a nested variable. Overall, the major deficiencies that were routinely observed revolve around:

1. Type I and II error considerations,
2. litter allocation and analysis,
3. analysis using sex as a factor,
4. analysis using repeated measures,
5. statistical analysis assumptions.

We identified a number of statistical principles and issues which were prominent in all these studies. A similar, but more rigorous, re-examination of endocrine disruption data was undertaken by Haseman et al. [22]. They identified important guidelines regarding appropriate design and analysis of studies which are as applicable to these studies. Specifically, the issues were divided into: 1) experimental design issues, such as power, replication, litter allocation, potential investigator bias, quality control and control groups, and 2) data analysis issues, such as choice of statistical methodology, heterogeneity of the data, covariance, and regression versus ANOVA, biological interpretation and data selectivity [22]. The following provides background and advice on these issues with regards to DNT studies.

3. Study design and analysis plan

3.1. General considerations

3.1.1. Hypothesis testing vs. hypothesis generating

Conceptually, two types of experimental studies can be contrasted: hypothesis testing (confirmatory) and hypothesis generating studies (exploratory). In the former, a hypothesis (or a number of them) is (are) stated; in the latter, data are generated without any specific preconceived hypothesis. In the former, a conclusion is derived, whereas in the latter, a hypothesis is proposed and another study (on a new data set) has to be designed to test it. Other more stringent conditions also define hypothesis testing vs. generating (see below).

As Muller et al. [45] stated, a hypothesis testing study practically requires that a detailed written protocol be developed with a clear statement of purpose, study objectives, research questions and hypotheses, methods, study design and specific statistical analysis information. Any deviation from a preset analysis plan (e.g., different statistical tests, further transformations, different data grouping or data adjustments, additional variables, different error rate criteria, new hypothesis evaluation, etc.) will make such a study a hypothesis-generating study if a decision has been made after examining the data. In the context of a hypothesis-generating study, the Type I error rate can be set at higher values (e.g., 0.2) than is conventionally done. Most re-analyses of existing data fall within the realm of hypothesis-generating studies (except if a detailed and preset analysis protocol is available before the data are examined); the conclusive value of such an analysis would be questionable at best. To say the least, it is misleading to report a hypothesis-generating analysis as if it were hypothesis-testing.

Looking at data after the fact in a number of different ways can often show some unanticipated (and sometimes nonexistent)

relationship among variables. An interesting illustration is given by Freedman [18] who ran a multiple regression on 51 columns of random data (one of them being arbitrarily designated as the "dependent variable") and a p-value of 0.53 was obtained. After selected columns with small regression coefficients were removed, a multiple regression was rerun and provided a p-value of 0.0005. This paper illustrates one example of the potential danger of creating something from nothing when data are reanalyzed in light of previous analyses. If this practice is acceptable in hypothesis-generating studies where no conclusion is drawn, it is not acceptable in hypothesis-testing studies.

One could wonder whether the DNT study run as per guideline is a hypothesis-generating or a hypothesis-testing study. On one side, it can be argued that there can be no hypotheses, given that the same standard testing is conducted irrespective of the chemical; on the other side, the DNT proposes areas of study and asks standard questions about a fixed number of endpoints (e.g., motor activity, learning and memory). The crux of the problem lies in the definition of "hypothesis". In the present context, it is argued that the DNT qualifies under hypothesis-testing studies, as far as all the criteria stated above are ascertained [45].

3.1.2. Multiplicity

The US EPA DNT study guideline requires the use of a number of procedures (15 or more). In each of these procedures, a number of dependent variables can be identified (e.g., number of brain areas to be examined for morphometrics). Some of these variables are in turn examined on a number of occasions, whether within or across test sessions (e.g., motor activity). The data collection takes place in two sexes and generally four dose groups (including control). Finally, all of these data are analyzed and a large number of p-values are generated (simple main effects and interaction terms). Table 1 illustrates the "multiplicity" problem by assuming that each treated group is compared with the control group for each variable at each time point. The number of 1302 derived p-values is only given as an example, and does not imply that it constitutes a recommended analysis method.

The multiplicity problem can be encountered at several levels. It is present in the context of group comparisons; for example, a number of papers have reported multiple t-tests in the evaluation of the statistical significance of a difference between treatment groups, where the accepted Type I error was set to 0.05 for each comparison. Multiple comparison procedures that maintain the overall Type I error rate are discussed below.

At a higher level, the analysis of variance (e.g., main effect of treatment, treatment-by-sex interaction) also controls the Type I error rate across the different groups and across factors (multiway ANOVA). This procedure contributes to a decrease in the number of generated p-values. At a still higher level, the multiplicity problem also originates from the fact that the previous analyses can be repeated over a number of dependent variables. Multiple ANOVAs present the same challenge as multiple t-tests. Most often, however, no correction is made for the numerous p-values that are generated by the repeated use of statistical analyses. This problem is also considered below.

Table 1
Example of numbers of potential p-values in a typical DNT test

Dependent variables	Testing times	# Time points	# p-values [a]
Dams			
Gestation body weight	GDs 0, 6, 10, 15, 20	5	3
Lactation body weight	LDs 0, 4, 11, 13, 17, 21	6	36
Observations	GDs 10, 18; LDs 6, 13	4	288
Pups			
Developmental landmarks			
Pup count	PNDs 0, 4, 11, 17, 21, 35, 60	7	42
Body weights	PNDs 0, 4, 11, 17, 21, 35, 60	7	42
Vaginal patency	Between 28 and 42	1	6
Preputial separation	Between 35 and 52	1	6
Behavioral tests			
Observations	PNDs 4, 11, 21, 35, 45, 60	6	432
Auditory startle amplitude	PNDs 22 and 61	2	12
Auditory startle habituation	PNDs 22 and 61	2	60
Motor activity (total counts)	PNDs 13, 17, 21, 60	4	24
Motor activity adaptation	PNDs 13, 17, 21, 60	4	144
Learning (latency)	PNDs 22 and 61	2	60
Memory (latency)	PNDs 22 and 61	2	60
Neuropathology			
Brain weights (absolute and relative)	PNDs 11 and 60	2	24
Morphometry	PNDs 11 and 60	2	36
Total			1302

Other assumptions:
Observations (# signs) 12.
Auditory startle (# within-session blocks) 5.
Motor activity (# within-session bins) 6.
Learning and memory (# test days) 5.
Morphometrics (minimum # of areas) 3.
[a] General assumption: 2 sexes and 3 comparisons.

One of the consequences of the "multiplicity" problem is that it increases the probability of false declarations of an effect. Controlling the Type I error rate at $\alpha = 0.05$ per comparison indicates that there are 5 chances out of 100 that an effect be declared when it is not present, given that one comparison is made. When the number of derived p-values increases, the probability of a false positive increases. If all the data were independent, the overall error rate (also referred to as α_{ew} for experiment-wise error rate) could be calculated with the following formula:

$$\alpha_{ew} = 1 - (1 - \alpha_c)^n$$

where α_c is the accepted error rate per comparison (e.g., 0.05) and n is the number of derived p-values [23], p. 611. For 10 and 100 p-values, the overall probability of falsely declaring at least one effect (α_{ew}) statistically significant is 0.40 and 0.99, respectively, under conditions of independence, while an α_c of

0.05 is maintained per comparison. Because of the lack of independence of data, the actual error rate falls somewhere between the per-comparison error rate and the theoretical experiment-wise error rate. It is very important to provide the reader with the total number of derived *p*-values when reporting the results of a study; e.g., if three statistically significant *p*-values are reported in a study, their meaning will be very different whether a total of 5 or 50 *p*-values had been derived in the study. It is therefore misleading to report only the significant *p*-values without at least giving the reader a very good estimate of the total number of derived *p*-values, whether significant or not.

It is important to recognize that looking summarily over a data set to see if there appears to be any effect constitutes *ipso facto* an implicit analysis. For example, with an ordinal dataset of observations (e.g., scores of 1–5), it is easy to survey the results and see that the individual scores of the control and experimental groups are mostly 3's, and the investigator might conclude by experience that the results would not be significant, were they to be statistically analyzed, and therefore the investigator may decide not to formally "analyze" them. However, if a number of 1's were to be detected in the high-dose group to a slightly greater proportion than in the control group, the investigator might initiate a formal statistical analysis and find out that the difference between control and high-dose group was (or was not) statistically significant. In both cases, an analysis had been conducted, either implicitly or explicitly.

3.1.3. p-values

The null hypothesis can always be rejected. Such a statement is meant to reflect the fact that, in a randomized study designed to compare multiple treatment groups, a statistically significant *p*-value can always be obtained if the sample size is large enough. A statistically significant *p*-value simply indicates that some relation exists between two or more variables, but does not provide any indication about the strength of association between these variables, no matter how low the *p*-value is. For example, it is not unusual to see a low correlation coefficient (e.g., 0.12) associated with a significant *p*-value of 0.001 in studies with large sample sizes. In other words, a low *p*-value *per se* does not mean that a useful degree of association exists. In such a situation, the investigator should realize that great importance should not be attributed to the significant *p*-value in light of the low correlation coefficient. A number of indices of strength of association have been proposed (e.g., R^2, η^2, ω^2) [23], pp. 413–422 and should be considered to evaluate the degree of relation between variables. Confidence intervals and prediction intervals can also be useful, especially in a graphical context. When the strength of association is low (no matter what the *p*-value might be), it indicates that little can be predicted from the independent variable to the dependent variable, i.e., that the variation in *X* does not explain much of the variation of *Y*.

For historical reasons, *p*-values have been expressed as inequalities by reference to a criterion (α), typically as $p<0.05$. Now that exact *p*-values can be provided for the majority of procedures by most standard statistical packages, it is suggested that they be reported as such, e.g., $p=0.08$ [67]. To help the reader better understand what has actually been done in an analysis, it is also very helpful to report the *F* value and its associated degrees of freedom in addition to the exact *p*-values, as appropriate.

3.2. Controlling experiment-wise Type I error rate

The multiplicity problem (i.e., analysis of a large number of dependent variables), though not always clearly recognized, results in the false declaration of effects and can be addressed at the level of the design and of the analysis of the study, e.g., by decreasing the total number of *p*-values, and/or by adjusting the α-criterion. A few miscellaneous examples are offered below for consideration (see also Section 3.8.5).

One should always define the questions in specific terms for each procedure and carefully choose the dependent variables that are going to answer them. It is important to distinguish between "primary" variables specifically collected to answer the question under investigation, and "ancillary" variables collected to help interpret some aspects of the results. Ancillary variables may be collected to assure that the test is functioning properly, but will not by themselves answer the question of interest. Consider, for example, a delayed matching-to-position procedure for the evaluation of short-term memory. The operant chamber has two retractable levers and a feed cup on the opposite wall. The rat is trained to press the extended lever which is then retracted, and the rat has to spend a variable delay with the snout in the feed cup on the opposite wall. After a delay is over, both levers are extended into the cage and the rat has to press the previously extended lever. Occasionally, the rat leaves the feed cup during the delay and presses the retracted (but still accessible) lever. This behavior, referred to as "rehearsal", typically increases the probability of correct responding, but should not be interpreted as reflecting improved memory. Such rehearsals are ancillary variables in the sense that they do not measure the endpoint of interest, but they can help in the interpretation of the primary retention data. These ancillary data need not be statistically analyzed if the focus of the test is to make a statement about the effects of a test substance on "memory". Whereas primary data should normally be statistically analyzed, ancillary data should not necessarily be.

The multivariate analysis also constitutes a way of reducing the total number of derived *p*-values by analyzing at the same time different dependent variables. It also has the advantage of being sensitive to trends among the different variables. The data can also be analyzed following a conditional scheme so that, for example, some analyses are only performed when some previous analysis is statistically significant. Since such uses of multivariate analyses are fraught with both design and interpretation issues, it is recommended that this option only be adopted with the help of a qualified statistician.

Violating some statistical assumptions can also contribute to the false declaration of effects, for example, when the sphericity assumption (see Section 3.4.2) is violated in a repeated-measures ANOVA [44,63] or when the litter is not used as the unit of statistical analysis (see Section 3.6).

3.3. Power and Type II error rates

Power is defined as the probability of not committing a Type II error, also known as a false negative error. That is, power is the statistical probability of correctly identifying a real effect, such as some side effects of a drug under development. Conversely, a Type II error involves not detecting the real drug effect or side effect, when one exists.

Power has long been a severely neglected aspect of hypothesis testing and experimental design. This is a rather mystifying oversight, since it incorporates a bias which is especially harmful for regulatory research.

Introductory statistics courses teach the convention of $\alpha < 0.05$. Setting the probability of a Type I error (α) at this level is simply a practice not dictated by the cost of such errors, and of course varies according to many factors. Despite the general acceptance of $\alpha < 0.05$, there is no general convention governing the protection levels desirable for Type II errors (β); however, committing a Type II error in regulatory research can have results which are devastating. It is clear that marketing drugs or environmental products which later turn out to have negative health consequences can have severely injurious human, environmental and economic consequences.

One might expect that companies, academia and regulatory agencies would agree that protection levels against Type II errors should be set at least equal to the $\alpha < 0.05$ levels used to protect against the arguably less-costly Type I errors. This would require experimental designs with a power ($1-\beta$) of 0.95. It is possible that such power levels are almost never achieved in animal research, even in a regulatory context. The reason is simple — this degree of protection against Type II errors is simply impractical, in that it requires more animals per treatment group than is normally affordable or acceptable. Consequently, it is becoming common practice to set power at or about 0.80, an error level 4 times higher than that tolerated for Type I errors [10,31,46].

