Translational neuroscience requires better design and analysis of preclinical studies

Stanley E. Lazic
Bioinformatics and Exploratory Data Analysis, F. Hoffmann-La Roche,
4070 Basel, Switzerland
stan.lazic@cantab.net

Abstract

Animal models are often used to obtain a better understanding of psychiatric, neurodegenerative, and neurodevelopmental disorders. Despite many years of research, these models have not led to many novel therapies or treatments. Translating results between species will always be difficult, and it is argued that inappropriate statistical analyses, failure to identify the experimental unit, lack of random assignment to treatment conditions, and unblinded assessment of outcomes contribute to the low rate of translating preclinical in vivo studies into successful therapies. It is known that these shortcomings can generate biased estimates, too many false positives and false negatives, and unreproducible results. These issues have been raised repeatedly, but have largely gone unheeded by scientists. Two recommendations are made to improve the situation.

Numerous animal models (lesion, transgenic, knockout, selective breeding, etc.) have been developed for a variety of psychiatric, neurodegenerative, and neurodevelopmental disorders. While many of these models have been helpful for understanding disease pathology, they have been less useful for discovering potential therapies, or for predicting which treatments will be useful in the clinic. In short, translation from in vivo animal models (typically rodent) has been poor, despite many years of research and effort. There are many reasons for this, including the inherent difference in biology between rodents and humans,[1] particularly relating to higher cognitive functions. In addition, there is the ever-present question of whether a particular animal model is even suitable; whether it captures the disease process of interest or faithfully mimics key aspects of the human condition. While important, these two considerations will be put aside, and the focus will be on the design, execution, and analysis of preclinical in vivo studies, and the role that these have on translational neuroscience.[2,3,4,5,6,7,8]

There are two issues that will be discussed. The first relates to litters, in particular, the natural litter-to-litter variation (i.e. litter effects) as well as designs where an experimental treatment is applied to whole litters rather than to the individual animals, usually because the treatment is applied to pregnant females and therefore to all of the offspring. The second issue relates to the quality of reported results, and how this relates to the quality of the underlying study.

Litter effects are ubiquitous, large, and important

It is known that on many variables and across many species, monozygotic twins are more similar than dizygotic twins, which are more similar than non-twin siblings, and which in turn are more similar than two unrelated individuals. What has not been fully appreciated is that all of the standard statistical methods (e.g. t-test, analysis of variance, regression, non-parametric methods) assume that the data come from unrelated individuals. However, rodents from the same litter are effectively dizyogtic twins; they share a large number of genes as well as prenatal and early postnatal environments, and therefore studies need to be designed and analysed in such a way that differences between litters do not bias or confound the results.[7,9,10,11,12,13,14,15,16,17] More specifically, this relates to the assumption of independence of the measurements. For example, measuring blood pressure (BP) from the left and right arm of ten randomly sampled and unrelated people only provides ten independent measurements of BP, not twenty. This is because the left and right BP values will be highly correlated—if the BP value measured from a person's left arm is high, then so will the value measured from their right arm. Similarly, animals within a litter will tend to have values that are more similar (i.e. correlated) than animals from different litters. This lack of independence needs to be handled appropriately in the analysis. Many animal models are derived from highly inbred strains, and this results in reduced genotypic and phenotypic variation. This is a different issue and unrelated to lack of independence. It does not mean that animals “are all the same” and that differences between litters do not exist.

