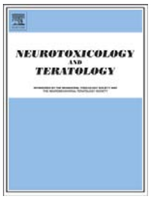
Neurotoxicology and Teratology 32 (2010) 121–123

Contents lists available at ScienceDirect

Neurotoxicology and Teratology

journal homepage: www.elsevier.com/locate/neutera



Open Peer Commentary

Practical considerations on the design, execution and analysis of developmental neurotoxicity studies to be published in *Neurotoxicology and Teratology*

Jacques Maurissen [*],[1]

ARTICLE INFO

Article history:
Received 11 June 2009
Received in revised form 12 August 2009
Accepted 9 September 2009
Available online 13 September 2009

Keywords:
Recommendation
Author
Reviewer
Disclosure
Interpretation
Multiplicity

ABSTRACT

The present article is an attempt to use the "ILSI Research Foundation/Risk Science Institute Reports from the Expert Working Group on Neurodevelopmental Endpoints" (2008) to help improve the quality of the manuscripts submitted to *Neurotoxicology and Teratology*, as well as the quality of their review. The points discussed in the present paper have been encountered during the peer-review process. A number of recommendations are proposed on the basis of general principles (clarity, full disclosure, and evidence-based interpretation) to help authors and reviewers. Clarity of language is a prerequisite, but clarity of purpose is essential. The methodology and statistical analysis for each dependent variable should be unambiguously presented and justified. Full disclosure encompasses a range of topics, such as defining the sample size for each experiment, clearly distinguishing between hypothesis-testing and hypothesis-generating (e.g., *a priori* vs. *a posteriori* analyses), clearly defining the statistical model appropriate to the study design and questions (e.g., repeated-measure approach), recognizing and addressing the multiplicity problem (e.g., conceptual unit for the error rate), identifying the appropriate unit for statistical analysis (e.g., litter), addressing the results of all analyses (e.g., "negative" results are important). Data interpretation should be evidence-based and not exceed the limits of the findings.

1. Introduction

In 2004 the International Life Science Institute (ILSI) gathered expert working groups to address a number of questions raised about developmental neurotoxicity endpoints that focused on four topics, including the use of positive controls, test variability, statistical analysis, and data interpretation. In 2008, *Neurotoxicology and Teratology* dedicated a special issue to the "ILSI Research Foundation/Risk Science Institute Reports from the Expert Working Group on Neurodevelopmental Endpoints" [2,4,12,14]. Even though the ILSI reports were mainly concerned with addressing and improving the quality of the developmental neurotoxicity studies as per regulatory Guidelines [11,15], they can be used as a resource and guidance for the conceptualization, design, and analysis of other neurotoxicity studies and their submission to the journal. Many of the principles and recommendations that came out of these reports can be used to improve the quality of the research and manuscripts submitted to this journal for developmental and adult neurotoxicity studies in general [4]. The intent of this article is to provide suggestions to improve the quality and the review of submissions to NT&T. Improvement in manuscript quality and clarity is very important not only for the advancement of the field, but also for the potential impact of developmental and adult neurotoxicity on other fields, future funding, as well as public health and policy.

A number of comments mainly based on the previous ILSI reports, have been assembled below to help authors and reviewers address or avoid a number of points of discussion most frequently encountered during the peer-review process. This list is not exhaustive. Comments are provided below on these points as they relate to different sections of a manuscript.

2. General

Communication clarity is of the utmost importance for all manuscripts. Manuscripts written by people whose primary language is not English may benefit from being reviewed by a scientist whose native language is English to assure proper use of syntax and text clarity.

Foremost, the question(s) to be addressed by the study should be clearly stated. Ideally, each study should have a clear statement of purpose and ask specific questions, be designed to address these questions and generate interpretable data that will answer them. Full disclosure of all the study parameters and all results should be

[*] Permanent address. JPM NeuroTox, LLC, 2415 N Woodland Estates Drive, Midland, MI 48642, USA. Tel.: +1 989 636 7145; fax: +1 989 636 7121.
E-mail address: JMaurissen@Yahoo.com.

[1] "On behalf of the Publications Committee of the Neurobehavioral Teratology Society".

0892-0362/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ntt.2009.09.002

provided. The reader should feel confident after reading the method section that he/she could duplicate the study.