If cost factors and animal use are potential explanations for the lack of a consensus on desirable power levels, the problem of effect sizes is certainly another explanation. Protection against Type I errors is conceptually rather simple: we assume that in reality there are no experimental effects, hence effect size is always set at zero. Power calculations are more demanding. A given power relates not just to some (hopefully plausible) β level. It is also based on some "adequate" effect size, selected more or less arbitrarily from an infinite range of possible effect sizes. Like β itself, there is no general answer to the question of how large an "adequate" effect size should be. Furthermore, it must be emphasized that decisions regarding "meaningful" effect sizes are not statistical questions, left to consulting statisticians. Rather, adequate effect size must be based on knowledge of the physiology of the test system. For example, a 30% drop in body weight can be lethal, and is of greater concern than a 30% drop in open field activity. Consequently, a test laboratory might choose different effect sizes for power calculations for different variables, based on such considerations.

It is practically unlikely that we set effect sizes according to a specific research problem. Instead, there are numerous published attempts to define "small", "medium" and "large" effect sizes. The most familiar of these utilizes Cohen's d [10]. In a simple t-test, 2-sample example with equal n, d is just the difference between the two means divided by the standard deviation of the population rather than the sample (thus dividing by n rather than $n-1$). Cohen's d is thus essentially a z score. In a simple t-test situation with two samples and equal n, Cohen suggests that a "small", "medium" and "large" effect would be in the neighborhood of $d = 0.2$, 0.5, and 0.8, respectively. To anchor this in concrete examples, a d of 0.5 is the difference in height between 14- and 18-year-old girls [10].

Regrettably, effect sizes are less intuitive when the concept is applied to the typical DNT design. Here we commonly deal with complex factorial designs, with at least a single untreated control and three increasing dose groups. In this case Cohen [6] calculates a slightly different measure of effect size, the f ratio (in the limiting case of 2 conditions, $f = d/2$). Cohen's f (not to be confused with Fisher's F ratio) is defined as the square root of the between-groups variance σ_m (but divided by k, the number of groups, not $k-1$ as in the traditional ANOVA), and divided by the square root of the ANOVA mean square error (σ).That is, $f = \sigma_m/\sigma$. Here Cohen defines an f of less than 0.1, 0.25 and 0.4 as small, medium and large, respectively [10].

To further complicate matters, in this situation the concept of effect size is dependent on how some "true" experimental effect increases with increasing dose. That is, for a given dependent variable we can have several different patterns of effects and each of which will produce a different value for σ_m. For example, one could wish to detect a difference between control and high-dose group, or a step-wise dose-response relationship, such as control > low > medium > high dose, or another pattern. Each of such patterns will have a slightly different σ_m and hence effect size. Thus we are practically concerned with a range of effect sizes, depending on the precise pattern of dose effects.

As an example, assume a standard DNT design, with a control group and three increasing dose groups. Assume further that we are measuring the body weight of young adult male rats. Mean weight of controls in this example is 360 g. Body weights are tightly controlled physiologically, with coefficients of variation (CV) seldom more than 10%. This implies a standard deviation of 36, and we will also assume homogeneity of variance and equal n in each of the four groups. Then $\sigma = 36$, but σ_m and hence f will vary according to the precise pattern of results.

Table 2 presents possible outcomes according to a range of effect sizes and any of four possible effect patterns. Assuming $\sigma = 36$, effect size involves either a "medium", "large" or "very large" effect with a 0.5, 1.0 or 1.5 σ difference between controls and high-dose groups (corresponding respectively to 18, 36 and 54 g of body weight). For all three effect sizes, the low and middle doses can independently differ or not differ from control values. Note that σ_m, Cohen's effect size, varies substantially according to the precise pattern of results. A step-wise dose-response decrease in body weight in equal increments from control to high-dose produces the smallest Cohen's effect size, while decreases restricted to the two highest doses produce a substantially larger σ_m. Since power is a function of effect

Table 2
Power and sample sizes (at β=0.80 or 0.95) for different patterns of effects and effect sizes, using a simple 1-way ANOVA test for treatment effects on rat body weight (assuming standard deviation of 36 for all four treatment levels, C = control, L = low, M = mid, and H = high)

Effect size	Mean weights				Cohen's σ_m	Effect	Power	Sample size	
	C	L	M	H				0.80	0.95
Medium	360	354	348	342	6.708	0.186	0.24	80	125
	360	360	351	342	7.462	0.207	0.30	65	101
	360	360	360	342	7.794	0.217	0.32	60	93
	360	360	345	342	8.318	0.231	0.36	53	82
Large	360	348	336	324	13.416	0.373	0.79	21	32
	360	360	342	324	14.925	0.415	0.87	16	24
	360	360	360	324	15.588	0.433	0.90	16	24
	360	360	330	324	16.636	0.462	0.94	14	22
Very large	360	342	324	306	20.125	0.559	0.99	10	15
	360	360	333	306	22.387	0.622	~1.0	9	13
	360	360	360	306	23.383	0.650	~1.0	8	12
	360	360	315	306	24.954	0.693	~1.0	7	10

"Medium", "Large" and "Very large" refer to effect size with 0.5, 1.0 and 1.5 SD difference between control and high dose, respectively.
σ_m: equivalent to the square root of the between-groups mean square error.

size, power also depends appreciably on the precise pattern of obtained results.

Table 2 shows that when there is a difference between control and high-dose groups of 0.5σ, effect sizes are just slightly less than Cohen's "medium" effect, while differences of a full σ give us a "large" effect size around Cohen's value of f>0.4. Thus a 5% drop in body weight between control and high dose is close to a "medium" effect, while a 10% drop is a "large" effect in this example. Note also that even if we accept a power of 0.80 as, if not optimal, at least minimally acceptable, then the common n=20/cell never achieves adequate power to detect "medium" effects, as defined by Cohen. In contrast, n=20 does achieve such power levels, but only for "large" or "very large" effect sizes. If we further assume that the usual high-dose level is set far above expected population exposure levels, thus guaranteeing very large effect sizes, we can be satisfied with the DNT study design from the narrow standpoint of power. A first conclusion, then, is that the DNT suggestion of n=20 animals per cell is adequate if not generous, while lower sample sizes provide acceptable protection against Type II errors only for very large differences.

Further to this discussion, Table 2 also provides estimates of sample sizes per cell required to attain power of 0.80 and 0.95, for the above effect patterns and sizes. Again it is obvious that detection of medium effects is not possible with anything like n=20 animals per cell, and that a power of 0.95 is approached but never quite attained even with large effects restricted to the highest dose groups. These data further illustrate that for practical reasons, it is improbable that we will achieve protection from Type II errors at less than β=0.2 except for very large effects.

To complicate matters, the experimental designs actually used in DNT studies are more complex than those shown in Table 2. This additional complexity is a consequence of the presence of sex in the experimental design. The DNT guidelines correctly specify that chemical effects be tested on both sexes, with up to 20 treated litters per dose group without specifying how sex should be included in this design.

Practically, three options have been utilized. One is to test one male and one female from each of, say, 20 litters per dose. Given the conventional four treatment levels, this means testing 40 animals per treatment level, or a total of 160 animals. A second alternative is to test just one animal per litter, with 10 of the 20 litters per treatment level contributing only one male each, and the other ten litters contributing one female each. In this design there are only 10 animals per cell, with 8 cells (2 sexes × 4 treatment levels), or 80 animals total to be tested. The third alternative has been to use only ten litters per treatment level, with each litter again contributing one male and one female. Here also a total of 80 animals are tested.

Statistically, the simplest of these three designs uses just one animal per litter, with half the litters contributing one male each, and the other half one female each. This is a classical 2-way ANOVA design, with both sex and treatment as fixed effects. There is a single error term, with (in our example of 20 litters/dose) 72 degrees of freedom, a main effect of sex (1 *df*) and treatment (3 *df*), and of course a sex-by-treatment interaction, also with 3 *df*. Again, a total of 80 animals are tested.

When each litter contributes two animals, one male and one female, the design is a more complex "split-plot" or mixed model. Sex must be treated as a correlated variable within litter (comparable to a repeated-measures design, where the subject rather than the litter is the unit of analysis), and tested by using the residual within-litters error term. The treatment effect, on the other hand, is a between-groups effect, and is tested using the between-litters error term. This has complex and sometimes surprising effects on statistical power. In this example, taken from the third alternative, again we test 80 animals, but we now have 10 and not 20 litters per treatment, with two animals (one of each sex) per litter. This model too has a total of 80 animals. However, there are now two error terms (one for testing treatment effects, the second for testing the correlated sex and sex by treatment effect). Each error term has 36 degrees of freedom.

Hence, while such mixed designs can be very powerful for detection of the correlated measure, in this case the effect of sex, they are generally less powerful for detecting treatment effects. Note, for example, that the degrees of freedom in the error term for the simpler 2-way sex-by-treatment ANOVA (72) is twice the degrees of freedom for either error term in the mixed model [39], although both designs contain an identical total of 80 animals. For this reason alone, the mixed model will provide lower power for the detection of treatment effects than does the 2-way model. Further discussion of these options is beyond the scope of this paper, but again it must be stressed that when litters contribute both sexes, the sex and the sex by treatment effect must be analyzed as a correlated variable.

Multiple tests have serious consequences for experiment-wise Type I protection; matters are more complex where Type II errors are involved. Increasing the number of statistically independent tests where there are true population differences between controls and some high-dose level has the paradoxical effect of both

increasing and decreasing power. Assume that there are three independent "real" effects in a DNT study, significant at $\alpha < 0.05$, and that the power of our DNT study to detect a real effect of this size is 0.80. Then it is obvious that our ability to detect all three effects is low, being $(0.80)^3$ or 0.512. Thus conducting multiple independent tests of true effects actually substantially reduces our ability to detect all such effects. In this narrow sense, then, multiple tests can reduce power. On the other hand, our ability to detect at least one of the three independent true effects is increased. This power is simply 1 minus the probability of detecting no effects, or $1-\beta^3$. In the case of a power of 0.8 to detect a single effect, our ability to detect at least one (or more) of three effects now goes up to $1-0.008$, i.e. 0.992.

This is a heartening result, since presumably detection of even one such effect would be sufficient to raise some sort of warning flag. However, it is by no means certain that we would ever have three truly independent treatment effects in a DNT study. For example, typically all 15 or so assessment procedures (e.g., body weights, morphometric measurements and a range of behavioral variables) are conducted on animals from the same 20 litters. It is likely that under these circumstances truly independent effects are impossible, even if the treatment produces different effects, via different mechanisms, on three different dependent variables. We can conclude that, unlike Type I errors, power (defined as our ability to detect at least one real effect) does not go down with increasing test number, and indeed may increase substantially, to the degree that such tests are truly independent.

One cannot discuss power without mentioning test variability. While high variability is not good, matters are never that simple. As we have seen, effect size is generally measured in z scores (recall that correlations too are simply the mean cross-product of two z scores). This removes all variability differences between dependent variables, by normalizing for such variability. Thus the dispute about whether a more variable measure is "worse" than a less variable measure does not pertain to power calculations when effect size is normalized. However, this normalized approach can be contrasted with the criterion-related approach, which expresses the effect of interest not in terms of standard deviations, but in terms of percentages. In such a case, variability matters. It is important to recall that all power calculations require estimation of two parameters, variance and difference from mean (criterion). Z-scores combine both measures. The criterion-related approach does not, and hence use of criteria (e.g., 20% change from baseline) is meaningless for power considerations without also tabulating a range of possible variances, an extremely clumsy approach which necessitates huge tables.

In any case, when the high variability is due to high measurement error, there is a problem. We must ever strive to minimize such error. On the other hand, in many cases high variability may not be due to measurement errors. Biologically, some variables must be kept within very narrow ranges, or survival is imperiled. Thus organ and body weights have relatively low variability, because a 50% drop in body weight is lethal. On the other hand, many variables do not require such precise control, and hence are physiologically free to vary widely. This may be true for some behavioral variables, for instance. The point is that in either case the practice of measuring effect size in standard deviation units may have biological validity. Presumably one can alter loosely controlled variables more substantially than tightly-controlled variables, and hence treatment effects can be expected to be proportional to standard deviations in both cases. Thus, inherent variability is of little consequence in the way power is often calculated.

As shown above, $n=20$ litters per treatment group is near a maximum practically feasible level, and does provide adequate protection against Type II errors where large effects are involved. However, in many DNTs (especially where time-consuming neuropathology examination is involved), all available litters are not always used, with sample size below 10 litters. Furthermore, there is often no clear explanation of how these litters were chosen. Based on the above power considerations, the current practice of sampling less than the full number of 20 litters is generally discouraged. Indeed, conducting analyses at such insufficient power levels may be worse than not conducting analyses at all, since it is all but sure to lull us into a sometimes unwarranted false sense of security.

3.4. Statistical analyses and assumptions

Statistical tests are typically designed to be used under a specified set of assumptions, and violations of these assumptions may have more or less severe consequences. One assumption that is paramount and common to all statistical analyses is that of randomness (or independence of observations). This assumption specifies that every sample is made of cases randomly chosen. For example, selecting the smallest or the heaviest pups for testing procedures would violate one of the most important and general assumptions of statistical practice, and very serious errors in inference could be made. It is important that analysis methodology be determined *a priori* based on historical data and knowledge rather than reliance on preliminary tests of assumptions that may prove to be overly sensitive.

3.4.1. Analysis of variance

Several studies from Box [6,7] on the analysis of variance (ANOVA) demonstrated that violation of the assumption of homogeneity of variance is not very serious when the number of cases in each group is similar. The author [5] stated:

> "To make a preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!" (p. 333).

More recently, Hays [23] also wrote:

> "… a test for homogeneity of variance before the analysis of variance has rather limited practical utility, and modern opinion holds it that the analysis of variance can and should be carried on without a preliminary test of variances, especially in situations where the number of cases in the various samples can be made equal." (p. 484).