This is not a minor issue that only statistical pedants worry about, with little practical importance for scientists. Holson and Pearce showed twenty years ago that if three treated and three control litters are used, with two offspring per litter (total number of offspring = 12) then the false positive rate (Type I error) is 20% rather than 5% (based on actual body weight data from their experiment).[10] Furthermore, the false positive rate increases with the number of offspring per litter: if the number of offspring per litter is 12 (total number of offspring = 72) then the false positive rate is 80%. The error rate is also influenced by the relative variability between and within litters, and thus will vary for each outcome. Nevertheless, given that papers report the results of multiple tests (multiple outcome variables and multiple comparisons), we can expect the literature to be rife with false positive results. It may seem paradoxical, but in addition to too many false positives, treating the individual offspring as the experimental unit can also lead to low power (too many false negatives) when true effects exist.[10,11] This is because differences between litters is unexplained variation, and thus the “noise” in the data is increased, potentially masking true treatment effects. In the same paper, Holson and Pearce reported that only 30% of papers in the behavioural neurotoxicology literature correctly accounted for litter effects. A subsequent study in 1997 using forty litters found “significant litter effects… in varying degrees, for almost every behavioural, morphologic, and neuroendocrine measure; they were evident across indices of neural, adrenal, thyroid, and immunologic functioning in adulthood”[11] (and see references therein for further studies supporting this conclusion). This paper also noted that 34% of papers in *Developmental Psychobiology* correctly accounted for litter effects and only 15% of papers in related journals. This issue has been brought up repeatedly for almost forty years,[9] but has largely been ignored by neuroscientists. One can only speculate on the number of erroneous conclusions that have been reached, and what an incredible waste of resources this has been.

One might argue that when many studies are conducted, including replications within and between labs, the evidence will eventually converge to the “truth”, and therefore these considerations are only of minor interest. Unfortunately, there is no guarantee of such convergence, as the literature on the superoxide dismutase (SOD1) transgenic mouse model of amyotrophic lateral sclerosis (ALS) demonstrates. Several treatments showed efficacy in this model and were advanced to clinical trials, where they proved to be ineffective.[18] A subsequent large-scale (>5000 mice) and properly executed study did not replicate the previous findings.[19] This study also identified litter as an important variable which affected survival (the main outcome), and which was not taken into account in the earlier studies. The authors also demonstrated how false positive results can arise with inappropriate experimental designs and analyses. Litter effects were not the only contributing factor; a meta-analysis of the preclinical SOD1 literature revealed that only 31% of studies reported randomly assigning treatments, and even fewer reported blind assessment of outcomes.[20] Lack of randomisation and blinding are known to overstate the size of treatment effects.[3,5,6,21] In addition, there was evidence of publication bias, where studies with positive results were more likely to be published.[20] Thus, the combination of poor experimental design, analysis, and publication bias contributed to numerous incorrect decisions regarding treatment efficacy.

When treatments are applied to whole litters

Some disease models have a distinctive experimental design feature: the treatment is applied to pregnant females (and therefore to all of the unborn animals within that female), but the scientific interest is in the individual offspring (Figure 1). This design is common in toxicology and nutrition studies, as well as psychiatric studies which examine the effects of maternal stress on offspring and in the valproic acid (VPA) model of autism. Here, the "treatment" refers to the experimental manipulation that induces the disease features, and it does not refer to a therapeutic treatment. Difficulties arise because the experimental unit ("n"; defined as the smallest physical unit that can be randomly assigned to a treatment condition) in these cases are the pregnant dams and not the individual offspring.[7,9,10,11,12,13,14,15,16,17] In other words, the sample size is the number of dams, and the offspring are considered subsamples, much like the left and right kidney from a single animal do not represent a sample size of two (n = 1, but there are two replicate measurements). This may come as a surprise, and it is irrelevant that the scientific interest is in the offspring, or that the offspring eventually become individual entities (unlike the kidneys). Regulatory authorities have clear guidelines on the matter;[22] for example, the Organisation for Economic Co-operation and Development (OECD) has made a firm statement in their guidelines for chemical testing: "Developmental studies using multiparous species where multiple pups per litter are tested should include the litter in the statistical model to guard against an inflated Type I error rates. The statistical unit of measure should be the litter and not the pup. Experiments should be designed such that littermates are not treated as independent observations [p. 12]".[23] There is a restriction on randomisation because only whole litters can be assigned to the treatment or control conditions, and this needs to be reflected in the analysis. This could be accomplished by using only one animal per litter, and then standard methods (e.g. t-test, ANOVA, etc.) can be employed. This is not the most efficient design, unless the excess animals can be used for other experiments. A second option is to use more than one animal per litter, and then average the values of the animals within a litter. These mean values are then taken forward and can be analysed using standard methods. A third option is to use multiple animals per litter, and then analysis is performed with a nested or hierarchical model, which properly handles the structure of the data and avoids artificially inflating the sample size (also known as pseudoreplication[7,24]). This method is preferred because litter is entered as a variable in the analysis and the magnitude of the litter effect can be quantified. When using the first two options, it is clear that to increase the sample size and thus power, the number of litters needs to be increased. This is also true for the third option, but may not be so readily apparent [pp. 3–4].[16]

