Improving basic and translational science by accounting for litter-to-litter variation in animal models

Stanley E Lazic1* and Laurent Essioux2

Abstract

Background: Animals from the same litter are often more alike compared with animals from different litters. This litter-to-litter variation, or "litter effects", can influence the results in addition to the experimental factors of interest. Furthermore, sometimes an experimental treatment can only be applied to whole litters rather than to individual offspring. An example is the valproic acid (VPA) model of autism, where VPA is administered to pregnant females thereby inducing the disease phenotype in the offspring. With this type of experiment the sample size is the number of litters and not the total number of offspring. If such experiments are not appropriately designed and analysed, the results can be severely biased as well as extremely underpowered.

Results: A review of the VPA literature showed that only 9% (3/34) of studies correctly determined that the experimental unit (n) was the litter and therefore made valid statistical inferences. In addition, litter effects accounted for up to 61% (p <0.001) of the variation in behavioural outcomes, which was larger than the treatment effects. In addition, few studies reported using randomisation (12%) or blinding (18%), and none indicated that a sample size calculation or power analysis had been conducted.

Conclusions: Litter effects are common, large, and ignoring them can make replication of findings difficult and can contribute to the low rate of translating preclinical in vivo studies into successful therapies. Only a minority of studies reported using rigorous experimental methods, which is consistent with much of the preclinical in vivo literature.

Keywords: Autism, Experimental design, Litter-effects, Mixed-effects model, Multiparous, Nested model, Valproic acid

Background

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 predicting clinical efficacy. 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 recapitulates 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 and analysis of preclinical studies using multiparous species, and how this affects the validity and reproducibility of results. There are two issues that will be discussed. The first deals with 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 is the natural litter-to-litter variation that is often present, which means that the value of a measured experimental outcome can potentially be influenced by the litter that the animal came from.

*Correspondence: stan.lazic@cantab.net
1 In Silico Lead Discovery, Novartis Institutes for Biomedical Research, Basel, Switzerland
Full list of author information is available at the end of the article
Applying treatments 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). Here, the “treatment” refers to the experimental manipulation that induces the disease features, and it does not refer to a therapeutic treatment. This design is common in toxicology and nutrition studies, but is also used in neuroscience studies when examining the effects of maternal stress and in the valproic acid (VPA) model of autism. Difficulties arise because the experimental unit (“n”; defined as the smallest physical unit that can be randomly assigned to a treatment condition) is the pregnant dam and not the individual offspring [2-12]. 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, even though there are two “repeated” 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 kidneys). Regulatory authorities have clear guidelines on the matter [13,14]; 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 rate. The statistical unit of measure should be the litter and not the pup. Experiments should be designed such that litter-mates are not treated as independent observations [p. 12]” [14]. There is a restriction on randomisation because only whole litters can be assigned to the treatment or control conditions, which has implications for how studies are designed and analysed. An appropriate analysis can be conducted by using only one animal per litter (randomly selected), which allows standard methods to be used (e.g. t-test, ANOVA, etc.). This is often not the most efficient design in terms of animal usage, 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 these values for each litter. These mean values can then be taken forward and analysed using standard methods. A third option is to use multiple animals per litter, and then use a mixed-effects model for analysis, which properly handles the structure of the data.

![Figure 1 Defining the experimental unit.](image.png)

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 alike than offspring from different litters (between litter variation > 1) 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) with standard statistical analyses. Instead of averaging, one could randomly select only one animal from each litter, or use a mixed-effects 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.
(i.e. animals are nested within litters) and avoids artificially inflating the sample size (also known as pseudoreplication [12,15]). The third method is preferred to averaging values within a litter because the magnitude of the litter effect can be quantified. In addition, information on the precision of estimates will be lost by averaging, but is retained and made use of in the mixed-effects model. 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 [10, pp. 3–4], and is discussed further below.