One rule of thumb regarding heterogeneity of variance in the F test is that, so long as group sizes are reasonably similar, one can use ANOVA methods based on the assumption of homogeneity of variance with confidence so long as the largest standard deviation is less than twice the smallest standard deviation [43], p. 752.

As far as the normality of the distributions is concerned, Box [5] showed that the ANOVA is also robust against "general" departure from normality as far as the distributions have about the same shape, i.e. same skewness. If the skewness differs from one group to another, false declarations of an effect increase.

Parametric techniques are usually robust against departure from normality and equality of variances, except in severe cases. Norton (pp. 78–86) [34] looked at what happens to F in the case of non-normal distributions and when variances are not equal. She showed that, unless form and variance heterogeneity are extreme among treatment populations, the F test is not markedly affected (with sample sizes ranging from 3 to 10 used in the simulations), but, in general, to address a potential effect on the F ratio, allowance can be made to operate at a lower α level due to the increased false positive rate. Transformations can also be considered (Section 3.4.4.1). A thorough discussion of the consequences of assumption violations on Type I and Type II errors can be found in Glass et al. [20].

3.4.2. Repeated-measure analyses

Analysis methods designed to take into account repeated measurements on the same animal introduce unique data characteristics and additional analysis assumptions. In addition to random error attributable to between-animal variation, repeated measures include random error attributable to within-animal variation. The within-animal variation across time may influence the statistical analysis. For example, a univariate repeated-measures analysis of variance assumes sphericity (also known as circularity) of the variance–covariance structure and deviations thereof can result in increased false positives [63]. The "sphericity" condition is said to be fulfilled when the variances of the differences across repeated levels are equal. One means by which to address the lack of sphericity issue is through the Greenhouse–Geisser or Huynh–Feldt corrections which have been recommended by Tamura and Buelke-Sam [59]. The multivariate approach to repeated-measures analysis represents another alternative that is free from this circularity assumption [48]. A third approach is to employ mixed-model methodology. The appeal of this approach is that, while the aforementioned techniques address specific assumptions about the covariance structure, mixed-model methodology allows for the evaluation of multiple structures in order to construct a model that best fits the data. Recent computer software development has made mixed-model methodology more readily available (see Section 3.7 for more details about repeated-measures analysis).

3.4.3. Analysis of covariance

The analysis of covariance (ANCOVA) enables removal of the effects of an uncontrolled source of variation in one variable (called "covariate") from the analysis of another variable. For example, a test compound may affect nerve conduction velocity, which is known to be affected by temperature. If the test compound does not itself affect temperature, the ANCOVA will allow for the removal of the normal temperature variations (covariate) from the combined effects of temperature and test compound. Other examples may include using a baseline as a covariate to adjust for any pre-existing differences in one variable, or using litter size to adjust pup weight.

The analysis of covariance has additional important assumptions compared to the ANOVA [13], such as:

a. linearity of regressions: the standard covariance analysis assumes that the relationship between covariate x and criterion variable y is linear. Simply, an x–y scatter plot for each treatment group could be generated; or a test for linearity of regression could be performed [23], pp. 684–686;
b. homogeneity of slopes: the slopes of the regressions are parallel (i.e., same for all treatment groups). In other words, there are no statistically significant treatment-by-slope interactions;
c. covariate independence of treatment: covariate x is statistically independent of criterion variable y; in other words, treatment does not affect the covariate. If it does, part of the treatment effect may be removed by the regression adjustment or produce spurious meaningless results. An ANOVA of the covariate will help decide whether the covariate is affected by treatment.

Haseman et al. [22] provides an excellent example of the latter issue in which body weight is used as an adjustment in the analysis of organ weights. A comparison is made of the analysis of organ/body weight ratios vs. the analysis of organ weights using body weight as a covariate in an ANCOVA. While showing a preference for the ANCOVA approach, the authors caution against the potential difficulties when the test chemical affects both organ and body weight. When this is the case, the relative impact on organ weight of the test chemical and reduced body weight may be confounded. That is, it may be difficult to distinguish whether a chemical reduces organ weight simply by making the animal smaller or it has a direct effect on the organ itself. Shirley [54], on the other hand, argues that the use of relative organ weights has its own (often violated) assumptions and that this analysis is often misleading. The author continues by stating that, after simulations, the analysis of covariance was greatly superior to the analysis of relative organ weights. As noted by the authors, the investigator should be aware of this when interpreting organ weight changes using either an ANCOVA or an analysis of relative body weights.

Overall, the conclusions about the consequences of violating the ANOVA assumptions carry over to the ANCOVA [20]. But ANCOVA does not appear to be robust against violations of equal slope and normality with groups of unequal sample sizes [32]. Lord [38] also cautioned that errors of measurement in the covariate may either create the appearance of an effect or hide it. Nevertheless, ANCOVA can still often be a tool of choice, but considerable care is needed in applying ANCOVA procedures and in interpreting them. Because of the intricacies and the complexity of its assumptions, ANCOVA should not be used without expert advice.

3.4.4. Other considerations

3.4.4.1. Transformations. As appropriate, transformations can be used to stabilize variances and/or normalize distributions, and to linearize regressions. They are best used *a priori*, for

example, when the data are expected to follow a non-normal distribution: for example, Pryor et al. [49] used a square root transformation for motor activity data. Such a transformation has both a variance stabilizing and normalizing effect. Data that present a large proportion of values in the low or high ranges (e.g., 0–30% or 70–100%) usually benefit from an arcsine transformation (also known as angular transformation) that stretches both tails of the distribution. A very general transformation to achieve normality is the Box–Cox transformation, which allows selection of an optimal normality transformation based on any particular non-normal distribution [47]. Transformations, however, should not be chosen *a posteriori* on the basis of the statistical significance of the results. A far better approach is for laboratories to develop and routinely apply appropriate transformations for those DNT tests that reliably generate distributions that violate assumptions such as normality or homogeneity of variance.

It should be realized that, in the acoustic startle test, for example, some "raw" data are themselves transformations; e.g. the sound pressure level (SPL) of an auditory stimulus is often characterized as dB(SPL), where the reference sound pressure is 20 μPa. It can also be expressed as dB(A) when the "A weighting filter" is used to express the sound level. In any case, these relative dBs are really a measure of ratios on a logarithmic scale, i.e. a transformation. Decibels (without a reference) are dimensionless units, however, and cannot be used to express the sound level of an auditory stimulus.

3.4.4.2. Measurement scales. Treating data at a higher measurement level than warranted has also some potential consequences. The type of data represented by an analysis endpoint can be defined in several different ways. In a very broad sense, variables can be classified as continuous or discrete. A continuous variable can assume any value within a reasonable range defined by the limits of a measurement instrument. Body weights are classic examples of continuous variables. By contrast, discrete variables can only assume a limited number of values. Most observational endpoints are discrete in that an observer chooses from a short, predefined list of possible outcomes to categorize an observation (e.g., ranking of rat's reactivity).

There are several levels of measurement scales from the most limited to the most complex. Classification of data types can be refined by considering the amount of information or detail represented by the data [56,64].

a. Nominal (unordered categorical, descriptive): this scale distinguishes between categories. The values are unique identifiers, but do not reflect numerical relationships, e.g., presence or absence of eye opening, description of abnormal movements categorized into gait disturbance, tremors, and convulsions. If numbers are assigned to any of these categories, they are arbitrary.
b. Ordinal (ordered categorical, graded, rank-ordered): the classes are ordered along some continuum and eventually assigned a numerical rank or order. For example, five categories (numbered 1 through 5) from completely constricted to completely dilated pupils. An equal difference between assigned numbers does not represent an equal difference in the magnitude of the observation.
c. Interval: this scale ranks the relative order of the measure and contains equal units, but does not have an absolute zero, e.g., temperature, but 30 °C is not twice as warm as 15 °C.
d. Ratio: this scale is an interval scale with an absolute zero which may be used as a reference point. For example, motor activity counts or body weights range from zero to x, and $x/2$ is twice as many counts or grams as $x/4$.

Different types of statistical methods have been designed to analyze different types of data (Section 3.9). The same statistical test can, however, be eventually used with different types of data; however, many potential pitfalls and limitations must be recognized by the investigator. For example, analyzing nominal data with techniques designed to treat ratio data may either affect the power of the test, or generate uninterpretable results. Expert advice should be sought in the analyses of these data types.

3.5. Analyses of sex effects

One important aspect of the DNT guideline is the requirement that treatment effects be assessed in both sexes. With today's knowledge of sex differences in disease and toxicity, much clinical and animal research should be conducted in both sexes. Still, while mandating the inclusion of both sexes in toxicity testing, the DNT guideline does not stipulate how experiments should be designed and analyzed to derive maximum benefit from the inclusion of both sexes.

There are many advantages to properly including sex as a factor in the statistical analyses. This reduces the number of significance tests by half, while conferring substantial benefits in interpretation of the study findings. The main effects of treatment are measured across sex, not simply by sex. Not infrequently we find significant treatment effects in one sex, with trends in the opposite sex. Testing of the sex-by-treatment interaction will reveal whether there actually is a sex difference in treatment effects.

There are currently three approaches within conventional analysis of variance to the measurement of sex differences in response to potentially toxic compounds. The first, and by far the most problematic, of these approaches is to simply analyze each sex independently for dose effects. This is also the approach universally conducted in DNT studies. A second approach is to draw at least one male and one female from each exposed litter, while the third approach draws each sex from exactly half of all exposed litters, so that litter contributes subjects of only one sex (Section 3.3).

While the current DNT practice is to analyze the data from each sex separately, that approach is fraught with problems. First, in some studies it is unclear how sexes are drawn within litters. Thus this design may not always respect the litter as the fundamental unit of statistical analysis in all prenatal exposure designs. Second, analyzing the sexes separately fails to address the fundamental question, which is whether there are sex differences in treatment effects. The only way this important

question can be answered is by testing for the presence of a statistically significant sex-by-dose interaction. A third problem created by the practice of analyzing each sex independently is that this doubles the number of experiment-wise statistical tests for the main effect of treatment. Since as we have seen the alpha inflation caused by conducting hundreds of such significance tests is one of the greatest problems in DNT studies, this practice should be discouraged on the grounds of alpha inflation alone. A fourth problem is that the practice of analyzing exposure effects independently by sex does not provide any measure of the effects of sex itself on dependent variables. These sex effects are frequently sizeable, and hence provide a convenient internal measure of the reliability of the dependent variable.

In summary, in analyses of DNT data, sex must be included as a fixed effect variable in all analyses of treatment effects. The current practice of analyzing results separately by sex is simply not appropriate.

3.6. Litter effects

Treating multiple offspring from the same litter as independent subjects is a fundamental violation of assumptions that can severely inflate alpha levels [26,55,66]. The current DNT practice sometimes recognizes this principle in young preweaning animals, but not always in adults, on the false assumption that litter effects do not extend beyond infancy or weaning. This is, however, a mistaken assumption. Litter effects, which is to say correlations across littermates, exist and are large in young adult rats [22]. Indeed, if humans are any measure, "litter effects" (sib–sib or parent–child correlations) exist throughout life, and increase with advancing age.

As to the existence of litter effects in adult rats, Table 3 shows litter effects seen in a recent experiment (Sobrian, personal communication). The study design included both males and females, drawn from separate litters. Each litter contributed three same-sex siblings, and these animals were reared under different conditions until testing began at an age ≥ PND 110. Rearing and sex effects were removed by converting all scores to within-group z scores prior to analysis of litter effects. As the table shows, organ and body weights and a range of behavioral measures showed significant litter effects in these adult males and females. The obvious recommendation, then, is that litter must be tracked over time, and treated appropriately in statistical analyses.

Table 3
Litter correlations for adult rats (>110 days of age) using sample data (Sobrian, personal communication)

Dependent variable	F	p
Plus maze: entries into lighted arms	$F(23,48)=2.56$	0.003
Radial 8-arm maze: latency over 3 weeks of testing	$F(23,48)=1.93$	0.028
Morris water maze: latency over 3 weeks of testing	$F(23,48)=0.95$	0.535
Open field: mean square entries, 4 consecutive daily sessions	$F(23,48)=2.47$	0.004
Body weight	$F(23,48)=2.17$	0.012
Brain weight	$F(23,48)=3.70$	0.0001

Undeniably, then, litter effects are real, and often large, even in adult rats. Consequently, it is necessary to track litter throughout all stages of the DNT, and to adjust statistical data analyses accordingly. This is hardly an onerous requirement, but here, too, one caveat is in order. Some seem to have gone overboard in their respect for this principle, to the degree that there have been attempts to track the litter from the supplier of experimental animals. While there is no harm in such a practice, it is not essential, especially if potential dams are assigned to conditions under sound matching or randomization procedures.

Finally, we add a very brief word regarding recent reports on the use of the litter as the basic unit of analysis in toxicology studies. In an excellent paper, Elswick and colleagues [15] have analyzed the effects of using 1 or more ventral prostate weights per litter on experimental outcome and power. This paper correctly used litter as a random factor in all analyses, and not surprisingly concluded that drawing ventral prostate weights from more than one pup per litter was preferable to the use of a single pup per litter. Hence this paper does not in any way question the use of the litter as the fundamental unit in analysis; it only shows that litter means based on a larger n are more accurate, an inarguable conclusion. A second result, published as a recent abstract [16], is more problematic. This abstract appears to report that drawing data for spontaneous motor activity from 3 mice from each of three litters and treating this as an n of 9 has the same power as using animals from 9 litters, evidently one animal per litter. It would appear that this was not a true Monte Carlo study, because statistically this conclusion is inaccurate, and would not be obtained in a true Monte Carlo simulation when, as is generally the case (see Table 3), there are litter effects on spontaneous motor activity (Monte Carlo analysis is a statistical method used for simulating reality that takes into account randomness by testing a very large number of scenarios).