A related design issue is that greater statistical power can be achieved when littermates are used to test a therapeutic compound versus a placebo. If the therapeutic treatment is applied to the individual animals postnatally, then the individual animal is the experimental unit *for this comparison*. This is referred to as a split-plot design and has more than one type of experimental unit (litters for some comparisons and individual animals for others). These studies therefore require careful planning and analysis, but scientists are rarely introduced to these methods during the course of their training. It is not surprising therefore that the quality of many such studies is low.

Quality of preclinical studies

Previous studies have shown that general quality of preclinical animal experiments is low,[3,5,8,25,26,27] and there is increasing evidence that the quality of the design, analysis, and interpretation of basic neuroscience research is no better. For example, Nieuwenhuis et al. recently reported that 50% of papers misinterpret interaction effects.[28] In addition, the issue of "inflated n", or pseudoreplication, shows up in other guises,[7,29] and whole fields can misattribute cause-and-effect relationships.[30,31] There is also the concept of "researcher degrees of freedom", which refers to the flexibility that scientists have in choosing the main outcome variables, statistical models, data transformations, how outliers are handled, when to stop collecting data, and what is reported in the final paper.[32] Various permutations of the above options greatly increases the chances that at least something will be statistically significant, and this is what tends to get reported as the sole analysis that was conducted. Given the above, it is not surprising that the pharmaceutical industry has difficulty reproducing many published results.[33,34]

We now take a closer look at some of these issues in the VPA model of autism. It is a relatively new model and thus one would hope that the lessons of the past (e.g. SOD1 model) have been considered. Thirty-five primary research articles were identified (up to the end of 2011) which injected pregnant dams with VPA and then analysed the effects in the offspring. One study was excluded as key information was located in the supplementary material, but this was not available online.[35] Of the remaining studies, only four (12%) reported randomly assigning pregnant females to the VPA or control group. Many studies also used only a subset of the offspring from each litter, but often it was not mentioned how the offspring were selected. It is possible that many studies did actually randomise the dams and then randomly select a subset of the offspring, but simply did not report it; however, randomisation is such a crucial aspect for the validity of the results that it seems odd to omit commenting on it. Only six studies (18%) reported that the investigator was blind to the experimental condition when collecting the data. For fourteen studies (41%) it was not possible to determine how many dams were actually used (i.e. the sample size), and in four studies (12%) the number of offspring used were not indicated. Ten studies (29%) did not indicate whether both male and female offspring were used. No study mentioned performing a power analysis to determine a suitable sample size to detect effects of a given magnitude—but this is probably fortuitous because only three studies (9%) correctly identified the experimental unit! One of these used a nested design,[36] the second mentioned that litter was the experimental unit,[37] and the third used one animal from each litter.[38] A list of studies can be found in the supplementary material.

A number of papers had additional statistical or experimental design issues, ranging from trivial (e.g. reporting total degrees of freedom rather than residual degrees of freedom for an F-statistic) to serious. These include treating individual neurons as the experimental unit, which is distressingly common in electrophysiological studies but just as inappropriate as treating blood pressure values taken from left and right arms as $n = 2$, or chopping a single liver sample into ten pieces and thinking that measuring the expression of a gene in each piece gives $n = 10$.[7] If only it were so easy; clinical trials could be conducted with tens of patients rather than hundreds or thousands. Simply take a blood sample from twenty patients, divide it into twenty aliquots, and you now have an apparent sample size of $20 \times 20 = 400$! Need more power? Just divide those blood samples into fifty aliquots, now the apparent sample size is $20 \times 50 = 1000$; error bars are minuscule, p-values are < 0.05, and publication in a top journal looks promising. Regulatory authorities are not fooled by such stratagems, but is seems many journal editors and peer-reviewers are.

In addition, the wording of two studies[39,40] suggests that control dams did not receive a vehicle injection, and thus any differences between groups may be partly due to the stress of handling and injection. For some studies, the reported degrees of freedom did not correspond to what would be expected based on the verbal description of the analysis, and a number of studies did not correctly distinguish between “within-subjects” and “between-subjects” effects.

