Increasing the statistical power of animal experiments with historical control data

RELACS Consortium; Bonapersona, V.; Hoijtink, H.; Sarabdjitsingh, R. A.; Joels, M.

Published in:
Nature neuroscience

DOI:
10.1038/s41593-020-00792-3

Publication date:
2021

ARTICLES
https://doi.org/10.1038/s41593-020-00792-3
nature neuroscience

Increasing the statistical power of animal experiments with historical control data

V. Bonapersona[1]✉, H. Hoijtink[2], RELACS Consortium*, R. A. Sarabdjitsingh[1,13] and M. Joëls[1,3,13]

Low statistical power reduces the reliability of animal research; yet, increasing sample sizes to increase statistical power is problematic for both ethical and practical reasons. We present an alternative solution using Bayesian priors based on historical control data, which capitalizes on the observation that control groups in general are expected to be similar to each other. In a simulation study, we show that including data from control groups of previous studies could halve the minimum sample size required to reach the canonical 80% power or increase power when using the same number of animals. We validated the approach on a dataset based on seven independent rodent studies on the cognitive effects of early-life adversity. We present an open-source tool, RePAIR, that can be widely used to apply this approach and increase statistical power, thereby improving the reliability of animal experiments.

Before embarking on a new animal study, researchers must decide how many animals per group are needed to optimize the chance of detecting a real effect rather than a chance finding. When performing a statistical power calculation, power is commonly set a priori at 80% (prospective power); that is, the expectation is that 80 of 100 studies investigating a real effect will correctly conclude that the effect exists (true positive), while 20 will not (false negative). As power decreases, the rate of false positive results as well as that of false negative results will increase[1]. Prospective study power therefore directly affects the reliability of the subsequent research findings.

However, a landmark paper by Button et al.[2] estimated, based on 48 meta-analyses of neuroscience studies, that the median power, in reality, is around 21%, in agreement with previous reports in psychology[3]. Although Button's report was based mainly on studies in humans, a similar discrepancy between prospective and actual power likely exists in animal studies. If so, this would contribute substantially to the reproducibility crisis[4] in animal research[5–8], as single, underpowered studies have a low likelihood of detecting a real effect[1], although they can still be informative when included in meta-analyses[9,10].

To improve reproducibility, previous reports suggested using systemic heterogenization[7,11], multiple batches[12] or prospective multicenter studies[8,13], alongside changes in research practice and education[8]. These suggestions involve substantial logistical issues and resources; for the foreseeable future, it is likely that the majority of animal experiments will remain single laboratory. In a single-laboratory setting, an obvious solution to enhancing statistical power would be to increase the number of animals per experiment. For example, for a common effect size of Hedge's $g = 0.5$ (Welch's independent samples t-test, $\alpha = 0.05$), ten animals per group would correspond to a statistical power of 18%, 30 animals per group to 48% power and 65 animals per group to 81% power. Clearly, this is not a feasible solution, not only in terms of the space requirements and financial costs but also in light of continuing efforts to reduce the number of animals used in research.

How can one ensure that a study has sufficient power without increasing the number of animals per group to unrealistically high levels? An appealing approach would be to recycle data from past experiments, as implemented both in human and animal research[14,15]. In research practice, new studies often build on earlier ones, performed in one's own lab or elsewhere. Here, we focus on the specific example of studies using the same experimental endpoint. The data from similar previous studies can be incorporated into new experiments by using Bayesian priors, that is, distributions that describe the mean and variance of an experimental outcome from previous studies. This incorporation can occur already when planning an experiment in the power calculation or exclusively when analyzing the collected data (although this would require preregistration). Transforming information from previous studies in a mathematical function is not trivial, and it was suggested to be one of the most difficult aspects of Bayesian analysis[16]. Priors can be developed by incorporating data from multiple sources (for example, one's own and others' experiments or expert knowledge) and through various methodologies (reviewed in ref. [16]). Bayesian priors are used in the clinical literature and have already been applied to decrease sample sizes in new experiments (for example, refs. [17,18]). Yet, they have been adopted in very few animal studies (for example, ref. [19]; reviewed in ref. [14]), which, although remarkable, received limited attention. As a consequence, the powerful message of using historical controls in new experiments has not reached yet the end beneficiary: researchers performing animal experiments.