In summary, ignoring litter effects in the statistical analysis of DNT studies is simply not an acceptable practice. Standard ANOVA models make inclusion of litter as a correlated variable straight-forward, and failures to use such models risk unacceptable levels of alpha inflation.

3.7. Repeated measures

Common among many DNT designs are multiple measurements of the same endpoint on the same animal at different time points. For example, motor activity might be tested in the same group of animals on PND 13, 17, 21 and 60. Likewise, auditory startle might be tested in the same group of animals on PND 22 and 61. In such cases, the data present "repeated measures" on the same experimental unit (individual animal) and the statistical analysis should reflect that aspect of the study design.

The motor activity and auditory startle examples described above are ones in which the repeated measures represent sessions conducted on different days of the study (across-sessions). In both these endpoints, repeated measures are often artificially created for the statistical evaluation of within-session data. For example, data from a one-hour motor activity session is often broken down into 10 or 15-minute intervals for presentation and statistical

analysis. Likewise, 50 trials of auditory startle are often broken down into blocks of 10 trials each. The objective of the within-session evaluation is to assess adaptation that normally occurs during the test session. The objective of the across-session evaluation is to assess changes that occur over days across sessions. In both cases, a repeated-measures analysis provides the means by which to evaluate the changes in motor activity over two time frames (session and days).

Technically, repeated measures imply multiple measurements on the same experimental unit and therefore may occur in many different forms. For example, repeated measures may occur as dose level (many pharmacological studies are designed such that the same animal receives different dose levels) or within-session activity (several consecutive bins of motor activity recorded within a one-hour session). For simplicity's sake, the following discussion assumes a typical DNT study design in which parallel groups of animals are administered different dose levels (treatment groups) and measured at multiple time points (time).

From a statistical standpoint, the repeated-measures design introduces an additional dimension to the analysis. The primary evaluation of interest is that of the different treatment groups. Treatment groups are considered to be a between-subject (between-animal or, more commonly, between-litter) effect because the treatments are administered to the individual animal (or dam), i.e., each animal (or litter) is contained in one, and only one, treatment group. Time is considered a within-subject effect because the same animals are measured at the different time points. Animals (or dams) are randomly assigned to the treatment groups and therefore it can be assumed that measurements made on the different animals (between-animal or between-litter) are independent. However, the repeated measurements on the same animal (time) are not randomly assigned and therefore are not independent. That is, the measurements taken on the same animal are correlated and this must be taken into account in the analysis. As a result, there are two sources of random variation: between-animal (or litter) and within-animal."

There are numerous statistical techniques for analysis of repeated-measures design. One approach utilizes a univariate analysis of variance. If correlations between responses from the same animal are the same regardless of time proximity, then the univariate approach provides a valid method for repeated-measures analysis. However, this assumption may not always be realistic. For example, in an across-session motor activity analysis it may not be reasonable to assume that measurements taken on days 13 and 17 have the same correlation as those taken on days 13 and 60. Likewise, in a within-session motor activity analysis one might expect measurements taken in the 0–10 minute interval to be more highly correlated with those in the 11–20 minute interval than with those in the 51–60 minute interval when adaptation may have occurred.

The Huynh–Feldt condition (sphericity) refers to the differences between each pair of responses having equal variances and is a necessary condition for the univariate approach to repeated-measures analysis. If the Huynh–Feldt condition is not met, the test statistics for the within-subject effects (time and treatment-by-time interaction) will be overestimated, thus increasing the Type I error rates. There are analysis adjustments available (e.g., Greenhouse–Geisser and Huynh–Feldt) to account for the within-subjects correlations but some consider them inadequate in that they simply make general adjustments to degrees of freedom. Other more advanced alternatives may be preferable (see below).

A multivariate approach to repeated measures provides an alternative approach to evaluating within-subject factors when the Huynh–Feldt condition is not met. Although this approach may be preferred when the Huynh–Feldt condition is not met, it also makes assumptions that may be too general. For example, the multivariate approach assumes that the correlation between all pairs of responses for an animal is unique. As a result, the power of the multivariate approach is reduced if this assumption is not true. In addition, if for some reason a response for an animal is missing for a single time point, the multivariate methodology excludes all data from that animal in the analysis if an implied value cannot be supplied for the missing information.

More recent software and statistical methodology development provides the mixed-model approach to repeated-measures analysis. The name "mixed-model" refers to the mix of fixed effects (treatment group) with random effects (time) in the statistical model. The advantage of this approach is that it provides flexibility in modeling the correlated data presented by within-animal measurements (the covariance structure of the statistical model). The method involves a sequential approach to the repeated-measures analysis:

1. define the statistical model,
2. evaluate various covariance structures to determine the "best fit",
3. make statistical inferences based on the model determined in step 2.

What makes the mixed-model approach unique is its flexibility to evaluate various covariance structures to determine a model that best describes the within-subject correlations. Covariance structures range from very simple (compound symmetric: correlations between responses from the same animal are the same regardless of time proximity) to very complex (unstructured: within-subject correlations are unique for every pair of time points). The multivariate approach assumes an unstructured covariance structure and can therefore be less powerful when some correlation between observations does indeed exist. The univariate approach assumes the Huynh–Feldt condition (a general form of compound symmetry) and can lead to inflated Type I error rates when in fact the correlations are more complex. In truth, the correlation patterns may fall somewhere between these two extremes. For example, the first-order autoregressive structure assumes that the time points are equally spaced and the correlation between observations is a function of their distance in time. This seems intuitive in that one might expect a higher correlation between measurements that are closer together in time than those that are further apart. There are many other structures making different assumptions (e.g., unequally spaced time points).

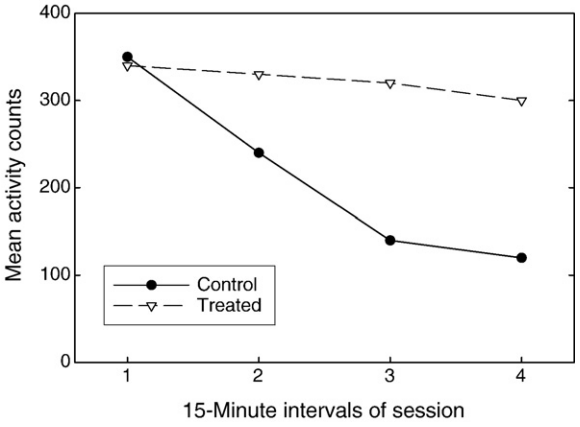

Fig. 1. Interpreting interactions: statistically significant main effect of treatment and significant time-by-treatment interaction. This figure (illustrating a plausible but fictitious outcome for a control and a treated group) shows a failure of motor activity to adapt in the treated group after the first 15 min, giving the appearance of hyperactivity and a significant main effect of treatment when activity is collapsed over trials.

In practice, various structures can be evaluated and, over time, the structure that best describes the within-subject correlations may be narrowed down to one or two for specific endpoints. For example, over the course of several studies it may be determined that the compound symmetric and first-order autoregressive structures consistently provide a better fit for the equally spaced time points over the narrow time frame of the within-session motor activity data. Likewise, one or two other structures might consistently provide better fits for the unequally spaced time points of the wider time frame provided by the between-session motor activity data. The historical evidence would suffice for narrowing the covariance evaluation to two specific structures for each endpoint. A thorough discussion of the advantages and disadvantages of the univariate, multivariate, and mixed-model repeated-measures analyses are provided by Littell and colleagues [35–37].

The advantage of repeated-measures analysis, regardless of the approach utilized, is that it allows for the evaluation of treatment group effects across time. For this evaluation it must first be determined if the treatment group effects remain constant across time. That is, are the observed effects of the different treatment groups dependent on the time point at which they were observed or are they basically the same at all time points? This question is addressed by the interaction of treatment group and time (treatment-by-time). A nonsignificant treatment-by-time interaction indicates that the treatment group effect remains constant across the time points and, therefore, it is reasonable to draw conclusions across the pooled time points by evaluating the treatment main effect. A significant treatment-by-time interaction indicates that the treatment group effect differs depending on the time point. In the presence of a treatment-by-time interaction, it is necessary to evaluate the nature of the interaction to determine if conclusions can be drawn across the pooled time points or if the individual time points should be considered individually.

Fig. 1 illustrates a significant treatment-by-time interaction for a within-session motor activity test. (For illustrative purposes only a control and one treated group are shown.) There is very little treatment effect in the first 15 min following dose. However, while the control group shows adaptation ("habituation") by 60 min after dose, the high-dose group remains very active. That is, there is no treatment effect at the first time point but a very large treatment effect thereafter. Because this treatment effect is monotonic (nondecreasing or nonincreasing), one could evaluate the treatment effect across the pooled time points and come to the conclusion that there was a significant treatment effect for the one-hour session.

However, simply ignoring the treatment-by-time interaction can adversely affect the interpretation of results. Consider the motor activity session depicted in Fig. 2. While the expected adaptation has occurred in the control group, the treated group has maintained a constant level of activity that is below that of the control at the 15- and 30-minute intervals, but above that of the control at the 45- and 60-minute intervals. Again, this illustrates a significant treatment-by-time interaction in that the treatment effect depends on the time point. However, to ignore the significant interaction and evaluate the treatment effect across the pooled time points would lead one to the false conclusion that there is no treatment effect for the one-hour session. That is, the true effect of treatment on adaptation is "masked" because, on average, the treated group's activity was about equal to that of the control group for the entire session as a whole.

There are different approaches on how to proceed in light of a significant treatment-by-time interaction. For some, the presence of a significant interaction is sufficient statistical evidence that there is a treatment effect on motor activity during the 60-minute session. That is, in the examples above the response curve of the treated group differs from that of the control group. From this point, one would examine graphical displays and individual data in order to determine the scientific relevance of the treatment effect.

For others, a significant treatment-by-time interaction would trigger statistical evaluations at each individual time point. In

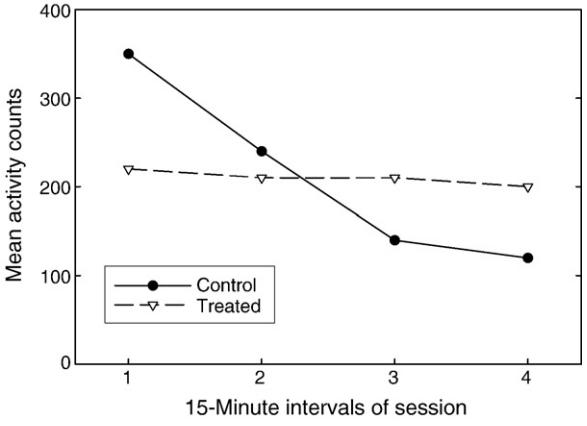

Fig. 2. Interpreting interactions: statistically significant time-by-treatment interaction without significant main effect of treatment. In this example, there is still a significant interaction between time and treatment, but it takes the form of lower initial activity, again as in Fig. 1 without adaptation. Clearly in this instance the significant interaction is not accompanied by a significant overall main effect of treatment.

the first example above this would lead to the conclusion that the treatment effect was statistically significant at the 30-minute time interval and continued through the rest of the session. In the second example, this would lead to the conclusion that, relative to control, the treatment effect was a significant decrease in activity at the 15-minute time interval and a significant increase at the 45- and 60-minute intervals. From this point, one would determine the scientific relevance of the statistical findings.

Both approaches have merit and are valid from a statistical standpoint. The important concept that they share is the recognition that there is a statistically significant treatment effect that is dependent on the time point. The basic difference in the approaches, as with any formal statistical analysis, centers on the point at which one abandons statistical inference in favor of scientific interpretation. As discussed in Section 4, statistical analysis must be viewed as a supplemental tool rather than an end in itself. As such, the scientist must decide how best to utilize the repeated-measures analysis tool to interpret the data.

3.8. Multiple pairwise comparison procedures

This section emphasizes control of inflation of Type I errors through the use of complex multifactorial experimental designs. Clearly in such complex multifactorial designs, the overall experimental objective determines where to halt the analysis, and sometimes simply obtaining a main effect of treatment will suffice. Yet undeniably the use of fewer, more complex statistical analyses can create substantial analytical problems. Foremost among these is the likelihood that, following overall multifactorial analyses, experimenters may well want to conduct many multiple comparisons between treatment means in a manner which adequately protects alpha levels and power. By way of example, a single procedure, such as brain morphometrics, can produce separate measures of treatment effects for each brain slice at every age. Thus, if thickness of six brain layers is assessed at two ages (PND 11 and 60) in a DNT with one control, three treatment groups and two sexes, there is a total of 72 possible pairwise comparisons between control and treatment means. Clearly this produces substantial problems for control of Type I and Type II errors. These problems are only accentuated by the dizzying multiplicity of multiple comparison techniques in the literature, leading to lack of adequate guidance on these problems by most statistics texts, and equally a lack of agreement among experts in this area.

It is also true that analysis of treatment effects in DNT studies may want to concentrate not on multiple pairwise comparisons, but rather on trend analyses across dose levels. Tukey et al. [61] provides an approach for evaluating response trends in an increasing dose design. One concern that investigators have with a trend analysis is that it may overlook meaningful responses which may not follow a dose-response pattern (e.g. *U*-shaped or inverted *U*-shaped response curves). However, trend analyses are easily adapted to evaluate such responses. For example, the trend analysis can be designed to evaluate both linear and quadratic dose responses. Another approach is to combine a linear trend analysis with pairwise comparisons to evaluate non-linear responses.