A related design issue is that greater statistical power can be achieved when litter-mates are used to test a therapeutic compound versus a placebo. If the therapeutic can be achieved when litter-mates are used to test a therapeutic section can be used. Many animal models are derived from different litters. One can think of it as within-litter values that are more alike (i.e. correlated) than two animals sowill the valuemeasuredfromtheirrightarm.Similarly, the BP value measured from a person's left arm is high, then left and right BP values will be highly correlated—if the dent measurements of BP, not twenty. This is because the precision of estimates will be lost by averaging, but information on the precision of estimates will be lost by averaging, and thus the “noise” in the data is increased, potentially masking true treatment effects. A subsequent study using forty litters found “significant litter effects… in varying degrees, for almost every index of neural, adrenal, thyroid, and immunologic functioning in adulthood” [5] (and see references therein for further studies supporting this conclusion). Holson and Pearce reported that only 30% of papers in the behavioural neurotoxicology literature correctly accounted for litter effects [4] and Zorrilla noted that 34% of papers in Developmental Psychobiology and 15% of papers in related journals correctly accounted for litter effects [5]. This issue has been discussed repeatedly for almost forty years [2], but experimental biologists seem unaware (or choose to ignore) the importance of dealing with litter effects. One can only speculate on the number of erroneous conclusions that have been reached and the resources that have been wasted.

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 [16]. A subsequent large-scale and properly executed replication study did not support the previous findings [17]. This study also identified litter as an important variable that 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 animals to treatment conditions and even fewer reported blind assessment of outcomes [18]. Lack of randomisation and blinding are known to overstate the size of treatment effects [19-23]. In addition, there was evidence of publication bias, where studies with positive results were more likely to be published [18]. Thus, the combination of poor experimental design, analysis, and publication bias contributed to numerous incorrect decisions regarding treatment efficacy.

**General quality of preclinical animal studies**

Previous studies have shown that the general quality of the design, analysis, and interpretation of preclinical animal experiments is low [20,21,23-30]. For example, Nieuwenhuis et al. recently reported that 50% of papers in the neuroscience literature misinterpret interaction effects [31]. In addition, the issue of “inflated n”; or pseudoreplication, shows up in other guises [12,32], and whole fields can misattribute cause-and-effect relationships [33,34]. There is also the concept of “researcher degrees of freedom”, which refers to the post hoc flexibility 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 [35]. Various permutations of the above options greatly increases the chances that something statistically significant can be found, and this gets reported as the sole analysis that was conducted. Given the above concerns, it is not surprising that the pharmaceutical industry has difficulty reproducing many published results [36-39].

**Methods**

**Literature review**

Ninety-five papers were identified on PubMed using the search term “(VPA [tiab] OR valproic acid [tiab]) AND autism [tiab]” (up to the end of 2011). Reference lists from these articles were then examined for further relevant studies, and one was found. Only primary research articles that injected pregnant dams with VPA and subsequently analysed the effects in the offspring were selected (in vitro studies were excluded). A total of thirty-five studies were found, and one was excluded as key information was located in the supplementary material, but this was not available online [40]. Two key pieces of information were extracted: (1) whether the analysis correctly identified the experimental unit as the litter and (2) whether important features of good experimental design were mentioned, including randomisation, blinding, sample size calculation, and whether the total sample size (i.e. number of pregnant dams) was indicated or could be determined.

**Estimating the importance of litter-to-litter variation**

Data from Mehta et al. [41] were used to estimate the magnitude of differences between litters on a number of outcome variables. This study was chosen because it included animals from fourteen litters (five saline, nine VPA) and therefore it was possible to get a reasonable estimate of the litter-to-litter variation. In addition, the study mentioned using randomisation and blind assessment of outcomes. Half of the animals in each condition were also given MPEP (2-methyl-6-phenylethyl-pyrididine), a metabotropic glutamate receptor 5 antagonist. To assess the magnitude of the litter effects, the effect of VPA, MPEP, and sex (if relevant) were removed, and the remaining variability in the data that could be attributed to differences between litters was estimated. More specifically, models with and without a random effect of litter were compared with a likelihood ratio test. This analysis is testing whether the variance between litters is zero, and it is known that p-values will be too large because of “testing on the boundary”, and therefore the simple method of dividing the resulting p-values by two was used as recommended by Zuur et al [42]. The exact specification of the models is provided as R code in Additional file 1 and the data are provided in Additional file 2.