3. Material and methods

3.1. Subjects

The total number of subjects (sample size) and their biological and demographic characteristics (e.g., age, weight, and sex) should be given, as well as their allocations to each of the different treatment groups. The method used for group assignment should also be named and described, e.g. randomization using a computerized pseudorandom number generator, sequential assignment in order of time of birth, assignment by exposure, etc. The assignment of subjects to groups should be unambiguous. For data analysis it is important to distinguish between "subjects" and "experimental units". The distinction between them is based on the concept of "independence". Whereas subjects, as units of measurement, do not have to be independent, experimental units have to. In other words, the response of one experimental unit should be independent of the response of another one. The distinction is crucial for statistical analysis because the experimental unit (e.g., litter, school), not the subject, should be the unit of analysis for proper data interpretation [3–5]. Positive and negative controls should be used, as appropriate.

3.2. Apparatus

The equipment should be described with enough details so that the reader would be able to visualize it, comprehend its functioning and limitations (e.g., minimum detection limits), and have a clear understanding of the nature of the variables it generates and of the confounders potentially affecting them. Equipment calibration should be addressed. Appropriate units of measurement should be used. For example, the decibel (dB) is a dimensionless unit; therefore, to express the absolute magnitude of a physical quantity, a reference level has to be specified, such as dB(A), dB(HL), dB re 20 μPa or other. It is necessary to report any calibration issues, equipment malfunction, measurement errors and other events potentially affecting the data integrity and interpretation.

3.3. Procedures and study design

Some extraneous factors that are not treatment factors (e.g., time of the day) can impact the response of interest and confounds its interpretation. Such factors should be accounted for either in the study design or in the data analysis. The study should be designed such that confounders are eliminated or, if unavoidable, can be assessed to enable appropriate data interpretation (e.g., performance in a learning task without elementary motor assessment should be avoided). When the subjects cannot all be tested at the same time or with the same device, the testing sequence and/or the subject allocation to the devices should be clearly described so that the reader can understand when and where males and females, as well as members of all the treatment groups are being tested. Ideally, males and females should be tested at the same time if a comparison between them is desirable; otherwise, testing males and females at different times, for example, leads to a situation where sex and time would be inextricably intertwined and the interpretation of the effect would be hopelessly confounded. No statistical method could disentangle the sex and time effects. The same is also true for all the treatment groups. Usually, the most practical solution is to counterbalance for all relevant factors within and across test sessions. Device equivalency should be documented as part of the calibration process. Minor differences between devices can also be addressed through counterbalanced assignment of subjects to devices (e.g., auditory startle devices, homogenizers). As a general rule, the investigator conducting the testing should be "blind" to treatment, particularly when behavioral observations are recorded. When several investigators have contributed to observations, the inter-observer reliability should be addressed.

Both hypothesis-generating and hypothesis-testing types of studies have their place in science. Whereas the former explores data in search of unexpected clues or novel hypotheses, the latter assesses *a priori* hypotheses according to a predetermined design and analysis. It should be made very clear whether the manuscript is describing an hypothesis-generating or an hypothesis-testing study [17]. In Muller et al.'s words [10], an hypothesis-testing study "demands that all hypothesis tests have been specified exactly and [the] α [value] fixed *a priori*." (p. 114). If any aspect of the design or analysis is modified *post hoc* (i.e., after examination of the data), that part of the study must be considered as hypothesis-generating. Presenting *post hoc* analyses as planned *a priori* evaluations is inappropriate because it is misleading, it distorts the true probability value of the analysis, it gives credence to a potentially illusory event and makes it look like a well-founded, expected effect [1,7].

3.4. Statistical analysis

The manuscript should unambiguously state which type of analysis is used for each of the variables analyzed, and give a rationale for its selection. General statements, such as "ANOVAs were used where appropriate", are not informative and should not be used. The statistical analysis should be guided by the experimental design (e.g., repeated-measure designs should be analyzed by a repeated-measure type of analysis) and should be reported in detail (e.g., repeated-measures ANOVA by the univariate approach with Greenhouse–Geisser correction, or by the multivariate approach [with selected test criterion], or by the mixed-effects model approach [with model selection for goodness of fit, for example]). All the between- and within-subject factors should be clearly identified.