Planned trend analyses are certainly a viable alternative to multiple pairwise comparisons. The primary advantage, relative to alternative methods that test for homogeneity of groups (e.g., ANOVA or pairwise group comparisons), is that trend analyses are more powerful when the response truly does follow a dose-response trend. However, for the sake of brevity, trend tests will not be discussed herein in any greater detail. To properly tailor trend analyses to address study design and objectives, the investigator should employ the assistance of a knowledgeable statistician.

3.8.1. Terminology

Several distinctions need to be made at the outset of this discussion. Perhaps first is the distinction between planned and *post-hoc* tests. Planned multiple comparison procedures (MCPs) are a part of the experimental design, and must be in place before the first data are analyzed. Conversely, *post-hoc* tests are MCPs conducted after the fact, which is to say following at least a preliminary analysis of the data. This is an important distinction for scientific research. Planned comparisons can deal adequately with problems of power and alpha inflation. *Post-hoc* measures are often applied to data which appear "trendy", and hence are some subset of a large and uncontrolled family of multiple comparisons. Whatever the choice of multiple comparison procedures, we emphasize that in DNT studies, all such comparisons must be planned at the outset of the study, i.e., during the design stage, prior to any experimental interventions. In such cases planned pairwise comparisons are not "*post-hoc*", although we often incorrectly so refer to multiple pairwise planned comparisons.

A second distinction is that of pairwise comparisons between treatment levels and other comparisons. Perhaps the most important planned comparison will be unweighted pairwise comparisons between means of the treatment levels and the control group mean. Here we will refer to these pairwise comparisons as multiple pairwise comparison procedures, or MPCPs. This section will consider MPCPs, and especially those between control and all dose levels. This is in no way to discourage trends analyses, for example, but the topic is simply too vast to consider every possible planned comparison.

Still another important distinction is between protected and unprotected MPCPs. A protected MPCP is conducted only if the overall *F* test for treatment effects, or perhaps an interaction of some other factor(s) with treatment, is found to be statistically significant. Conversely, an unprotected MPCP is conducted regardless of overall *F* tests for significant treatment effects.

3.8.2. Choice of MPCPs

As always with inferential statistics, the choice of MPCP involves considerations of both alpha protection and power. For DNT purposes, two attractive alternatives appear to be either Dunnett's or Fisher's Least Significant Difference (LSD) [9,28,52,53], although as usual not all authorities are in agreement on this as on every other aspect of this topic [50,60]. Both tests have high power combined with reasonable Type I error rates, both are readily available in standard statistics packages, and both are extremely simple. Two other choices, neither quite as attractive, are Newman–Keuls and Duncan's. Newman–Keuls

tends to be more conservative than the Fisher's LSD, while Duncan's has exceptionally high Type I error rates. Other choices, including Bonferroni's, Tukey's HSD and Scheffé's, trade excellent alpha protection for substantial reductions in power, and generally cannot be recommended for DNT studies.

It is imperative that the study objective be the driving force in the selection of an MPCP. What makes Dunnett's test unique among the array of available MPCP is that it is designed to control alpha specifically for the subset of pairwise comparisons that consist of each individual treated group vs. a single control group. That is, in a study consisting of a control group and three treated groups, Dunnett's test controls alpha for the three pairwise comparisons of each treated group vs. control. Since this MPCP protects against only a specific subset of all possible pairwise comparisons, it is quite powerful and widely applicable to DNT studies, in which the comparison of treated groups with the control is a common objective.

By contrast, Fisher's protected LSD is designed to adjust alpha for all possible pairwise comparisons. In the example of a study with a control group and three treated groups, Fisher's protected LSD controls alpha for the six pairwise comparisons of the four groups. Thus, Fisher's LSD would be the preferred MPCP if the study's objective is to evaluate all possible group comparisons but would be somewhat conservative (less powerful), relative to Dunnett's test, if only comparisons with the control are of interest since it adjusts for additional comparisons that are not of interest.

Another distinguishing characteristic of the two tests is that Fisher's LSD is a protected test while Dunnett's test is designed as an unprotected test. That is, Fisher's LSD test is used as a *post-hoc* test only when the analysis of variance *F* test for treatment effect is statistically significant. By contrast, Dunnett's test is designed as a stand-alone test to be conducted without regard to the outcome of the analysis of variance *F* test for treatment effect [12]. It is not uncommon for DNT testing laboratories to misuse Dunnett's test as a protected test, i.e., dependent on the outcome of the analyses of variance. However, because the critical values and alpha adjustments are calculated based on Dunnett's being an unprotected stand-alone test, to use it as a protected *post-hoc* test reduces its power and may yield conservative results.

Table 4 presents a comparison of the results from Dunnett's and Fisher's protected LSD for five simulated experimental outcomes. For simulation purposes, all examples utilized a sample size of $n=20$/group and maintained a within-group standard deviation of 36. All tests in this simulation were conducted at the 0.05 significance level. Because it is a protected MPCP, Fisher's LSD was only conducted in the examples producing a significant treatment effect in the ANOVA test, whereas Dunnett's test was conducted for all examples since it is an unprotected test.

Table 4
Example results of the *F* test, a protected Fisher's LSD and the unprotected Dunnetts for a range of treatment outcomes

Example	Group means				ANOVA *p*-value	Multiple comparison	
	C	L	M	H		Fisher's LSD	Dunnett
1	360	370	380	386	0.104	NSD	NSD
2	360	370	380	389	0.062	NSD	H>C
3	360	389	380	370	0.062	NSD	L>C
4	360	370	380	391	0.041	H>C	H>C
5	370	380	360	394	0.021	H>M	NSD

Simulation uses $n=20$/group with a within-group SD=36, with four treatment groups (C = control, L = low, M = mid, H = high dose).
NSD=no significant difference.

In example 1, where the group means ranged from 360 to 386, there were no significant effects with either the ANOVA or Dunnett's test. Examples 2 and 3 show a slightly higher mean in either the high-dose or low-dose group, and Dunnett's test was conducted and those dose groups (with the mean of 389) were significantly greater than the control. This illustrates the increase in power that comes with Dunnett's test when the only group comparisons of interest are the pairwise comparisons of treated groups with the control. In both cases, the ANOVA, which tests the simultaneous equality of all four treatment group means, was not significant. Furthermore, since the ANOVA *F* test does not assign any order to the group means, the results were identical in both examples ($p=0.062$). In example 4, the high-dose group was even higher, and both tests used (Fisher's LSD and Dunnett's) identified the high-dose mean as being higher than that of control. Finally, in example 5, the mid-dose group mean is lower than, and the high dose is higher than that of control. The ANOVA followed by Fisher's LSD is not affected by this ordering and identifies the two extreme means as being different. However, when compared with the control mean, none of the treatment group means are different when tested with Dunnett's.

This last characteristic of Dunnett's test is of particular note in study designs that utilize dual control groups (e.g., both a vehicle and a pair-fed control group). Often times, laboratories will analyze such study designs by using Dunnett's test for comparison of treated groups with one control and then repeat the process for comparison with the second control. However, this practice should be avoided since, to do so employs Dunnett's test for twice the number of comparisons for which it was designed and thus increases the chances of a false positive.

In summary, if the only comparisons of interest are those of the individual treated groups with a single control, then Dunnett's MPCP is an efficient and powerful test that addresses the specific objective. If any other group comparisons are of interest, then Fisher's protected LSD is a powerful test that adjusts for all possible comparisons.

If a dose-response evaluation is the primary objective, then the reader is referred to tests designed for that task such as William's test [68] or linear contrasts. These tests gain power by restricting the alternative hypothesis (monotonic increasing or decreasing) as opposed to not equal, which is the alternative hypothesis in ANOVA and the MPCPs discussed here. These trend tests are not considered multiple pairwise comparison procedures in that, much like the ANOVA *F* test, they provide a simultaneous global conclusion about all groups rather than for specific pairs of groups. As discussed earlier, such trend tests are beyond the scope of this paper but the reader is referred to Tukey et al. [61].

Here it is also important to emphasize that these recommendations are not iron-clad laws. So long as adequate alpha protection is provided, investigators are free to choose among other alternatives. For instance, many DNT studies use Tukey's MPCP. It is certainly not wrong to do so, and indeed this procedure provides excellent protection against Type I errors. Yet equally clearly, Tukey's is less than optimal in terms of the overall protection provided against all forms of experimental error (Types I and II combined — Carmer and Swanson [9]) due to rather low power. So, as always, choice of an MPCP involves balancing relative costs and benefits, a balance which will vary according to experimental objectives.

3.8.3. MPCP testing when interactions with treatment are significant

Another issue is the use of MPCPs with several common multifactorial designs. In order to understand what is to follow, it is first necessary to remind the reader of the concept of *simple main effects*. This is an essential concept in interpreting most multifactorial designs, not least those using multiple correlated or repeated measures. If the multifactorial analysis reveals a significant interaction between treatment and another factor, it is necessary to unravel that interaction by testing treatment effects for every level of the interacting factor. For example, if a 2-way treatment-by-sex design produces a significant treatment-by-sex interaction, we will want to test treatment effects for every level of sex, which is to say separately for males and females. Here the test of the treatment effect for males alone is a test of a simple main effect of treatment. To understand the application of MPCPs to simple main effects of treatment, we begin by discussing the ubiquitous 2-way sex-by-treatment design. If none of the three effects (the main effects of treatment and sex and the treatment-by-sex interaction) are significant, analysis clearly stops here. Similarly, if the treatment effect is the sole significant effect, pairwise comparisons of treatment group means averaged across sex can be conducted using the recommended unprotected Dunnett's or protected Fisher's LSD.

Matters become more complex when both treatment and sex main effects are significant, or when there is a significant interaction between sex and treatment, independent of whether the two main effects are significant. To understand why this is so, consider analysis of body weight at some age. If both sex and treatment but not their interaction are significant, we are faced with a dilemma. The main effect of treatment in this design is simply the average of the weights of the two sexes for each of k treatment groups. However, this average is chimerical, in the sense that there is no animal whose weight is the average of male weights and female weights. Thus the experimenter will necessarily want to present treatment effects by sex, so that the reader can see how treatment actually affected real animals, not statistical chimeras. Presenting simple main effects by sex is also required for interpretation of a sex-by-treatment interaction. Thus there are two experimental outcomes necessitating multiple pairwise comparisons of the simple effect of treatment — a significant interaction, or significant main effects of both sex and treatment. In either case, we conduct MPCPs of two groups, males and females. How should we control alpha under these circumstances? Two possible approaches are discussed here. One alternative, virtually always utilized in current DNT designs, is to simply ignore the problem. This strategy increases power, albeit at the expense of increased Type I errors (doubled number of pairwise comparisons). This alternative will continue to be popular, especially when sex is not nested within litter. In such designs, simple main effects of sex involve sample sizes half as large as those for the main effect of treatment, and reductions of alpha to compensate for multiple comparisons becomes excessively conservative. A second alternative is to control family-wise alpha for the doubling of pairwise comparisons in some fashion. Several approaches less conservative than Bonferroni's are discussed in the following section. However, all still reduce alpha and hence power.

3.8.4. Correlated measures

The MPCP problem is aggravated when correlated measures are involved, as they usually are in DNT studies. By "correlated" measures we mean within-subject measures, multiple measures taken on each subject. These will usually take one or both of two forms. For instance, it is common to measure a number of dependent variables in a single procedure. Thus we might collect body weight, or thickness of six brain layers, or a range of behavioral variables including activity, rearing and stereotypy in a motor activity setup. The traditional repeated measures are another example of correlated dependent variables, since the same dependent variable is measured repeatedly. For example, motor activity may be measured every five minutes for an hour, and at three different ages. We suggest including as many such correlated measures as possible in single multifactorial analyses, even though this clearly greatly complicates the problem of alpha protection in such complex experimental designs. These problems are only further aggravated by the typical, and necessary, inclusion of both sexes in the analysis.

The best way to deal with the problem of multiple correlated measures is to begin by simplifying designs wherever possible. This approach is rare, but holds great promise for making MPCPs in multifactorial designs both practicable and interpretable. It seems to be rare for researchers to look at the correlation structure of their correlated dependent variables. For instance, automated startle tests or automated open field tests can today record many supposedly distinct variables. Yet assessment of correlations between such behavioral variables reveals that often these different variables are so highly correlated as to be virtually identical. Table 5 contains an example (Dr. S. Sobrian, personal communication). In this case, the seven machine-generated variables listed fall into two groups, horizontal activity and rearing (vertical activity). Several of the variables show significant correlations both within and between the two primary variables. Hence only two variables underlie this group of seven, and only two simple main effects of treatment need be assessed for these two variables, not the full seven.

Just as multiple dependent variables may be simplified by reducing to one variable for each correlated set, we can sometimes simplify repeated measures by the simple expedient of reducing the number of time intervals assessed [65]. For example, it may be reasonable to simplify repeated measures of

Table 5
Correlations between dependent variables for spontaneous motor activity ($n=33$ subjects)

Movement type	Variable	Horizontal movement				Vertical movement		
		Hactv	Totdist	Movtime	Restime	Vactv	Vmovno	Vmovtime
Horizontal	Hactv	–	0.923 *	0.922 *	−0.922 *	0.448	0.399	0.428
	Totdist		–	0.856 *	−0.856 *	0.296	0.180	0.308
	Movtime			–	−1.00	0.35	0.328	0.359
	Restime				–	−0.351	−0.328	−0.359
Vertical	Vactv					–	0.905 *	0.982 *
	Vmovno						–	0.865 *
	Vmovtime							–

Hactv = horizontal activity.
Totdist = total distance.
Movtime = movement time.
Restime = resting time.
Vactv = vertical activity.
Vmovno = vertical number of movements.
Vmovtime = vertical movement time.
* Statistically significant correlations.

adaptation by reducing the number of intervals ("bins") used in the data analyses. The DNT guidelines do not allow such reductions to proceed beyond a certain point, usually five blocks of ten trials for startle, or not less than five temporal periods in motor activity. The optimal strategy may be to divide temporal intervals or trials into a number of bins which, consonant with DNT guidelines, adequately describe the change over time. Such decisions should be made based on historical data, and need to be made during the test validation stage, and not separately for each study.