**Power analysis**

In these types of designs, power (the ability to detect an effect that is actually present) is influenced by the (1) number of litters, (2) variability between litters, (3) number of animals within litters, (4) variability of animals within litters, (5) difference between the means of the treatment groups (effect size), (6) significance cutoff (traditionally $\alpha = 0.05$), and (7) statistical test used. In order to illustrate the importance of the number of litters relative to the number of animals within litters and how an inappropriate analysis can lead to p-values that are too small, a power analysis was conducted with the number of litters per group varying from three to ten, and the number of animals per litter varying from one to ten. The other factors were held constant. Variability between litters (SD = 0.8803) and the variability of animals within litters (SD = 0.8142) was estimated from the locomotor activity data from Mehta et al. [41]. For each combination of litters and animals, 5000 simulated datasets were created.
with a mean difference between groups of 0.15. Once the
datasets were generated, the power for three types of anal-
yses were calculated. The first analysis averaged the values
of the animals within each litter and then groups were
compared with a t-test. The second analysis used a mixed-
effects model, and the third ignored litter and compared
all of the values with a t-test. The last analysis is incorrect
and only presented to demonstrate how artificially inflat-
sample size affects power. The power for each analysis
was determined as the proportion of tests that had \( p < 0.05 \). The R code is provided in Additional file 1 and is
adapted from Gelman and Hill [43].

Results and discussion
Low quality of the published literature
The VPA model of autism is relatively new and poten-
tial therapeutic compounds tested in this model have not
yet advanced to human trials. The opportunity therefore
exists to clean up the literature and prevent a repeat of
the SOD1 story. The main finding is that only 9% (3/34)
of studies correctly identified the experimental unit and
thus made valid inferences from the data. One study used
a nested design [44], the second mentioned that litter was
the experimental unit [45], and the third used one animal
from each litter, thus bypassing the issue [46]. In fourteen
studies (41%) it was not possible to determine the num-
ber of dams that were used (i.e. the sample size) and in
four studies (12%) the number of offspring used were not
indicated. In addition, 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 off-
spring were selected. Only six studies (18%) reported that
the investigator was blind to the experimental condition
when collecting the data. Ten studies (29%) did not indi-
cate whether both male and female offspring were used.
No study mentioned performing a power analysis to deter-
mine a suitable sample size to detect effects of a given
magnitude—but this is probably fortuitous, given that
only three studies correctly identified the experimental
unit. It is possible that many studies did use randomi-
sation and assess outcomes blindly, but simply did not
report it. However, randomisation and blinding are crucial
aspects for the validity of the results and their omission in
manuscripts suggests that they were not used. This is fur-
ther supported by studies showing that when manuscripts
do not mention using randomisation or blinding the esti-
imated effects sizes are larger compared to studies that do
mention using these methods, which is suggestive of bias
[19-23,29].

A number of papers had additional statistical or exper-
imental 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 common in electrophysiological studies, but just
as inappropriate as treating blood pressure values taken
from left and right arms as \( n = 2 \), or dissecting a sin-
gle liver sample into ten pieces and treating the expression
of a gene measured in each piece as \( n = 10 \) [12]. If it
were that easy, clinical trials could be conducted with tens
of patients rather than hundreds or thousands. Regula-
tory authorities are not fooled by such stratagems, but is
seems many journal editors and peer-reviewers are. A list
of studies can be found in Additional file 3.

Estimating the magnitude of litter effects
To illustrate the extent to which litter effects can influ-
ence the results, data originally published by Mehta et
al. [41] were used and experimental details can be found
therein. Locomotor activity in the open field is shown in
Figure 2 for nine VPA and five saline injected control lit-
ters. Half of the animals from each condition were given
MPEP (a mGluR5 receptor antagonist) or saline. Visually,
there do not appear to be differences between VPA and
control groups and there is a slight increase in activity due
to MPEP. The effect of MPEP was not significant when
litter effects were ignored (Figure 2A; p = 0.082), but it
was when adjusting for litter (Figure 2B; p = 0.011). In this
case the shift in p-value was not large, but it happened to
decrease it below the 0.05 threshold after the excess noise
caused by litter-to-litter variation was removed.