Solutions

A number of solutions have been proposed in the past,[41] but some are difficult to implement. For example, scientists could develop the skills and knowledge to design and analyse experiments. But if this has not happened by now, it is unlikely to happen in the foreseeable future. An other option is to have a statistician associated with most preclinical studies. However, there are not enough statisticians to meet this demand, and this type of “project support” is often viewed by academic statisticians as a secondary activity. More manuscripts could go to a statistician for peer review, but again, a lack of statisticians, and particularly a lack of statisticians with the appropriate subject matter knowledge (just as it is difficult to do good science without a knowledge of statistics, it is difficult to perform a good analysis without knowledge of the science). Some have recommended registering animal studies,[42,43,44,45] but as argued elsewhere,[7] this does not fit well with the goals of preclinical discovery research.

There are however two practical solutions which can be easily implemented. The first is to follow the ARRIVE (Animals in Research: Reporting In Vivo Experiments) guidelines for reporting experimental results.[46] More specifically, items 6 (Study design), 10 (Sample size), 11 (Allocating animals to experimental groups), and 13 (Statistical methods) should a be mandatory requirement for all publications involving animals and should be included as a separate checklist that is submitted along with the manuscript, much like a conflict of interest or a transfer of copyright form. This would be signed by the authors (either all authors or only the primary/senior author), would go out for peer review, and would also be published online along with the manuscript. This would have many beneficial effects. First, it would make it easier to spot any design and analysis issues by reviewers, editors, and other readers (who may only skim the published paper for the main findings and assume that the design and analyses were appropriate). Currently, this can only be done by reading through verbal descriptions located in many different parts of a manuscript and supplementary material—assuming that the relevant information is even provided. Second, and more importantly, if scientists are *required* to comment on how they randomised treatment allocation, or how they ensured that assessment of outcomes was blinded, then they will conduct their experiments accordingly if they plan on publishing in a journal that has these reporting requirements. Similarly, if researchers are required to state what the experimental unit is (e.g. litter, cage, individual animal, etc.), then they will be prompted to think hard about the issue and design better experiments, or seek advice. This recommendation will not only improve the quality of reporting, but it will also improve the quality of experiments, which is the real benefit. In addition, many design weaknesses and analytical flaws can be spotted at a glance, and editors and reviewers can give this the necessary weight when evaluating manuscripts. A final benefit is that it will make quantitative reviews/meta-analyses easier, because much of the key information will be on a single page.

The second solution is to make the provision of raw data a requirement for acceptance of a manuscript; not “to make it available if someone asks for it”, which is the current requirement for many journals, but uploaded as supplementary material or hosted by a third party data repository. None of the VPA studies provided the data that the conclusions were based on, making reanalysis impossible. Remarkably, of the thirty-five studies published, only one provided the necessary information to conduct a power analysis to plan a future study,[38] and this was only because one animal per litter was used and the necessary values could be extracted from the graphs. Datasets used in preclinical animal studies are typically small, do not have confidentiality issues associated with them, are unlikely to be used for further analyses by the original authors, and have no additional intellectual property issues associated with them given that the manuscript itself has been published. It is noteworthy that many journals require microarray data to be uploaded to a publicly available repository (e.g. Gene Expression Omnibus or ArrayExpress), but not the corresponding behavioural or histological data. It is perhaps not surprising that there is a relationship between study quality and the willingness to share data.[47,48,49] Publishing raw data can be taken as a signal that researchers stand behind their data and therefore their conclusions. Funding bodies should encourage this by requiring that data arising from the grant are made publicly available (with penalties for non-adherence). There is an inherent conflict of interest between advancing knowledge and advancing one's career, and it is inevitable that the latter will occasionally come at the expense of the former. Even though most researchers are honest and well-intentioned, the appropriate structures should be put in place to ensure that appropriate design and analyses were used, and to make it easy to verify claims or to reanalyse data. Currently, it is often difficult to establish the former and almost impossible to perform the latter. Moreover, it is clear that appropriate designs and analyses are often not used, making it difficult to give the benefit of the doubt to those studies with incomplete reporting of how experiments were conducted and data analysed.

