Comment on "Stress in Puberty Unmasks Latent Neuropathological Consequences of Prenatal Immune Activation in Mice"

Stanley E. Lazic

Science **340**, 811 (2013);
DOI: 10.1126/science.1237793

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 16, 2013):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/content/340/6134/811.2.full.html

A list of selected additional articles on the Science Web sites **related to this article** can be found at:
http://www.sciencemag.org/content/340/6134/811.2.full.html#related

This article **cites 5 articles**, 1 of which can be accessed free:
http://www.sciencemag.org/content/340/6134/811.2.full.html#ref-list-1

This article has been **cited by** 1 articles hosted by HighWire Press; see:
http://www.sciencemag.org/content/340/6134/811.2.full.html#related-urls

This article appears in the following **subject collections:**
Neuroscience
http://www.sciencemag.org/cgi/collection/neuroscience
Technical Comments
http://www.sciencemag.org/cgi/collection/tech_comment

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Comment on "Stress in Puberty Unmasks Latent Neuropathological Consequences of Prenatal Immune Activation in Mice"

Stanley E. Lazic

Giovanoli *et al*. (Reports, 1 March 2013, p. 1095) applied an immune challenge to pregnant females, and therefore to all offspring, and subsequently applied stress to offspring on a per cage basis. The data, however, were analyzed as a completely randomized design, which is inappropriate given these restrictions on randomization. This will increase both false positives and false negatives.

Giovanoli *et al*. (*1*) used an experimental design where pregnant female mice were assigned to receive an injection of poly (I:C) (polyriboinosinic-polyribocytidilic acid) or vehicle control (presumably the procedure was randomized, but this was not mentioned). After weaning, littermates were housed two to three per cage and were then (presumably randomly) assigned on a per cage basis to a stress or control condition. There were thus two randomizations, one at the level of litter (all of the animals within each litter were randomized together) and one at the level of the cage (all of the animals within a cage were randomized together). This is known as a split-plot or split-unit design in the statistical literature and is characterized by experimental treatments being applied at different levels (litters and cages within litters) and by restrictions on complete randomization (*2*–*4*). Split-unit and nested designs in general are common in neuroscience research but are typically analyzed as if they were completely randomized designs, where each individual animal can be randomly assigned to the various treatment combinations. For example, only 9% of studies using the valproic acid model of autism correctly identified the hierarchical nature of the design and performed a sensible analysis (*5*). Giovanoli *et al*. provided a detailed description of their methods (statistical test used, sample size per group, and degrees of freedom), which makes it clear that the analyses were inappropriate.

Ignoring the actual design of the experiment and analyzing it as a completely randomized design (e.g., two-way analysis of variance, as the authors have done) has two negative consequences. The first is that the sample size (*n*) is artificially inflated. When comparing the effect of poly(I:C) versus control, the sample size is the number of litters (~6 to 10, depending on the cohort) and not the number of offspring (~35 to 50). The result is that *P* values will be too small and the number of false positives will be greater than the nominal 5% level. The second consequence is that differences between litters will end up as unexplained variation and thus true differences will be harder to detect (reduced power). It is irrelevant that the scientific interest is in the individual offspring, or that previous studies using the poly(I:C) model have been published without comment. When treatments are applied to whole litters, regulatory authorities require that litters (not individual offspring) are treated as the experimental units (*6*, *7*), as the numeric output from a statistical analysis (i.e., *P* values) is only meaningful if the analysis is correctly applied. The same applies to cages when testing the effects of stress. Strictly speaking, since the two to three animals within a cage were randomized as a group and not individually, they are considered subsamples, pseudoreplicates, or "technical replicates," and they do not contribute to the overall sample size (*2*–*5*, *8*). One might argue that individual animals within a cage could have been individually randomized to stress or control conditions, or that the offspring were randomized to cages, and then the cages were randomized to one of the two conditions, which would allow the sample size to be the number of offspring for the stress versus control comparison. One approach might be to test for cage effects, and if not significant (often at a less-strict 0.1 level), then the cage can be excluded as a variable in the analysis, and the number of offspring can be used as *n* for the stress versus control comparison. Although there is some debate about this "sometimes pooling" approach, there is often little to be gained (*9*) and some very strong opinions against it (*2*).

Giovanoli *et al*. stated that an equal number of offspring from each litter were used in the stress and control groups, and this wise design ensures that differences between litters will not be confounded with the effect of stress. However, unless the analysis appropriately reflects the design, the *P* values and confidence intervals will not be valid. One might be tempted to examine the figures, note the difference between group means relative to the error bars, and think that the results are obvious for some outcomes (the calculation of *P* values being a box-ticking exercise to satisfy the editors and reviewers). However, mean ± SEM graphs are not suitable for such visual estimates with split-unit designs because the error bars reflect a tangled mess of sources of variation, with the incorrect *n* used in their calculation, and which do not correspond to the error variance used in the corresponding statistical test. Some effects were large, and the overall qualitative conclusion may not change if reanalyzed with an appropriate model. However, the authors performed a thorough examination of these animals and had many outcome variables with modest effect sizes. In addition, nonsignificant effects may become significant when previously unexplained variation (i.e., noise) is attributed to litters or cages and thus removed.

The authors have developed an interesting animal model and designed a nice series of experiments; however, the analyses do not lead to valid *P* values, making interpretation of the results difficult. This paper will likely influence many future studies; therefore, it is important that correct inferences are made. More important, it is imperative that future investigators are aware of the design and analysis issues when applying treatments to whole litters rather than to individual offspring (*5*). Other groups may not be as careful in balancing litters across treatment groups, which would confound treatment effects with litter effects and lead to biased inferences, lack of reproducibility, and delays in translating preclinical results into the clinic (*10*). Given the complex nature of these experimental designs and subsequent analyses, collaboration with statistical colleagues during the planning and analysis phases would be highly beneficial.

In Silico Lead Discovery, Novartis Institutes for Biomedical Research, Basel, Switzerland.

E-mail: stan.lazic@cantab.net

References

1. S. Giovanoli *et al*., *Science* **339**, 1095 (2013).
2. R. Mead, S. G. Gilmour, A. Mead, *Statistical Principles for the Design of Experiments: Applications to Real Experiments* (Cambridge Univ. Press, Cambridge, UK, 2012).
3. G. E. P. Box, J. S. Hunter, W. G. Hunter, *Statistics for Experimenters: Design, Innovation, and Discovery* (Wiley-Blackwell, Hoboken, ed. 2, 2005).
4. G. Casella, *Statistical Design* (Springer, New York, 2008).
5. S. E. Lazic, L. Essioux, *BMC Neurosci.* **14**, 37 (2013).
6. International Conference on Harmonisation. Detection of Toxicity to Reproduction for Medicinal Products and Toxicity to Male Fertility. S5(R2) (2000); www.ich.org/products/guidelines/safety/safety-single/article/detection-of-toxicity-to-reproduction-for-medicinal-products-toxicity-to-male-fertility.html.
7. Organisation for Economic Cooperation and Development, OECD Guidelines for the Testing of Chemicals: Developmental Neurotoxicity Study (2007); www.oecdbookshop.org/oecd/display.asp?lang=EN&sf1=identifiers&st1=5l4fg25mnkxs.
8. S. E. Lazic, *BMC Neurosci.* **11**, 5 (2010).
9. D. G. Janky, *Am. Stat.* **54**, 269 (2000).
10. S. C. Landis *et al*., *Nature* **490**, 187 (2012).

13 March 2013; accepted 24 April 2013
10.1126/science.1237793