It may be difficult to determine whether litter effects
are present by simply plotting the data by litter because
they may be obscured by the experimental effects. For
a visual check, it is preferable to remove the effect of
the experimental factors first and then plot the residual
values versus litter. The y-axis for Figure 3 shows the
residuals, which are defined as the difference between the
observed locomotor activity for each animal and the value
predicted from the model containing group (VPA/saline)
and condition (MPEP/saline) as factors (from Figure 2A).
The residuals should be pure noise, centred at zero, and
should not be associated with any other variable. However,
it is clear that there are large differences between litters
(Figure 3A), indicating heterogeneity in the response from
one litter to the next. When litter effects are taken into
account, the mean of each litter is closer to zero. Also note
that variance of the residuals (\( \sigma^2 \)) is reduced by 61% when
litter is taken into account (\( p < 0.001 \)). This is shown by
the spread of the grey points around zero on the right
side of each graph, which are clustered closer together
in the second analysis. The interpretation is that litter
accounted for 61% of the previously unexplained variation
in the data. Note that it would be impossible to determine
whether litter effects are present if only one litter per treat-
ment group was used because litter and treatment would
be completely confounded.
Analysis with and without litter taken into account. Nine pregnant female C57BL/6 mice were injected with 600 mg/kg VPA subcutaneously on embryonic day 13, and five control females received vehicle injections. Half of the animals in each condition were also injected with either a mGluR5 receptor antagonist (MPEP) or saline postnatally. Total locomotor activity in the open field over a 30 min period at 8–9 weeks of age is shown. There was a slight increase in activity due to MPEP, but it was not significant when differences between litters were ignored (A; two-way ANOVA, mean difference = 0.60, F(1,44) = 3.17, p = 0.082). Adjusting for litter removed unexplained variation in the data, allowing the small difference between groups to become statistically significant (B; mixed-effects model, mean difference = 0.64, F(1,32) = 7.19, p = 0.011). Note how the values in the second graph have less variability around the group means; this increased precision leads to greater power of the statistical tests. Lines go through the mean of each group and points are jittered in the x direction.

Visualising litter-to-litter variation. The residuals represent the unexplained variation in the data after the effects of VPA and MPEP have been taken into account; they should be pure noise and therefore not associated with any other variable. However, the standard analysis (A) shows that when residuals are plotted against litter (x-axis) there are large differences between litters. In other words, there is another factor affecting the outcome besides the experimental factors of interest. The variance of the residuals (grey points on the right) is high (\(\sigma^2 = 1.29\)). The proper analysis (B) reduces the unexplained variation in the data by 61% (\(\sigma^2 = 0.50; p < 0.001\)), which can be seen by the narrower spread of the grey points around zero, and the large differences between the litters have been removed. This reduction in noise allows smaller true signals to be detected. Error bars are SEM. Litters F and L only have one observation and thus no error bars.
A similar analysis was performed for other variables and the results are displayed in Table 1. It is clear that litter-to-litter variation is important for a number of behavioural outcomes. It is also clear from Figure 3A how one could obtain false positives with an inappropriate design and analysis. Suppose an experiment was conducted with only one VPA and one saline litter, with ten animals from each, and that there is no overall effect of VPA on a particular outcome. If the experimenter happened to select Litter A (saline) and Litter M (VPA) there would be a significant increase due to VPA, but if Litter D (saline) and Litter G (VPA) were selected, there would be a significant effect in the opposite direction! There are many combinations of a single saline and VPA litter that would lead to a significant difference between conditions. Having two or three litters per group instead of one will reduce the false positive rate, but it will still be much higher than 0.05 [4]. In addition, these apparent differences would not replicate with a properly designed follow-up experiment.

How power is affected by the number of litters and animals

Figure 4 shows the power for various combinations of number of litters and number of animals per litter. This analysis is based on averaging the values for the animals within a litter and then comparing the groups with a t-test. It is clear that increasing the number of animals per litter has little effect on power (the lines in Figure 4A are nearly flat after two animals per litter), whereas increasing the number of litters results in a large increase in power. The results for the mixed-effect model are nearly identical and the results of the inappropriate analysis which ignores litter shows increasing power with increasing number of animals per litter (Additional file 4). This is false power however, and is due to an artificially inflated sample size (pseudoreplication) that will lead to many false positive results.

Some may object on ethical grounds to using so many litters and then selecting only one or a few animals from each, as there will be many additional animals that will not be used and presumably culled. Certainly all of the animals could be used, but there is almost no increase in power after three animals per litter (at least for the locomotor data) and therefore it is a poor use of time and resources to include all of the animals. One could argue therefore that it is unethical to submit a greater number of animals to the experimental procedure if they contribute little or nothing to the result. One could also argue that it is even more unethical to use any animals for a severely underpowered (or flawed) study in the first place and then to clutter the scientific literature with the results. One way to deal with the excess animals is to use them for other experiments. This requires greater planning, organisation, and coordination, but it is possible. Another option is to purchase animals from a commercial supplier and request that the animals come from different litters rather than have an in-house colony. As a side note, suppliers do not routinely provide information on the litters that the animals come from and thus an important variable is not under the experimenter’s control and cannot even be checked whether it is influencing the results.

How does litter-to-litter variation arise?