In summary, wherever possible the number of correlated within-subject variables should be reduced. This simple expedient will often do more to protect against alpha inflation without substantial loss of power than will any amount of statistical manipulation of MPCPs. Such simplification is rare in the current DNT practice.

3.8.5. MPCPs for multiple dependent variables

Once all practical data simplifications have been conducted, the experimenter (and the reader/assessor) will need to resolve the problem of alpha control for multiple MPCPs. For example, an analysis of morphometric brain measurement data may have to deal with as many as 6 to 8 endpoints. This actually presents two problems. First, since each subject contributes data for all measurements, these data typically do not meet the sphericity requirement for the conventional analysis of variance. Thus, as with repeated measures, the experimenter is well advised to use multivariate analysis techniques which do not require the sphericity assumption. A good choice is profile analysis [58]. Profile analysis is a multivariate technique which does not require sphericity, and which looks at the effects of treatment (in the above example) on the profile of all morphometric brain data. Typically, use of profile analysis is preceded by conversion of all raw data to z scores to accommodate for measurements of different magnitude. Profile analysis is then conducted, and provides tests of differences between weights, tests of the main effect of treatment, and a test of the treatment-by-organ weight interaction. When a significant interaction between organs and treatment is obtained, it will be necessary to conduct MPCPs on simple main effects of treatment for each organ.

An even simpler approach to multivariate analyses of multiple correlated endpoints is provided by Heyse [24]. This technique adjusts alpha levels based on the degree of correlation between variables and the most statistically significant p value obtained from tests of each of the k correlated variables. The p value is adjusted as follows:

$$p(\text{adjusted}) = 1 - (1 - p_0)^r$$

where p_0 is the smallest of k obtained p values, and r is an adjustment for correlation between variables (set at the square root of k if the actual value is unknown). If the obtained p (adjusted) is greater than alpha, then the interaction is not significant.

For either of the above approaches, a significant interaction will typically be followed by k individual tests for simple main effects of treatment. Given the relatively large number of such measures, any approach which adequately protects alpha will also substantially reduce the power of each MPCP. A Bonferroni-style adjustment may be utilized in such cases, and several such adjustments are presented below. The above approach extends to all correlated within-subject measures, including repeated measures. In all cases, a three-step procedure may be undertaken.

1. Wherever possible, simplify the data by pooling highly correlated variables and reducing the number of temporal intervals assessed (again, such simplifications need to be conducted prior to actual data analyses, not in a *post-hoc* fashion).
2. Conduct a multivariate repeated measures, a profile analysis of the results, or perhaps a Heyse adjustment.
3. When there is a significant interaction between treatment and correlated within-subject variables, then test simple main effects for treatment at each level of the correlated measure. Adjust family-wise alpha protection using one or another of Bonferroni-type adjustments to alpha, including several such

adjustments discussed below. Then use Dunnett's or Fisher's LSD for pairwise comparisons at the adjusted alpha level produced by one of these Bonferroni-style methods, for each simple main effect of the correlated within-subjects variable.

Several *p*-value or alpha adjustment procedures have been proposed. One extreme correction is the Bonferroni correction; it assumes independence of data and is conservative (i.e., low power). This correction simply divides alpha (typically 0.05) by *n*, the number of measures. This is excessively conservative, not least because all such adjustments are protected, that is, occur only if there is a significant interaction term. There have been a number of published Bonferroni-type adjustments to alpha which are at least slightly less conservative [25,27]. Another variation has been proposed by Tukey et al. [61] with the understanding that if making a Bonferroni correction to address the question of multiplicity of statistical tests is too unlikely to find significance (especially when the number of tests increases), then making no correction at all is equally unacceptable. Therefore, Tukey has proposed to divide α, not by the number of tests within a class of variables (as in the Bonferroni correction), but by its square root (in the absence of a known correlation structure) [40].

More recently, a different approach has been proposed, the "false discovery rate" (FDR) method [2,3]. Controlling the FDR was applied by Ellis et al. [14] to the analysis of neurochemical maps. Briefly, it consists of sorting the *p*-values and calculating an adjusted *p*-value as a function of its rank in the series. Practically, the unadjusted *p*-values are sorted in ascending order. An adjusted *p*-value is calculated by multiplying the unadjusted *p*-value by the ratio of its corresponding index (i) over the total number of *p*-values (n). This adjusted value is appropriate for independent tests. The following correction is necessary for dependent tests. It consists in multiplying the adjusted *p*-value by the sum of $1/i$ for $i=1$ and n.

Whatever method is used, it is important to note that any adjustment to an α or a p value will result in a modification of the power (for more details see Section 3.3).

3.9. Analyses of different data types

The amount of information provided by the data increases from the lowest level (nominal) to the highest level (ratio) (Section 3.4.4.2). As the level of measurement increases, so does the sophistication and availability of statistical methodology. Most of the analyses described in this paper are based on parametric methods (evaluation of parameters from known distributions such as the normal distribution) and are applicable to continuous data or data with the interval and ratio level of measurement. Parametric methods, such as analysis of variance-based methodology, are fully developed to address some of the major objectives of this paper (litter and sex effects, repeated measurements) and are readily available in statistical software packages.

Nonparametric methods, which do not rely on parameters of known distributions, are generally employed for the analysis of discrete or categorical variables, such as those described by the nominal or ordinal level of measurements. To make full use of available nonparametric methodology, one must first distinguish between nominal and ordinal variables. Statistical methods designed for ordinal variables address directional shifts in the response outcome. For example, the administration of treatment may result in a general shift in the ease-of-removal-from-cage response from easy to difficult. For nominal variables, the statistical methodology should simply address whether or not there is a general association between the treatment group classification and the pattern of response outcomes.

As the level of information provided by the data decreases, the sophistication and availability of statistical methodology becomes more limited. While the effects of litter, sex, and repeated measures are easily incorporated into statistical models utilized in parametric analyses, they are less easily addressed in nonparametric analyses used in the analysis of discrete variables. In practice, these effects are often held fixed while analyzing the effect of treatment groups. For example, the effect of treatment groups might be evaluated separately for each sex and time point. While less than desirable, this is often seen as a necessary sacrifice for the analysis of discrete variables. While this may be somewhat true for nominal variables, there are alternatives available when the discrete variable is ordinal in nature.

Methodology designed for ranked data analysis, such as Kruskal–Wallis and Wilcoxon, are often employed for the analysis of ordinal variables. These methods do not readily address repeated measures and, as the number of actual observed outcomes decreases resulting in a majority of response "ties" (same response outcome), the appropriateness of these ranked data analyses decreases. Furthermore, while it is true that the Friedman test provides a nonparametric approach to repeated measures, it too lacks generality in that it is only appropriate for a single repeated measure.

Another approach to discrete variable analysis employs the Mantel–Haenszel strategy [30,39,41]. The approach is often referred to as a "strategy" rather than a test because it provides flexibility to address general association (nominal variables) and directional shifts (ordinal variables) as well as directional shifts in response associated with directional shifts in the treatment variable (dose-response evaluation).

Methods for addressing repeated measures in discrete variables have been developed by Grizzle et al. [21] and Koch et al. [29], and are referred to as the GSK method (for Grizzle, Starmer and Koch). This methodology has been integrated into the SAS analysis procedure PROC CATMOD, so called because it utilizes categorical modeling methodology. The application of CATMOD to functional observational battery (FOB) data has been described by Creason [11]. This approach becomes less practical in the presence of sparse or "zero" cells (response levels for which very few or no animals in a treatment group were categorized at a given time point).

More recently, advances have been made in analytical models for ordinal data types. The generalized estimating equation (GEE) approach [33] provides another alternative to modeling categorical data. This methodology has been integrated into the SAS analysis procedure PROC GENMOD. While promising in its ability to address multiple factors such as treatment, litter, and sex, as well as repeated measures of ordinal data, the GEE

methodology can be quite complex and should only be employed with the assistance of a knowledgeable statistician.

Fleiss [17] and Stokes et al. [57] provide excellent resources for statistical analysis of discrete or categorical data analysis. As noted in Section 3.4.4.2, failure to distinguish the different data types and utilize the most appropriate statistical methodology may result in loss of power or uninterpretable results.

3.10. Censored data

A relatively unique data characteristic is presented by some DNT tests that measure the time it takes for a certain event to occur. These include tests such as passive avoidance and Morris water maze, both of which are often used as tests of learning and memory. In a passive avoidance test, the animal learns to associate entering into an area with a shock, and subsequently learns to not enter that area and thereby passively avoids the shock. In the Morris water maze, the animal swims in a tank of water and eventually learns the location of an escape platform under the surface, using the spatial cues throughout the room. In these types of studies, latency is a measure of time-to-event, for example, crossing into shock compartment, or finding the escape platform. There is typically a cut-off for the latency measure, to provide a practical upper limit for the test. The consequence of this maximum latency is possibly more important for passive avoidance, since the longer latency is the dependent measure which directly addresses how well the animal has learned. On the other hand, swim latencies decrease as the animal learns, and acquisition is evaluated by shorter latencies. The rest of this section will address passive avoidance latencies, but are applicable to any test that provides these types of data.

In statistical terms, the latency of an animal is considered to be right-censored if the event never occurs in the allotted time. For that particular animal, the actual time to cross over is not known, only that it did not occur. The latency measurement combines two data characteristics: a time-dependent quantitative measure of how long it takes for an animal to cross over and a dichotomous categorical measure of whether or not an animal crossed over in the allotted time. A common approach to statistical analysis is to ignore the censored data as such and conduct standard analyses for quantitative data. For example, in a study for which the allotted time for crossover was 180 s, the analysis of latency might be conducted with an analysis of variance or possibly a nonparametric Kruskal–Wallis test. The maximum allotted time of 180 s is substituted for those values that are censored. One obvious problem with this approach is that for the censored values, the animal never actually crossed over. Therefore their inclusion as 180 s will underestimate the true latency.

The effect of censored data on the statistical analysis may be negligible in the first passive avoidance training trial, where most or all animals cross over and receive the shock. However, with increasing trials there may actually be as much or more information in the proportion of animals that do not cross over than there is in those that do.

A statistical analysis approach designed for time-dependent data that includes censored data is referred to as "survival analysis." The name stems from its common use in the analysis of survival data in which the time-to-event of interest is death. For example, in a two-year rodent carcinogenicity study, a certain proportion of the data is censored in that not all of the animals die in the two-year period. The same concepts can be applied to latency in the passive avoidance test in which the time-to-event of interest is the crossover; not all animals will cross over in the allotted time.

One simple but intuitive method for applying the survival analysis approach to latency data is provided by the Kaplan–Meier (KM) method of estimating the survival curve. The KM method estimates the probability that the time-to-crossover is greater than a specific time *t*. For those values that are not censored, the KM estimate is simply the proportion of animals that have not crossed over by time *t*. For censored values KM is undefined. The KM estimates can then be compared among groups using either the log-rank test or generalized version of the Wilcoxon test (for example, SAS PROC LIFETEST).

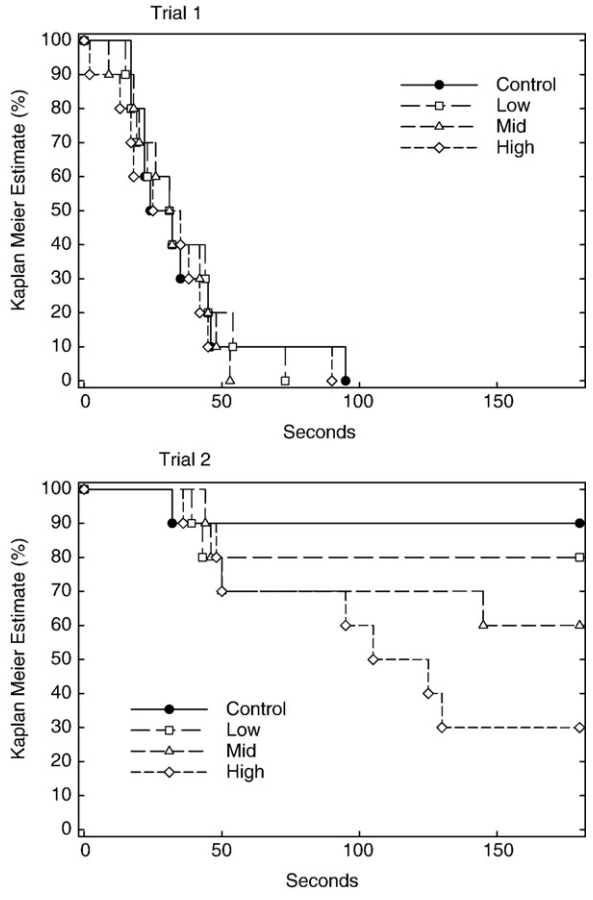

Fig. 3. Kaplan–Meier method of estimation of censored data. Passive avoidance data are notorious for censored data, because commonly a significant proportion of animals fail to cross on trial 2, after the initial shock, and are simply assigned the latency at which the trial is terminated. The top panel gives data for trial 1, initial latency to cross into the shock box. Since no aversive stimuli have as yet been administered, all animals cross over within the allotted time. Such data (trial 1) can be analyzed using either conventional ANOVA on latency scores or the Kaplan–Meier approach, with closely comparable results. Panel 2 shows latency to cross into the shock box on trial 2, following being shocked in this box on trial 1. Many rats freeze and are timed out, producing a substantial amount of censored data. In this example, using a conventional analysis suggests that there is no treatment effect, while the Kaplan–Meier approach using log-rank survival tests shows that in fact there is a significant main effect of treatment.