In this study, we first evaluate the extent of the power problem in animal research by examining a much larger sample of animal studies than previously reported[2]. Next, we show how historical data can be used to limit the number of animals used in a study by tailoring the Bayesian prior approach to animal experiments. We validate the method and provide an example of how this approach can be applied in daily research practice. We then estimate the impact of the approach on the statistical power of future animal experiments. Lastly, we present RePAIR (Reduction by Prior Animal Informed

[1]Department of Translational Neuroscience, University Medical Center Utrecht Brain Center, Utrecht University, Utrecht, The Netherlands. [2]Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands. [3]University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [13]These authors contributed equally: R. A. Sarabdjitsingh, M. Joëls. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: v.bonapersona-2@umcutrecht.nl

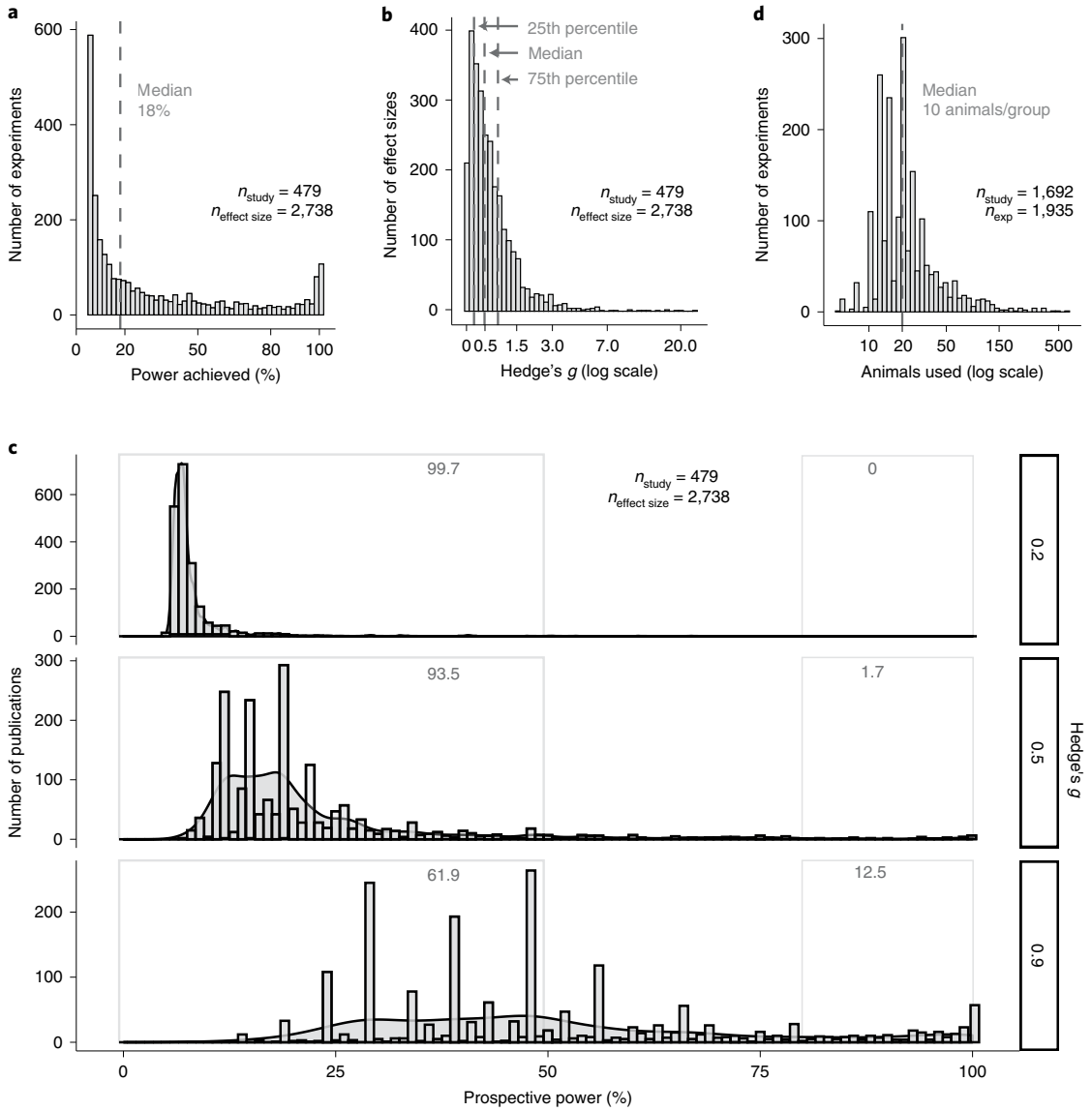

Fig. 1 | Many animal experiments are severely underpowered. a, Power of identified experiments (two-tailed Welch's *t*-test, effect sizes as reported in published papers, 'data B' in Supplementary Fig. 1). Dashed line indicates median, equal to 18%. **b**, Range of common effect sizes in animal literature ('data B' in Supplementary Fig. 1). Dashed lines indicate percentiles. The related quantities (Hedge's *g* values of 0.2, 0.5, 0.9) were defined as 'small', 'medium' and 'large' effect sizes, respectively. **c**, Prospective powers of studies when considering a range of common effect sizes (**b**) and assuming at least one sufficiently powered experiment per publication ('data B' in Supplementary Fig. 1). The highest peaks in the histograms are due to a non-uniform distribution of animals used, as shown in **d**. Histograms and density plots of the same data are overlapping. Left: power ≤50%; right: power ≥80%. **d**, Animals per study when considering the two largest independent groups ('data A' and 'data B' in Supplementary Fig. 1). Dashed line indicates median, equal to 20 animals (~10 animals per group).

Research), a user interface optimized for easy use, to facilitate the implementation of the methodology.

Results

Many animal experiments are severely underpowered. A common approach to estimate the extent of the power problem in animal research is to calculate statistical power from published literature. Through a systematic search (Supplementary Notes 1 and 2), we identified a large sample of animal studies in the areas of 'neuroscience' and 'metabolism' (n_{study} = 1,935) that were previously included in meta-analyses (n_{ma} = 69). These animal studies had an overall median statistical power of 18% (Fig. 1a), which was roughly equal in the two fields (neuroscience, 15%; metabolism, 22%).