Differences between litters could exist for a variety of reasons, including shared genes and shared prenatal and early postnatal environments, but also due to age differences (it is difficult to control the time of mating), and because litters are convenient units to work with. For example, it is not unusual for litter-mates to be housed in the same cage, which means that animals within a litter also share not just their early, but also their adult environment. It is also often administratively easier to apply experimental treatments on a per cage (and thus per litter) basis rather than per animal basis. For example, animals in cage A and C are treated while cage B and D are controls. Animals may also undergo behavioural testing on a per cage basis; for example, animals are taken from the housing room to the testing room one cage at a time, tested, and then returned. Larger experiments may need to be conducted over several days and it is often easier to test all the animals in a subset of cages on each day, rather than a subset of animals from all of the cages. At the end of the experiment animals may also be killed on a per cage basis. Given that it may take many hours to kill the animals, remove the brains, collect blood, etc., the values of many outcomes such as gene expression, hormone and metabolite concentrations, and physiological parameters may change due to circadian rhythms. All of these can lead to systematic differences between litters and can thus bias results and/or add noise to the data.

There is an important distinction to be made between applying treatments to whole litters versus “natural” variation between litters. When a treatment is applied to a whole litter such as the VPA model of autism or maternal stress models, then the litter is the experimental unit and the sample size is the number of litters. Therefore,
by definition, litter needs to be included in the analysis if more than one animal per litter is used (or the values within a litter can be averaged). However, if multiple litters are used but the treatment(s) are applied to the individual animals, experiments should be designed so that if litter effects exist, then valid inferences can still be made. In other words, litters should not be confounded with other experimental variables because it would be difficult or impossible to detect their influence and remove their effects. Whether litter is an important factor for any particular outcome is then an empirical question, and if it is not important then it need not be included in the analysis. However, the power to detect differences between litters will be low if only a few litters are used in the experiment and therefore a non-significant test for litter effects should not be interpreted as the absence of such effects. Analysing the data with and without litter and choosing the analysis that gives the “right” answer should of course be avoided [35]. Flood et al. provide a nice example in the autism literature of an appropriate design followed by a check for litter effects, and then the results for the experimental effect were reported when litter was both included and excluded [47]. Consistent with other studies demonstrating litter-effects, this paper found a strong effect of litter on brain mass.

Four ways to improve basic and translational research

Better training for biologists

Most experimental biologists are not provided with sufficient training in experimental design and data analysis to be able to plan, conduct, and interpret the results of scientific investigations at the level required to consistently obtain valid results. The solution is straightforward, but requires major changes in the education and training of biologists and it will take many years to implement. Nevertheless, this should be a long-term goal for the biomedical research community.

Make better use of statistical expertise

A second solution is to have statisticians play a greater role in preclinical studies, including peer reviewing grant applications and manuscripts, as well as being part of scientific teams [48]. However, there are not enough statisticians with the appropriate subject matter knowledge to fully meet this demand—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. In addition, this type of “project support” is often viewed by academic statisticians as a secondary activity. Despite this, there is still scope for improving the quality of studies by making better use of statistical expertise.
More detailed reporting of experimental methods

Detailed reporting of how experiments were conducted, how data were analysed, how outliers were handled, whether all animals that entered the study completed it, and how the sample size was determined are all required to assess whether the results of the study are valid, and a number of guidelines have been proposed which cover these points, including the National Institute of Neurological Disorders and Stroke (NINDS) guidelines [49], the Gold Standard Publication Checklist [50], and the ARRIVE (Animals in Research: Reporting In Vivo Experiments) guidelines [51]. For example, ARRIVE items 6 (Study design), 10 (Sample size), 11 (Allocating animals to experimental groups), and 13 (Statistical methods) should be a mandatory requirement for all publications involving animals and could 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. Something similar has recently been introduced by Nature Neuroscience [52]. This would make it easier to spot any design and analysis issues by reviewers, editors, and other readers. In addition, 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 submitting to a journal with 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. A final advantage is that it will make quantitative reviews/meta-analyses easier because much of the key information will be on a single page.

Make raw data available

Another 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 [46], and this was only because one animal per litter was used and the necessary values could be extracted from the figures. 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 [53-55]. Publishing raw data can be taken as a signal that researchers stand behind their data, analysis, and conclusions. Funding bodies should encourage this by requiring that data arising from the grant are made publicly available (with penalties for non-adherence).

The above suggestions would help ensure that appropriate design and analyses are 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.