Consider a study in which 10 animals from each of 4 groups were subject to the passive avoidance test with the maximum allotted time of 180 s. Fig. 3 shows the KM estimates for the two trials. Table 6 presents the data summaries and results of several different statistical approaches. The curves show graphically an estimate of the proportion of animals that have not yet crossed over at a given time point during the test. For example, 30% of the control animals had not yet crossed over 40 s into trial 1. The curves illustrate graphically that there is little distinction between the latencies of the four groups in the first trial. However, differences in latency during trial 2 are exemplified by the notable separation between curves beginning at about 50 s into the trial.

In trial 1 all animals from all groups crossed over and thus there were no censored values. The latency data from trial 1 is conducive to a parametric analysis approach and the ANOVA results in a nonsignificant overall test for group effect ($p=0.977$). Likewise, the nonparametric Kruskal–Wallis test resulted in a nonsignificant overall test of group effect ($p=0.943$). Although there were no censored values in trial 1, survival analysis methodology can still be applied. The log-rank test for comparing group KM estimates of survival for trial 1 was not statistically significant ($p=0.964$) as was a dose-related trend based on the log-rank ($p=0.598$). Thus there is general agreement in the different analyses in the presence of no censored values.

In trial 2, numerous animals did not cross over resulting in a large proportion of censored values. Note that because the data are heavily skewed toward the censored values, the means are distorted and the medians provide a truer picture of the relative latency of the four groups. Although only 3 out of 10 (30%) group 4 animals completed the 180 s without crossing compared with 9 out 10 (90%) group 1 animals, only the log-rank survival tests resulted in statistically significant results. The smaller *p*-value for the log-rank trend test compared to the test for homogeneity of groups reflects the greater power for the former when responses occur in a dose-related fashion. As illustrated by the KM graphs, such was the case for latencies in trial 2.

Table 6
Example results of analyses of data with censored values

	Trial 1		Trial 2	
Treatment	Mean±SD (median)	Censored	Mean±SD (median)	Censored
Group	Latency	Values	Latency	Values
C	36±23 (28)	0	165±47 (180)	9
L	35±19 (32)	0	152±59 (180)	8
M	32±14 (32)	0	137±63 (180)	7
H	33±25 (30)	0	113±56 (115)	3

Analyses	*p*-values	
ANOVA (H*)	0.977	0.209
Kruskal–Wallis (H)	0.943	0.114
Log-rank (H)	0.964	0.048
log-rank trend (T)	0.598	0.007

(H*) = test of homogeneity of groups.
(T) = test of dose-related trend.
Data for passive avoidance latency to cross, $n=20$/group (treatment groups, C = control, L = low, M = mid, H = high dose). Maximum test time 180 s, i.e., censored value.

For time-to-event data that include no or very few censored values, the survival test methodology provides results similar to parametric and nonparametric counterparts that do not account for censored values. However, because they are designed to account for censored values, the survival test methodology may be more appropriate for analysis of time-to-event data in the presence of numerous censored values.

For illustration purposes of the concept of censored data, the example presented here evaluated only the comparison of treatment groups for an individual sex. More complex parametric survival regression models can be utilized to evaluate additional factors such as sex, as can logistic regression techniques. However, in practice, the nonparametric survival methodology described here provides an efficient means by which to analyze and interpret results from data that include censored values.

4. Interpretation of statistical analyses

4.1. Considerations

The first requirement before attempting to interpret the statistical analysis is to look at the data. Tables of means and standard deviations (or other indices of central tendency and dispersion) offer one way of summarizing the data, but figures are also strongly encouraged because they often give a better appreciation than tables and offer a more concise way of presenting a large data set than tables. Going to the individual animal data may also be necessary to make sense of the statistical analysis. Close evaluation of the data can reveal patterns of effects that may not reach statistical significance, as well as a statistically significant difference when several baseline groups (i.e., before treatment) are compared. This may be very challenging when confronted with data for hundreds of rats on as many as 20 tests, but it is possible. As Bolles [4] wrote,

> "Perhaps the most basic thing I have to say is that rather than looking at the statistics, you should look at the data." (p. 83).

The statistical analysis should be viewed as a tool rather than an end in itself. It should help the investigator to formulate a conclusion or a hypothesis. If all the conditions have been filled out for a study to be a hypothesis testing study (i.e., *a priori* identification of all the elements of a study), conclusions can be drawn from the study. If some of these conditions are missing, a hypothesis or several hypotheses can be generated, but these would have to be tested in another sample.

Several considerations should be given about the meaning of a *p*-value. First of all, the data should be carefully looked at and the investigator should attempt to reconcile the obtained *p*-value with the examination of the data. Too often the statistical analysis has been erroneously reduced to a "star-seeking" exercise without which no decision could be taken [51].

As stated above, it is important to realize that the *p*-value is in part a reflection of the sample size so that a statistically significant relationship, for example, could in all likelihood be found between intelligence and shoe size if the sample size were large enough. Consideration should always be given to the strength of

association between dependent and independent variables. Yates [71] had the following comment on the significance tests:

> "… it has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are estimating." (p. 32).

p-values should serve as guidance. A statistically significant *p*-value does not indicate the magnitude of a difference; neither does it provide any information about the replicability of the finding. A statistically significant *p*-value can be considered as nonsignificant by the investigator for a number of reasons alluded too above (e.g., previous information, multiplicity problem, large sample size, lack of biological plausibility, etc.). Similarly, a nonsignificant statistical *p*-value can be construed as significant for the same or other reasons (e.g., previous information, low power, small sample size, biological consistency, etc.). It should be kept in mind that the distinction between 0.049 and 0.051 as statistically significant and nonsignificant, respectively, is untenable and is as arbitrary as the significance criterion used most often (i.e., 0.05).

A clear distinction needs to be made between statistical significance and biological significance. A finding can be statistically significant, and have no biological significance, and vice-versa. For example, a 2% difference in body weight with sample sizes of 50 rats per sex at the end of a chronic toxicity study may be statistically significant and real, but is this difference biologically important? Conversely, a study with 8 rats in the control and treated groups may not reach the traditional statistical significance level, but the data are consistent with previous information collected on this compound, are biologically plausible and provide support for the *a priori* stated scientific hypothesis. Should such data be declared toxicologically significant? The *p*-value alone should never overrule a decision based on consistency, examination of the data, strength of association, dose-response relationship, toxicological relevance, biological plausibility, etc.

4.2. Reporting and presentation

Sorting the data into a table format that allows evaluation of groups by dose and by test time allows a look at several trends, but most often graphs are more helpful, as remarked under Section 4.1. Whenever means of continuous data are graphed, always include some measure of data variability (e.g., standard deviation, semi-interquartile range). Without such information, it is impossible to determine what differences might exist between treatment groups. It should be noticed, however, that the standard error of the mean is not a measure of individual data variability, but that it represents the spread of the sampling distribution of the mean. In other words, it is a measure of uncertainty in the average value of all possible samples of the same size taken from a given population. The standard error of the mean gives some idea about the accuracy of the mean.

It is also possible to plot means of ranked data, but it should be understood that the group means do not have any real statistical meaning. Plots of binary data are easier to understand when presented as incidence, or percent of treatment group showing the effect. The graphs should reflect the level of the analyses. For example, if the total counts are analyzed for motor activity, plot the total counts as a function of dose.

When data are compiled in tables, results of the statistical analyses should also be indicated. This should include degrees of freedom, *F*-values or other appropriate statistics, as well as exact *p*-values where available. Depending on the table format, it may be possible to include that information on the same table, or present it separately. At the very least, this information should be presented for every endpoint (whether significant or not), for the overall analyses and for any subsequent step-down analyses.

The following real-life example is presented to show the visual power of graphs (Fig. 4). Chemical *X* decreased the open-field activity, and the overall statistical analysis was significant. *Post-hoc* comparisons showed that the low and the high doses were different from control, but not the middle dose. One could conclude that since this did not show a dose-response relationship, it was an anomalous finding. The graph of the mean group data actually showed more variability in the low dose group, but the graph of the individual values clearly explained the outcomes. The low dose had a few subjects with higher activity, but most of the group had lower activity; the middle dose group was less variable but had somewhat higher counts. An interpretation of these data could be that all doses were different from control, even though the middle dose did not quite reach significance. In this case, the graphical representation of the individual values was critical in making such a determination.

Current US EPA reporting requirements for the DNT guidelines [62] include: 1) Tables of data for each test animal (including ID number of each pup and the litter from which it came) for each day tested/observed, body weight and scores on

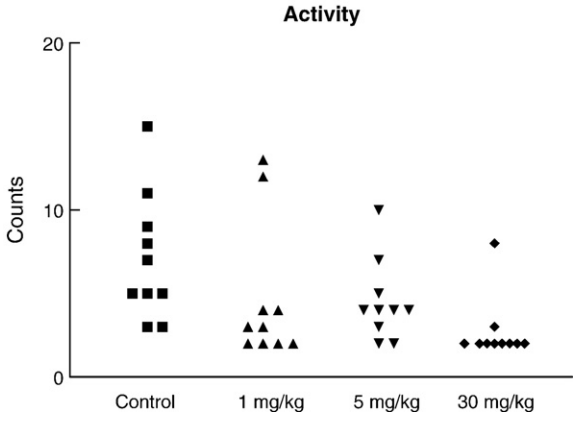

Fig. 4. The visual power of graphed data. In this graph, actual open field data are illustrated for controls and three increasing dose levels of chemical *X*. The ANOVA gives a significant main effect of treatment, but conventional *post-hoc* tests suggest that only the low and high doses differ from controls. Barring actual visual inspection of data, in the absence of a medium dose effect, experimenters might be tempted to reject such a non-monotonic dose effect as statistical error. However graphing all data points as shown tells another story — the effect is also seen at the middle dose, but escapes statistical significance.

each developmental landmark, total session activity counts and intrasession subtotals, data for each repeated trial/session showing acquisition and retention scores on the tests of learning and memory, time and cause of death, and any neurological signs seen, and for neuropathology, a list of structures examined as well as the locations, nature, frequency, and extent of lesions, and brain weights; 2) Summaries for each treatment and control group including the number of animals at the initiation of the study, body weight of the dams during gestation and lactation, litter size and mean weight at birth, number and percentage of animals showing each abnormal signs at each observation time, and mean and standard deviation for each continuous endpoint (e.g., body weight, activity counts, startle responses, etc.) at each observation time, and for neuropathology, the number of animals in which any lesions was found, the frequency and average grade for severity of the lesion for each animal, and the values of all morphometric measurements made for each animal listed by treatment group; 3) Evaluation of data, including "… appropriate statistical analyses. The choice of analyses should consider tests appropriate to the experimental design and need adjustments for multiple comparisons." Clearly, the suggestions made in this paper (including variability bars in graphs, *p*-values in tables and/or summaries, looking at individual data) are very much in line with the reporting requirements already set forth by the US EPA. The inclusion of such information would aid in the review of data sets by regulators or peer reviewers.

5. Summary of recommendations

Our general recommendations are as follows:

- evaluate the data and not just the level of significance. Wherever possible the graphing of data is encouraged;
- include sex and sex-by-treatment interaction as factors in the analysis of all dependent variables collected on both sexes;
- litter must remain a factor in analysis throughout the study, not just in young animals. If each litter contributes animals of both sexes, then sex and the sex by treatment interaction must be analyzed as a correlated variable;
- to assess adaptation or changes that occur over time, repeated-measures methodologies should be utilized to evaluate the effects of different dose groups while accounting for the correlated data resulting from multiple measurements on the same animal;
- clearly indicate the information to be provided by each procedure. Do not test hypotheses on the same data that generated them. Always test them in a new data set;
- identify the type of data that each endpoint represents (e.g., body weight is a continuous endpoint, degree of lacrimation is an ordinal endpoint) and utilize the most appropriate statistical methodology for that type. Describe in detail all the statistical analyses in the protocol and the study report;
- running a statistical analysis after seeing the data can only generate hypotheses to be confirmed, but no conclusions;
- consider strategies to address the multiplicity problem in the study, or at least indicate how the multiplicity problem will be addressed in the study. Use complex multifactorial statistical analyses to substantially reduce multiplicity of significance tests. This approach has the added benefit of allowing tests of interaction terms which are not addressed in simpler designs;
- provide the total count of derived *p*-values (significant and nonsignificant). Preferably, report exact *p*-values with their associated *F* values and degrees of freedom, as appropriate. If *p*-values are adjusted for multiple comparisons, unadjusted *p* values should also be provided and the method of adjustment should be identified;
- address the strength of association between dependent and independent variables;
- use pairwise comparison procedures that are optimal for the questions being addressed (e.g., use Dunnett's if the only concern is to compare treatment mean values to control);
- consider the use of statistical methodology specifically designed for censored data when the data include a substantial number of such measurements (e.g., passive avoidance latencies in which crossover never occurs).