While it is difficult to quantify the extent to which poor statistical practices hinder translational (as well as fundamental) research, it is clear that a large inflation of false positive and false negative rates will only slow progress down. In addition, because of publication bias and researcher degrees of freedom, it is possible for a field to converge to the wrong answer. Experimental design and statistical issues are, in principle, fixable. Improving these will allow scientists to focus on creating and assessing the suitability of disease models and the efficacy of therapeutic interventions, which is challenging enough.

Conflict of interest

The author declares no conflicts of interest.

References

[1] Geerts H. Of mice and men: bridging the translational disconnect in CNS drug discovery. *CNS Drugs* 2009; **23**: 915–926.

[2] Sena E, van der Worp HB, Howells D, Macleod M. How can we improve the preclinical development of drugs for stroke? *Trends Neurosci* 2007; **30**: 433–439.

[3] Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PMW, Macleod M, Dirnagl U. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. *Stroke* 2008; **39**: 929–934.

[4] Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PMW, Buchan A, van der Worp HB, Traystman R, Minematsu K, Donnan GA, Howells DW. Good laboratory practice: preventing introduction of bias at the bench. *Stroke* 2009; **40**: e50–e52.

[5] Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, Hutton J, Altman DG. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* 2009; **4**: e7824.

[6] Sena ES, van der Worp HB, Bath PMW, Howells DW, Macleod MR. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol* 2010; **8**: e1000344.

[7] Lazic SE. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci* 2010; **11**: 5.

[8] Shineman DW, Basi GS, Bizon JL, Colton CA, Greenberg BD, Hollister BA, Lincecum J, Leblanc GG, Lee LBH, Luo F, Morgan D, Morse I, Refolo LM, Riddell DR, Scearce-Levie K, Sweeney P, Yrjanheikki J, Fillit HM. Accelerating drug discovery for Alzheimer's disease: best practices for preclinical animal studies. *Alzheimers Res Ther* 2011; **3**: 28.

[9] Haseman JK, Hogan MD. Selection of the experimental unit in teratology studies. *Teratology* 1975; **12**: 165–171.

[10] Holson RR, Pearce B. Principles and pitfalls in the analysis of prenatal treatment effects in multiparous species. *Neurotoxicol Teratol* 1992; **14**: 221–228.

[11] Zorrilla EP. Multiparous species present problems (and possibilities) to developmentalists. *Dev Psychobiol* 1997; **30**: 141–150.

[12] Wainwright PE. Issues of design and analysis relating to the use of multiparous species in developmental nutritional studies. *J Nutr* 1998; **128**: 661–663.

[13] Festing MFW, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J* 2002; **43**: 244–258.

[14] Festing MFW. Principles: the need for better experimental design. *Trends Pharmacol Sci* 2003; **24**: 341–345.

[15] Festing MFW. Design and statistical methods in studies using animal models of development. *ILAR J* 2006; **47**: 5–14.

[16] Casella G. *Statistical Design*, Springer, New York 2008.

[17] Maurissen J. Practical considerations on the design, execution and analysis of developmental neurotoxicity studies to be published in Neurotoxicology and Teratology. *Neurotoxicol Teratol* 2010; **32**: 121–123.

[18] Schnabel J. Neuroscience: Standard model. *Nature* 2008; **454**: 682–685.

[19] Scott S, Kranz JE, Cole J, Lincecum JM, Thompson K, Kelly N, Bostrom A, Theodoss J, Al-Nakhala BM, Vieira FG, Ramasubbu J, Heywood JA. Design, power, and interpretation of studies in the standard murine model of ALS. *Amyotroph Lateral Scler* 2008; **9**: 4–15.

[20] Benatar M. Lost in translation: treatment trials in the SOD1 mouse and in human ALS. *Neurobiol Dis* 2007; **26**: 1–13.

[21] Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of randomization and blinding affect the results? *Acad Emerg Med* 2003; **10**: 684–687.

[22] International Conference on Harmonisation. Detection of toxicity to reproduction for medicinal products and toxicity to male fertility. *S5(R2)* 1993; .

[23] OECD. Guideline for the testing of chemicals: developmental neurotoxicity study 2007; pages 1–26.