Data collected from the same subject at different times under different technical conditions (e.g., when settings on the data collection system, or the assessment tools have changed from one time to another) may not be pooled into one repeated-measure analysis if the data are not normalized to the same testing conditions. For example, when startle data are collected in young and adult animals, the amplification setting is typically changed to optimize the data accuracy within the appropriate recording range. As a result the raw amplitude readings do not reflect the actual differences between young and adult animals, and cannot be meaningfully incorporated into a single repeated-measure analysis. However, if the data from both young and adult animals can be expressed by reference to the same scale (i.e., the same amplification setting) following the appropriate transformation, the data can be analyzed in a single repeated-measure analysis.

The authors should address the multiplicity problem associated with the generation of many p values [17]. The accepted Type I error rate (α) should be reported and the unit for its conceptual error rate identified (e.g., per comparison, familywise, experimentwise, etc. [8,9,13]). When power is defined, it should similarly be identified by its conceptual unit (e.g., one-pair, per-pair, all-pairs power [4,13]). The magnitude of a p value does not reflect the degree of relation between dependent and independent variables. Special indices measure strength of association, and authors are strongly encouraged to use them [4]. It is more meaningful to compare effect sizes rather than p values.

When litters or groups, such as schools, are used in the study, these factors should be used as the unit of statistical analysis [3–5]. When data are censored (i.e., an arbitrary maximal or minimum value is assigned to a variable, such as a score of 3 min if no response is given; often seen in passive avoidance tests as an example), the use of statistical methodology especially designed to analyze such data should be considered. Given that the tests designed to analyze censored data are few, mostly distribution free, and do not typically have the flexibility of parametric models, a general linear model can be considered especially when the number of censored data points is small. In such a case, the rationale for the selected choice should be provided. More details can be found in Holson et al. [4] and in Muller et al. [10].

4. Results

As stated above, whereas hypothesis-testing and hypothesis-generating studies have different legitimate purposes, it is wrong and misleading to present an hypothesis-generating study as hypothesis-testing [1,4,7,10]. All analyses should be reported. It is strongly recommended that exact p values be provided [16]. Expressing them as inequalities (e.g., $p<0.05$) or tabular "*" not only unnecessarily reduces their informational content, but can lead to different interpretations when, in fact, the p values may be the same for all practical purposes (e.g., $p=0.049$ or $p=0.051$ would be reported either as $p<0.05$ or non-significant, or be represented by the presence or absence of an "*"). If all p values are not provided, their total count (declared statistically significant or non-significant) should be provided to help the reader understand the extent of the analyses performed during the study. A single "extremely" statistically significant difference provides less scientific evidence than valid data replication [6] or complementary findings across test variables.

4.1. Tables and figures

Indices of central tendency should always be accompanied by an index of dispersion of raw data. When a mean is reported, the standard deviation (SD) should also be provided in preference to the standard error of the mean (SEM). The SD quantifies scatter of the raw data whereas the SEM estimates the range where the true mean of the population lies. The SEM is always smaller than the SD, and decreases with increased sample size, whereas the SD gets more precise with increased sample size, but can increase or decrease. The authors can always justify the use of the SEM if they feel that the SEM is more appropriate. As much as possible, graph axes should not be truncated; if truncation is needed, the legend should indicate so. Graphs should reflect data variability, not distort it. A graph is worth a thousand p values.

5. Conclusion

The over-interpretation of results is not acceptable, e.g., the interpretation should be logical, reasonable, evidence-based and not exceed the limits of the findings. Caution and conservativeness are proper here. Hypothesis-generating studies result in new hypotheses, and have a different purpose than hypothesis-testing studies. Statistical analysis should be used as a guide for data interpretation. It should be seen as an incentive for going back to the data and looking at them in the context of other data to detect patterns. The final interpretation should also be evidence-based and, for example, include dose–response relationships and its anticipated shape, available information on the toxicity of the chemical under study, its mode of action, historical control data, etc. [4,14]. A statistically significant difference from controls may either be a false positive, or may indicate a real treatment effect, which may or may not be adverse (i.e., statistical vs. biological significance). Similarly, a non-statistically significant difference may either be a false negative, or may hide a real treatment effect, which may or may not be adverse (e.g. small expected effect in study with reduced power). Interpretation should address biological plausibility and the biological and/or functional significance of the results. It should be based more on trends than on fragments. If the data are relevant to human health, it is strongly recommended that the authors compare the dose levels and route of administration used in their study with known or estimated human exposures or exposure guidelines (e.g., threshold limit values). Unreported analyses and selective data presentation and citations can mislead the reader and misdirect future science. As such, these practices are scientifically and ethically unacceptable.