Although this approach closely replicated the results of previous reports[2,3], it has major limitations[20]. An alternative approach is to estimate a reasonable prospective power to describe a plausible scenario for new experiments. Because real effect sizes are unknown, we estimated a common range by selecting the medians and quantiles of the distribution identified from published animal studies ($n_{effect\ size}$ = 2,738). These corresponded to Hedge's *g* values of 0.2, 0.5 and 0.9 (Fig. 1b), which is almost identical to Cohen's *d* rule of thumb for small, medium and large effect sizes[21]. Prospective study power was then calculated for this range of effect sizes directly derived from published studies. For large effect sizes, prospective power was sufficient (above 80%) only in 12.5% of studies. This percentage dramatically decreased if smaller effect sizes were considered (Fig. 1c).

Bayesian priors can increase statistical power while limiting sample size. Actual study power is much lower than is commonly assumed (Fig. 1c). The most obvious solution would be to increase sample sizes. Currently, a common sample size used is ten animals per group (Fig. 1d). When considering this common sample size and a Welch's independent samples *t*-test ($\alpha = 0.05$), one would need to assume an effect size of Hedge's $g = 1.4$ to reach a power of 80%. Such an expected effect size is far larger than what is commonly observed in rodent literature (Fig. 1b). If more realistic effect sizes are used, for example, Hedge's $g = 0.2$ or 0.9, the required sample size increases to 394 and 21 animals per group, respectively.

An alternative solution is to use data from past experiments in the form of Bayesian priors. We implement this here as a specific application of power priors[22], while adapting an equal-but-discounted[16] approach. Importantly, we applied priors only to the control group and not to the experimental group, as control animals can be more reasonably assumed to belong to the same population (Methods).

We first performed a simulation study to estimate how the use of Bayesian priors influences sample size and power (Fig. 2a). The simulation study was based on the formula

$$n_{con} = n_{exp} - n_{prior} \times index$$

where the number of animals in the control group (n_{con}) can be reduced by the number of control animals from prior studies (n_{prior}) multiplied by a weight (*index*, value between 0 and 1) that describes the similarity between control and prior groups. The experimental group (n_{exp}) remains the same. Based on this formula, the number of animals needed in the control group is effectively diminished (discounted) by the weighted prior. Conversely, if the number of animals remains the same, a further increase in n_{prior} can still be beneficial, as power could be enhanced up to its highest boundary, that is, approaching 100% with large effect sizes (Fig. 2a).