Acknowledgements

This paper is the product of an ILSI Research Foundation/Risk Science Institute expert working group formed to address the evaluation and interpretation of neurodevelopmental endpoints for human health risk assessment. The authors are a subgroup of the expert working group. All members of the expert working group were included in numerous discussions of the draft manuscript and were part of an extensive review process prior to submittal for publication. The members of the expert working group include: Dr. Jane Adams, University of Massachusetts — Boston; Dr. John M. Balbus, Environmental Defense; Dr. David Bellinger, Harvard Medical School; Dr. Kevin Crofton, US Environmental Protection Agency; Dr. Penny Fenner-Crisp, Consultant; Dr. J. Edward Fisher, Jr., US Food and Drug Administration; Dr. John Foss, Charles River Laboratories; Dr. Les Freshwater, BioSTAT Consultants, Inc.; Dr. Scott Hancock, Health Canada; Dr. Ulla Hass, Danish Institute of Food Safety and Toxicology; Dr. Keith Hazelden, Huntingdon Life Sciences, Inc.; Dr. R. Robert Holson, New Mexico Tech; Dr. Edward D. Levin, Duke University Medical Center; Dr. Susan Makris, US Environmental Protection Agency; Dr. Tim Marrs, Edentox Associates; Dr. Jacques P. J. Maurissen, The Dow Chemical Company; Dr. Elizabeth Mendez, US Environmental Protection Agency; Dr. Angelo Moretto, Universita di Padova, Italy (current affiliation, Department of Occupational and Environmental Medicine, International Center for Pesticides and Health Prevention, University of Milan and Luigi Sacco Hospital, Milan, Italy); Dr. Virginia Moser, US Environmental Protection Agency; Dr. Sherry Parker, OrbusNeich Medical, Inc.; Dr. Whang Phang, US Environmental Protection Agency; Dr. Kathleen Raffaele, US Environmental Protection Agency; Professor David Ray, University of Nottingham; Dr. Louis (Gino) Scarano, US Environmental Protection Agency; Dr. Larry Sheets, Bayer CropScience; Dr. Thomas J. Sobotka, US Food and Drug Administration (current affiliation, retired); Dr. Sonya K. Sobrian, Howard University College of Medicine; Dr.

Rochelle W. Tyl, RTI International; Dr. Isabel Walls, ILSI Research Foundation. ILSI Research Foundation gratefully acknowledges the entities listed above that supported this project by allowing their staff to serve on the expert working group. This report was developed under the Cooperative Agreement R-83049601 between ILSI Research Foundation and the US Environmental Protection Agency Office of Pesticide Programs and Cooperative Agreement X-82916701 between ILSI Research Foundation and the US Environmental Protection Agency Office of Pollution Prevention and Toxics.

References

[1] H.K. Bates, R.H. McKee, G.S. Bieler, T.H. Gardiner, M.W. Gill, D.E. Strother, L.W. Masten, Developmental neurotoxicity evaluation of orally administered isopropanol in rats, Fundam. Appl. Toxicol. 22 (1994) 152–158.
[2] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc., B 57 (1995) 289–300.
[3] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Ann. Stat., 29 (2001) 1165–1188.
[4] R.C. Bolles, Why you should avoid statistics, Biol. Psychiatry 23 (1988) 79–85.
[5] G.E.P. Box, Non-normality and test on variance, Biometrika 40 (1953) 318–335.
[6] G.E.P. Box, Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification, Ann. Math. Stat. 25 (1954) 290–302.
[7] G.E.P. Box, Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effect of inequality of variance and of correlation of errors in the two-way classification, Ann. Math. Stat. 25 (1954) 484–498.
[8] J.L. Bussiere, L.M. Hardy, M. Peterson, J.A. Foss, R.H. Garman, A.M. Hoberman, M.S. Christian, Lack of developmental neurotoxicity of MN rgp 120/HIV-1 administered subcutaneously to neonatal rats, Toxicol. Sci. 48 (1999) 90–99.
[9] S.G. Carmer, M.R. Swanson, An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods, J. Am. Stat. Assoc. 68 (1973) 66–74.
[10] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, New Jersey, 1988.
[11] J.P. Creason, Data evaluation and statistical analysis of functional observational battery data using a linear models approach, J. Amer. Coll. Toxicol. 8 (1989) 157–169.
[12] C.W. Dunnett, A multiple comparison procedures for comparing several treatments with a control, J. Am. Stat. Assoc. 75 (1955) 789–795.
[13] J.D. Elashoff, Analysis of covariance: a delicate instrument, Am. Educ. Res. J. 6 (1969) 383–401.
[14] S.P. Ellis, M.D. Underwood, V. Arango, J.J. Mann, Mixed models and multiple comparisons in analysis of human neurochemical maps, Psychiatry Res. 99 (2000) 111–119.
[15] B.A. Elswick, F. Welsch, D.B. Janszen, Effect of different sampling designs on outcome of endocrine disruptor studies, Reprod. Toxicol. 14 (2000) 359–367.
[16] P. Eriksson, D. Von Rosen, H. Viberg, A. Fredriksson, Developmental toxicology in the neonatal mouse: the use of randomly selected individuals as statistical unit compared to the litter in mice neonatally exposed to PBDE 99, Toxicologist 84 (2005) 219–220.
[17] J.L. Fleiss, Statistical Methods for Rates and Proportions, John Wiley & Sons, Inc., New York, 1981.
[18] D.A. Freedman, A note on screening regression equations, Am. Stat. 37 (1983) 152–155.
[19] M.W. Gill, M.S. Swanson, S.R. Murphy, G.P. Bailey, Two-generation reproduction and developmental neurotoxicity study with sodium chlorite in the rat, J. Appl. Toxicol. 20 (2000) 291–303.
[20] G.V. Glass, P.D. Peckham, J.R. Sanders, Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance, Rev. Educ. Res. 42 (1972) 237–288.
[21] J.E. Grizzle, C.F. Starmer, G.G. Koch, Analysis of categorical data by linear models, Biometrics 25 (1969) 489–504.
[22] J.K. Haseman, A.J. Bailer, R.L. Kodell, R. Morris, K. Portier, Statistical issues in the analysis of low-dose endocrine disruptor data, Toxicol. Sci. 61 (2001) 201–210.
[23] W.L. Hays, Statistics for the Social Sciences, Holt, Rinehart and Winston, New York, 1973.
[24] J.F. Heyse, Technical issues in the design and analysis of teratology/reproduction Studies, 1987, Annual Meeting of the Biostatistics Subsection of the Pharmaceutical Manufacturer's Association, 1987 San Diego, CA.
[25] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Statist. 6 (1979) 65–70.
[26] R.R. Holson, B. Pearce, Principles and pitfalls in the analysis of prenatal treatment effects in multiparous species, Neurotoxicol. Teratol. 14 (1992) 221–228.
[27] G. Hommel, A comparison of two modified Bonferroni procedures, Biometrika 76 (1989) 624–625.
[28] H.J. Keselman, P.A. Games, J.C. Rogan, Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus test statistic, Psychol. Bull. 86 (1979) 884–888.
[29] G.G. Koch, J.R. Landis, J.L. Freeman, D.H. Freeman Jr., R.C. Lehnen, A general methodology for the analysis of experiments with repeated measurement of categorical data, Biometrics 33 (1977) 133–158.
[30] J. Landis, E. Heyman, G. Koch, Average partial association in three-way contingency tables: a review and discussion of alternative tests, Int. Stat. Rev. 46 (1978) 237–254.
[31] J.R. Levin, Overcoming feelings of powerlessness in "aging" researchers: a primer on statistical power in analysis of variance designs, Psychol. Aging 12 (1997) 84–106.
[32] K.J. Levy, A Monte Carlo study of analysis of covariance under violations of the assumptions of normality an equal regression slopes, Educ. Psychol. Meas. 40 (1980) 835–840.
[33] K.Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.
[34] E.F. Lindquist, Design and Analysis of Experiments in Psychology and Education, Houghton Mifflin Company, Boston, 1956.
[35] R.C. Littell, G.A. Milliken, W.W. Stroup, R.D. Wolfinger, SAS System for Mixed Models, SAS Institute Inc., Cary, NC, 1996.
[36] R.C. Littell, J.P. Endergast, R. Natarajan, Modelling covariance structure in the analysis of repeated measures data, Stat. Med. 19 (2000) 1793–1819.
[37] R.C. Littell, W.W. Stroup, R.J. Freund, SAS for Linear Models, SAS Institute Inc., Cary, NC, 2002.
[38] F.M. Lord, Large-sample covariance analysis when the control variable is fallible, J. Am. Stat. Assoc. 55 (1960) 307–321.
[39] N. Mantel, Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure, J. Am. Stat. Assoc. 58 (1963) 690–700.
[40] N. Mantel, Assessing laboratory evidence for neoplastic activity, Biometrics 36 (1980) 381–399.
[41] N. Mantel, W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease, J. Natl. Cancer Inst. 22 (1959) 719–748.
[42] J.P. Maurissen, A.M. Hoberman, R.H. Garman, T.R. Hanley Jr., Lack of selective developmental neurotoxicity in rat pups from dams treated by gavage with chlorpyrifos, Toxicol. Sci. 57 (2000) 250–263.
[43] G.P. McCabe, D.S. Moore, Introduction to the Practice of Statistics, W. H. Freeman and Co., New York, 1999.
[44] R.B. McCall, M.I. Appelbaum, Bias in the analysis of repeated-measures designs: some alternative approaches, Child Dev. 44 (1973) 401–415.
[45] K.E. Muller, C.N. Barton, V.A. Benignus, Recommendations for appropriate statistical practice in toxicologic experiments, Neurotoxicology 5 (1984) 113–125.
[46] K.R. Murphy, B. Myors, Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, Lawrence Erlbaum Associates, New Jersey, 2004.
[47] NIST/SEMATECH e-Handbook of Statistical Methods, 2006.

[48] R.G. O'Brien, M.K. Kaiser, MANOVA method for analyzing repeated measures designs: an extensive primer, Psychol. Bull. 97 (1985) 316–333.
[49] G.T. Pryor, E.T. Uyeno, H.A. Tilson, C.L. Mitchell, Assessment of chemicals using a battery of neurobehavioral tests: a comparative study, Neurobehav. Toxicol. Teratol. 5 (1983) 91–117.
[50] T.A. Ryan, Comment on "protecting the overall rate of type I errors for pairwise comparison with an omnibus test statistic", Psychol. Bull. 88 (1980) 354–355.
[51] D.S. Salsburg, The religion of statistics as practiced in medical journals, Am. Stat. 39 (1985) 220–223.
[52] D.J. Saville, Multiple comparison procedures: the practical solution, Am. Stat. 44 (1990) 174–180.
[53] N.C. Schwertman, N.J. Carter, A more practical Scheffe-type multiple comparison procedure for commonly encountered numbers of comparisons, J. Stat. Comput. Simul. 53 (1995) 181–196.
[54] E. Shirley, The analysis of organ weight data, Toxicology 8 (1977) 13–22.
[55] L.P. Spear, S.E. File, Methodological considerations in neurobehavioral teratology, Pharmacol. Biochem. Behav. 55 (1996) 455–457.
[56] S.S. Stevens, On the theory of scales of measurement, Science 103 (1946) 677–680.
[57] M.E. Stokes, C.S. Davis, G.G. Koch, Categorical Data Analysis Using the SAS System, SAS Institute, Inc., Cary, NC, 2006.
[58] B.G. Tabachnik, L.S. Fidell, Using Multivariate Statistics, Allyn and Bacon, 2001.
[59] R.N. Tamura, J. Buelke-Sam, The use of repeated measures analyses in developmental toxicology studies, Neurotoxicol. Teratol. 14 (1992) 205–210.
[60] L.E. Toothaker, Multiple Comparisons for Researchers, Sage Publications, London, 1991.
[61] J.W. Tukey, J.L. Ciminera, J.F. Heyse, Testing the statistical certainty of a response to increasing doses of a drug, Biometrics 41 (1985) 295–301.
[62] U.S. EPA, Health Effects Guidelines OPPTS 870.6300 Developmental Neurotoxicity Study, 1998.
[63] M.W. Vasey, J.F. Thayer, The continuing problem of false positives in repeated measures ANOVA in psychophysiology: a multivariate solution, Psychophysiology 24 (1987) 479–486.
[64] P.F. Velleman, L. Wilkinson, Nominal, ordinal, interval, and ratio typologies are misleading, Am. Stat. 47 (1993) 65–72.
[65] A.J. Vickers, How many repeated measures in repeated measures designs? Statistical issues for comparative trials, BMC. Med. Res. Methodol. 3 (2003) 22.
[66] P.E. Wainwright, Issues of design and analysis relating to the use of multiparous species in developmental nutritional studies, J. Nutr. 128 (1998) 661–663.
[67] J.H. Ware, F. Mosteller, F. Delgado, C. Donnelly, J.A. Ingelfinger, Medical Uses of Statistics, NEJM Books, Boston, Massachusetts, 1992.
[68] D.A. Williams, A test for differences between treatment means when several dose levels are compared with a zero dose control, Biometrics 27 (1971) 103–117.
[69] L.D. Wise, L.R. Gordon, K.A. Soper, D.M. Duchai, R.E. Morrissey, Developmental neurotoxicity evaluation of acrylamide in Sprague–Dawley rats, Neurotoxicol. Teratol. 17 (1995) 189–198.
[70] L.D. Wise, H.L. Allen, C.M. Hoe, D.R. Verbeke, R.J. Gerson, Developmental neurotoxicity evaluation of the avermectin pesticide, emamectin benzoate, in Sprague–Dawley rats, Neurotoxicol. Teratol. 19 (1997) 315–326.
[71] F. Yates, The influence of statistical methods for research workers on the development of the science of statistics, J. Am. Stat. Assoc. 46 (1951) 19–34.
[72] R.G. York, J. Barnett Jr., W.R. Brown, R.H. Garman, D.R. Mattie, D. Dodd, A rat neurodevelopmental evaluation of offspring, including evaluation of adult and neonatal thyroid, from mothers treated with ammonium perchlorate in drinking water, Int. J. Toxicol. 23 (2004) 191–214.