[24] Hurlbert SH. Pseudoreplication and the design of ecological field experiments. *Ecol Monogr* 1984; **54**: 187–211.

[25] Dirnagl U. Bench to bedside: the quest for quality in experimental stroke research. *J Cereb Blood Flow Metab* 2006; **26**: 1465–1478.

[26] Philip M, Benatar M, Fisher M, Savitz SI. Methodological quality of animal studies of neuroprotective agents currently in phase II/III acute ischemic stroke trials. *Stroke* 2009; **40**: 577–581.

[27] Bart van der Worp H, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. Can animal models of disease reliably inform human studies? *PLoS Med* 2010; **7**: e1000245.

[28] Nieuwenhuis S, Forstmann BU, Wagenmakers EJ. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat Neurosci* 2011; **14**: 1105–1107.

[29] Cumming G, Fidler F, Vaux DL. Error bars in experimental biology. *J Cell Biol* 2007; **177**: 7–11.

[30] Lazic SE. Relating hippocampal neurogenesis to behavior: the dangers of ignoring confounding variables. *Neurobiol Aging* 2010; **31**: 2169–2171.

[31] Lazic SE. Using causal models to distinguish between neurogenesis-dependent and -independent effects on behaviour. *J R Soc Interface* 2011; e-pub ahead of print 28 Sep 2011; doi:101098/rsif20110510; .

[32] Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 2011; **22**: 1359–1366.

[33] Mullard A. Reliability of 'new drug target' claims called into question. *Nat Rev Drug Discov* 2011; **10**: 643–644.

[34] Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 2011; **10**: 712.

[35] Rinaldi T, Silberberg G, Markram H. Hyperconnectivity of local neocortical microcircuitry induced by prenatal exposure to valproic acid. *Cereb Cortex* 2008; **18**: 763–770.

[36] Stodgell CJ, Ingram JL, O'Bara M, Tisdale BK, Nau H, Rodier PM. Induction of the homeotic gene Hoxa1 through valproic acid's teratogenic mechanism of action. *Neurotoxicol Teratol* 2006; **28**: 617–624.

[37] Kuwagata M, Ogawa T, Shioda S, Nagata T. Observation of fetal brain in a rat valproate-induced autism model: a developmental neurotoxicity study. *Int J Dev Neurosci* 2009; **27**: 399–405.

[38] Murawski NJ, Brown KL, Stanton ME. Interstimulus interval (ISI) discrimination of the conditioned eyeblink response in a rodent model of autism. *Behav Brain Res* 2009; **196**: 297–303.

[39] Rodier PM, Ingram JL, Tisdale B, Nelson S, Romano J. Embryological origin for autism: developmental anomalies of the cranial nerve motor nuclei. *J Comp Neurol* 1996; **370**: 247–261.

[40] Ingram JL, Peckham SM, Tisdale B, Rodier PM. Prenatal exposure of rats to valproic acid reproduces the cerebellar anomalies associated with autism. *Neurotoxicol Teratol* 2000; **22**: 319–324.

[41] Ioannidis JPA. Evolution and translation of research findings: from bench to where? *PLoS Clin Trials* 2006; **1**: e36.

[42] Roberts I, Kwan I, Evans P, Haig S. Does animal experimentation inform human healthcare? observations from a systematic review of international animal experiments on fluid resuscitation. *BMJ* 2002; **324**: 474–476.

[43] Sandercock P, Roberts I. Systematic reviews of animal experiments. *Lancet* 2002; **360**: 586.

[44] Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, Macleod M, Mignini LE, Jayaram P, Khan KS. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ* 2007; **334**: 197.

[45] Hackam DG. Translating animal research into clinical benefit. *BMJ* 2007; **334**: 163–164.

[46] Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010; **8**: e1000412.

[47] Wicherts JM, Borsboom D, Kats J, Molenaar D. The poor availability of psychological research data for reanalysis. *Am Psychol* 2006; **61**: 726–728.

[48] Bakker M, Wicherts JM. The (mis)reporting of statistical results in psychology journals. *Behav Res Methods* 2011; **43**: 666–678.

[49] Wicherts JM, Bakker M, Molenaar D. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One* 2011; **6**: e26828.