Conclusions

While it is difficult to quantify the extent to which poor statistical practices hinder basic and translational research, it is clear that a large inflation of false positive and false negative rates will only slow progress. In addition, coupled with researcher degrees of freedom and publication bias, 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.

Additional files

Additional file 1: R code for the analyses and power calculations. Code for the analyses and power calculations are given as a plain text file.

Additional file 2: Raw data. Raw data from Mehta et al. [41], including body weight, locomotor activity and anxiety measures from the open field test, grooming behaviour, and number of marbles buried in the marble-burying test. Details can be found in the original publication.

Additional file 3: List of VPA studies. List of the thirty-four studies using the VPA rodent model of autism.

Additional file 4: Power analysis for the mixed-effects model and the incorrect analysis. The interpretation of the graphs is the same as Figure 4 (main text). Panels A and B are for the mixed-effects model and are nearly identical to the results for averaging the values within each litter and then using a t-test (Figure 4 main text). Panels C and D ignore litter and compare all of the data with a t-test, which results in an artificially inflated sample size and inappropriate high power.

Abbreviations

ANOVA: Analysis Of Variance; BP: Blood Pressure; MPEP: 2-Methyl-6-Phenylethyl-Pyrididine; SOD1: Superoxide Dismutase; VPA: Valproic Acid.
38. Ledford H. Drug candidates derailed in case of mistaken identity. Nature 2012, 483(7391):519. [http://dx.doi.org/10.1038/483519a].
39. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature 2012, 483(7391):531–533. [http://dx.doi.org/10.1038/483531a].
40. Rinaldi T, Silberberg G, Markram H. Hyperconnectivity of local neocortical microcircuitry induced by prenatal exposure to valproic acid. Cereb Cortex 2008, 18(4):763–770. [http://dx.doi.org/10.1093/cercor/bhm117].
41. Mehta MV, Gandal MJ, Siegel SJ. mGluR5-antagonist mediated reversal of elevated stereotyped, repetitive behaviors in the VPA model of autism. PLoS One 2011, 6(10):e26077. [http://dx.doi.org/10.1371/journal.pone.0026077].
42. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. New York: Springer; 2009.
43. Gelman A, Hill J. Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge, UK: Cambridge University Press; 2007.
44. Stodgell CJ, Ingram JL, O’Rara 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(3):617–624. [http://dx.doi.org/10.1016/j.ntt.2006.08.004].
45. 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(4):399–405. [http://dx.doi.org/10.1016/j.ijdevneu.2009.01.006].
46. Murawski NJ, Brown KL, Stanton ME. Interstimulus interval (ISI) discrimination of the conditioned eyelink response in a rodent model of autism. Behav Brain Res 2009, 196(2):297–303. [http://dx.doi.org/10.1016/j.bbr.2008.09.020].
47. Flood ZC, Engel DLI, Simon CC, Negherbon KR, Murphy LJ, Tamavimok W, Anderson GM, Janusonis S. Brain growth trajectories in mouse strains with central and peripheral serotonin differences: relevance to autism models. Neurosci 2012, 210:286–295. [http://dx.doi.org/10.1016/j.neuroscience.2012.03.010].
48. Peers IS, Ceuppens PR, Harbron C. In search of preclinical robustness. Nat Rev Drug Discov 2012, 11(10):733–734. [http://dx.doi.org/10.1038/nrd3849].
49. Landis SC, Amaral SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Damell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudeau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCullin JM, Moxley RT, Narasimhan K, Noble LJ, Perin S, Porter JD, Steward O, Unfer E, Utz U, Silberberg SD. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 2012, 490(7419):187–191. [http://dx.doi.org/10.1038/nature11556].
50. 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(4):399–405. [http://dx.doi.org/10.1016/j.ijdevneu.2009.01.006].
51. 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(6):e1000412. [http://dx.doi.org/10.1371/journal.pbio.1000412].
52. Making methods clearer. Nat Neurosci 2013, 16(1). [http://dx.doi.org/10.1038/nn1013-10].
53. Wicherts JM, Borsboom D, Kats J, Molenaar D. The poor availability of psychological research data for reanalysis. Am Psychol 2006, 61(7):726–728. [http://dx.doi.org/10.1037/0003-066X.61.7.726].
54. Bakker M, Wicherts JM. The (mis)reporting of statistical results in psychology journals. Behav Res Methods 2011, 43(3):666–678. [http://dx.doi.org/10.3758/s13428-011-0089-5].
55. 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(11):e26828. [http://dx.doi.org/10.1371/journal.pone.0026828].