Validation in a case study. To test the validity of the proposed method in a real-life scenario, we performed a case study involving experiments assessing the effect of early-life adversity (ELA[23]) on spatial learning in adult male mice. The experimental dataset was gathered by aggregating data from single experiments that, in principle, shared the same design but individually had low power, from several laboratories in the RELACS (Rodent Early Life Adversity Consortium on Stress) consortium. Overall, information from 275 animals ($n_{con} = 132$, $n_{ELA} = 143$) was collected, which was more than required by our prospective power calculation ($n_{con+ELA} = 200$). Spatial learning was operationalized as a discrimination ratio measured in the object-in-location test. In the RELACS dataset, the discrimination ratio was significantly lower in animals that experienced ELA than in control mice ($t_{272.99} = 3$, $P = 0.003$).

We then mimicked a prospective experiment by reducing the number of control animals from the RELACS dataset to one-third of the animals that experienced ELA. The new sample sizes would then be $n_{con} = 49$ and $n_{ELA} = 143$. This hypothetical experiment is underpowered, because the difference distribution (ELA distribution – control distribution) contains the value 0 in its 95% confidence interval (Fig. 2b). Normally, one would argue that the two groups are not different from each other. To 'rescue' the interpretation while still conducting a per se underpowered experiment with 49 control animals and 143 animals that experienced ELA, a Bayesian prior was used. A prior was specified based on relevant yet unrelated (non-ELA) published studies of spatial learning using the object-in-location test. This prior had a cumulative adjusted sample size of $n_{prior} = 50.9$, as measured by the equation described in the previous section. The analysis therefore contained the sample size of ~51 animals for the prior of the control group, 49 control animals and 143 animals that experienced ELA. Although the experiment now hypothesized is still underpowered, the prior rescues the interpretation: the value 0 is outside of the 95% confidence interval of the difference distribution (Fig. 2b), and one would conclude that there is evidence that the two groups are different from each other. In other words, this example shows that the same experiment could be conducted with 83 fewer animals (from the 132 control animals from the RELACS dataset to the subgroup of 49 animals for the hypothetical experiment) while maintaining a power >80%.

When specifying the prior, every effort was made to reduce subjectivity in selecting literature and defining the related indices. Yet, other experimenters might have selected different papers with the same task or assigned different weights. Although it is not possible to exclude this possibility, it is unlikely that it would have had major effects on the results. The distribution of the prior was very similar to the one from the control animals in the RELACS dataset (Fig. 3a), which suggests a certain consistency in the measurement values of the experimental endpoint across sources of data.

Nonetheless, the issue of subjectivity may arise when considering other experimental endpoints. We evaluated this concern by performing a sensitivity simulation study to mimic variation arising from different selections of literature (Fig. 3b,c). Here, we randomly sampled control experiments from an available pool, containing non-ELA literature studies as well as the control studies from the RELACS dataset. This analytical approach to estimate variation has limitations, as researchers would rightfully follow pre-specified criteria to select previous experiments, rather than picking them at random. With the estimated variation, we calculated how random control study selection would relate to study power (Fig. 3d). Overall, the prospective power when using a prior was always larger than the currently estimated 18% (Fig. 1a), despite the variations.

Bayesian priors can substantially improve statistical power. Whether Bayesian priors can be applied to new studies depends on the presence of suitable available data from previously performed, similar studies. Although it is difficult to estimate how much suitable data (for a particular experiment) exists in the literature, one could argue that if publications are similar enough to be included in a meta-analysis, they should also be sufficiently similar to be used to calculate a prior.

We recalculated the prospective power displayed in Fig. 1c for studies identified by our systematic literature search (Supplementary Fig. 1). This time, controls from other studies within the same meta-analysis were used to calculate the prior. New experiments were simulated with the same total number of animals (n_{total}) as the published studies but different distributions to the experimental and control groups. As the control group can be aided by the prior, more animals were allocated to the experimental group, according to the rule of thumb $n_{exp} = 2 \times n_{con}$ (Fig. 2c).

For Hedge's $g = 0.9$, application of Bayesian priors increased the percentage of sufficiently powered studies from 12.5% to 69%. These calculations were performed with an *index* of 0.3, which is quite conservative; using an *index* of 1 would yield similar results, with prospective power increasing to 72.5% for large effect sizes.

RePAIR can facilitate implementation. To facilitate the use of Bayesian priors in animal experiments, we created RePAIR, an open-source web-based tool (https://osf.io/wvs7m/) that enables anyone designing future experiments to improve the quality of the study design. With a user-friendly interface, one can calculate (multiple) prior parameters from summary statistics of existing data, perform sample size calculations and execute analyses.