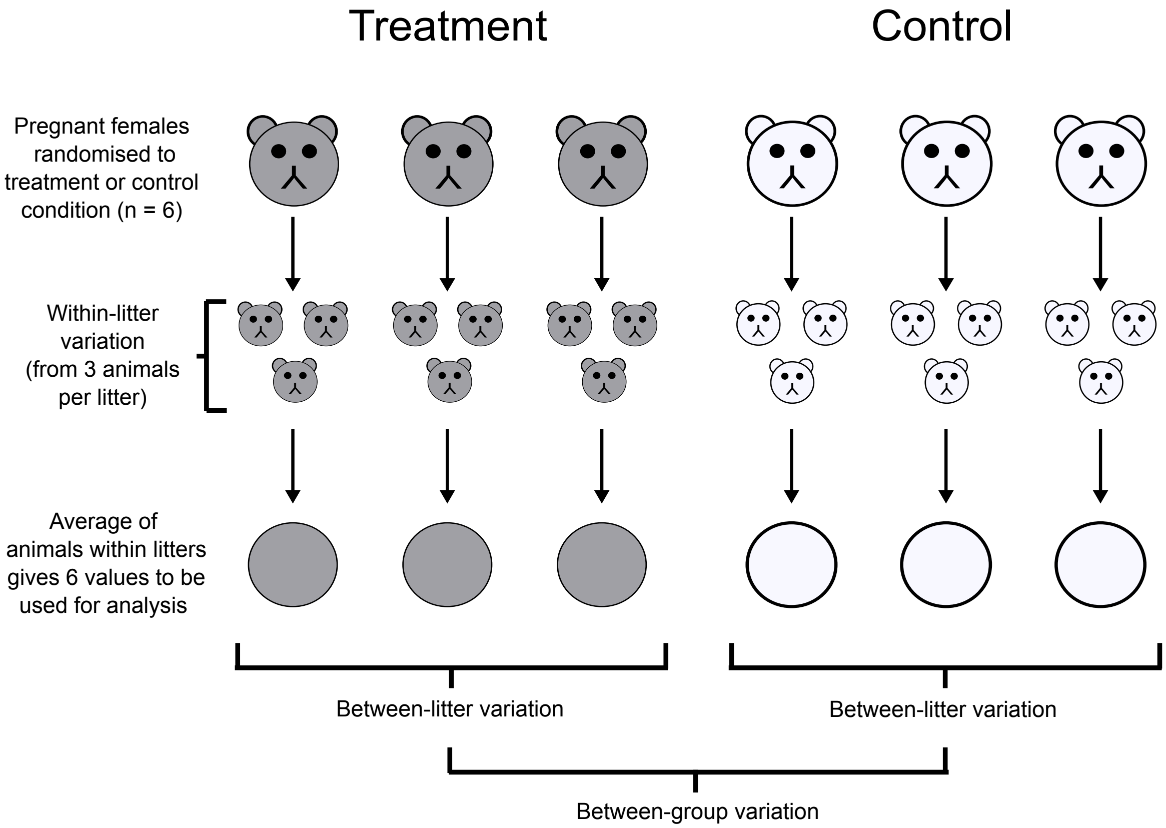
Treatment
Control
Pregnant females randomised to treatment or control condition (n = 6)
Within-litter variation (from 3 animals per litter)
Average of animals within litters gives 6 values to be used for analysis
Between-litter variation
Between-litter variation
Between-group variation

Figure 1: Defining the experimental unit. Pregnant females are the experimental units because they are randomised to the treatment (e.g. valproic acid) or control conditions, and therefore $n = 6$ in this example. The three offspring within a litter will often be more similar than offspring from different litters $\left(\frac{\text{Between}-\text{litter variation}}{\text{Within}-\text{litter variation}} > 1\right)$, and multiple offspring within a litter can be thought of as subsamples or "technical replicates", even though these are the scientific unit of interest. Only the mean of the within-litter values are important when comparing treated and control groups. Using all of the offspring without averaging will result in an inflated sample size (pseudoreplication) when using standard analyses. Instead of averaging, one could randomly select only one animal from each litter, or use a nested or hierarchical model to appropriately partition the different sources of variation. The only way to increase sample size, and thus power, is to increase the number of litters used.