6. Conflict of interest statement

The author has no conflicts of interest to declare.

Acknowledgments

The author acknowledges the constructive critiques provided by the Publication Committee of the Neurobehavioral Teratology Society (N. Day, C.D. Driscoll, J.E. Fisher, R.R. Holson, W.J. Pizzi, L.T. Singer, T.J. Sobotka, S.K. Sobrian, S.L. Wood). The Publication Committee has fully endorsed this paper as a supplement to the Instructions to Authors for *Neurotoxicology and Teratology* (except for R.R.H. who could not be contacted).

References

[1] J.C. Bailar III, Science, statistics, and deception, Ann. Intern. Med. 104 (1986) 259–260.

[2] K.M. Crofton, J.A. Foss, U. Hass, K.F. Jensen, E.D. Levin, S.P. Parker, Undertaking positive control studies as part of developmental neurotoxicity testing. A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints, Neurotoxicol. Teratol. 30 (2008) 266–287.

[3] J.K. Haseman, M.D. Hogan, Selection of the experimental unit in teratology studies, Teratology 12 (1975) 165–172.

[4] R.R. Holson, L. Freshwater, J.P.J. Maurissen, V.C. Moser, W. Phang, 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, Neurotoxicol. Teratol. 30 (2008) 326–348.

[5] R.R. Holson, B. Pearce, Principles and pitfalls in the analysis of prenatal treatment effects in multiparous species, Neurotoxicol. Teratol. 14 (1992) 221–228.

[6] D.H. Johnson, The insignificance of statistical significance testing, J. Wildl. Manag. 63 (3) (1999) 763–772.

[7] N.L. Kerr, HARKing: hypothesizing after the results are known, Pers. Soc. Psychol. Rev. 2 (1998) 196–217.

[8] R.E. Kirk, Experimental design: procedures for the behavioral sciences, Brooks/Cole Publishing Company, Belmont, CA, 1968 pp. 82–86.

[9] J. Ludbrook, Multiple comparison procedures updated, Clin. Exp. Pharmacol. Physiol. 25 (1998) 1032–1037.

[10] K.E. Muller, C.N. Barton, V.A. Benignus, Recommendations for appropriate statistical analysis in toxicologic experiments, NeuroToxicol. 5 (2) (1984) 113–126.

[11] Organisation for Economic Cooperation and Development (OECD, Guideline for the Testing of Chemicals, Test No. 426: Developmental Neurotoxicity Study, 2007.

[12] K.C. Raffaele, J.E. Fisher Jr., S. Hancock, K. Hazelden, S.K. Sobrian, Determining normal variability in a developmental neurotoxicity test. A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints, Neurotoxicol. Teratol. 30 (2008) 288–325.

[13] J.P. Shaffer, Multiple hypothesis testing, Annu. Rev. Psychol. 46 (1995) 561–584.

[14] R.W. Tyl, K. Crofton, A. Moretto, V. Moser, L.P. Sheets, T.J. Sobotka, A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints, Neurotoxicol. Teratol. 30 (2008) 349–381.

[15] USEPA, Health Effects Guidelines OPPTS 870.6300, Developmental Neurotoxicity Study, 1998.

[16] J.H. Ware, F. Mosteller, F. Delgado, C. Donnelly, J.A. Ingelfinger, P values, in: J.C. Bailar III, F. Mosteller (Eds.), Medical Uses of Statistics, 2nd Edition, NEJM Books, Boston, Massachusetts, 1992, pp. 181–200.

[17] L. Wilkinson, Statistical methods in psychology journals. Guidelines and explanations, Am. Psychol. 54 (1999) 594–604.