RePAIR can also be used to visualize the (potential) heterogeneity between one's own previously acquired control data and control data from other labs; if one's own data differ substantially from those obtained earlier in other laboratories, one could decide to use only one's own existing control data to calculate the prior or to not use historical controls at all and instead perform a fully powered experiment.

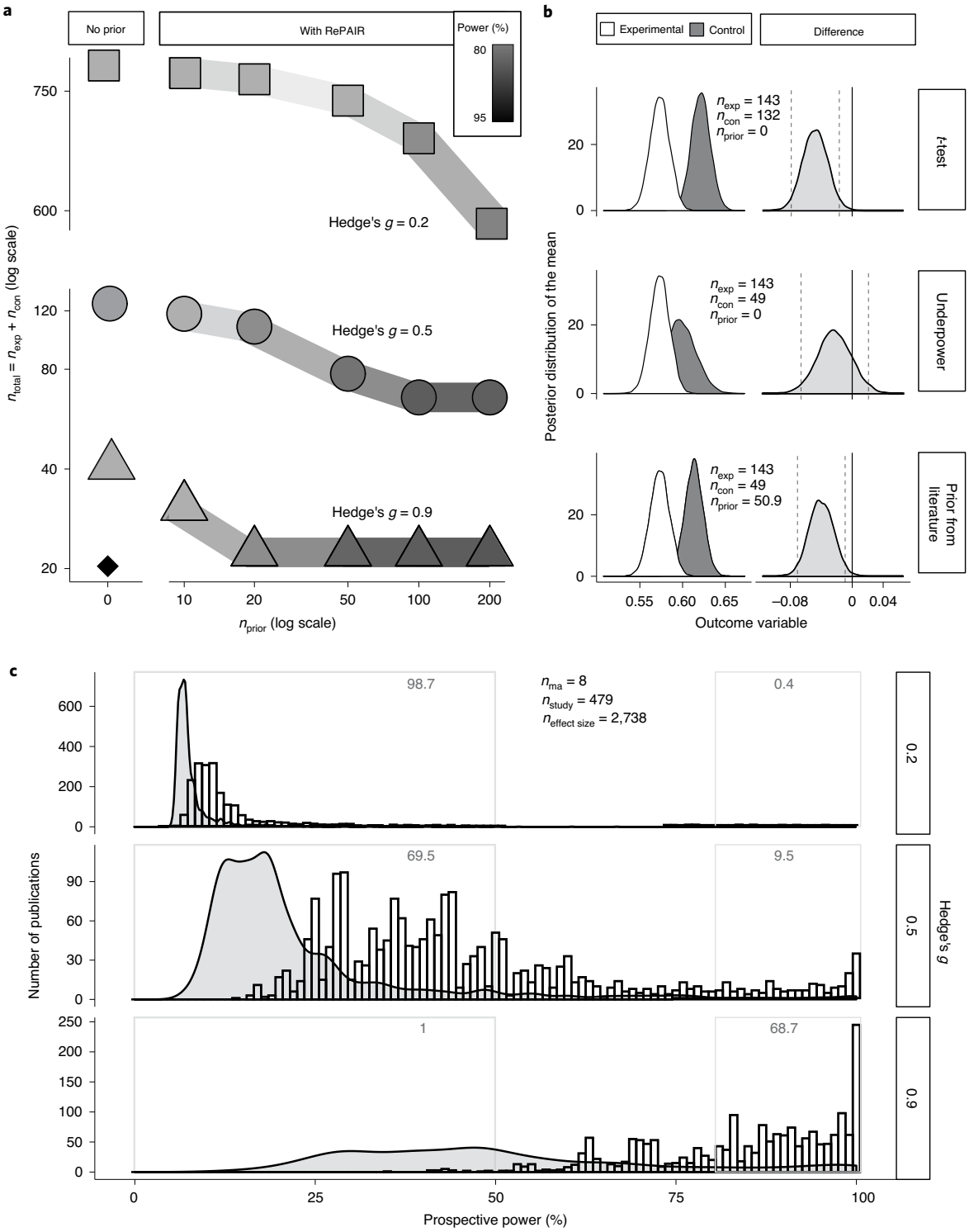

Fig. 2 | Historical controls can decrease the number of animals required for sufficiently powered research. a, Simulation study on the relationship between prior (*index* = 1), sample size and power. An n_{prior} value equal to 0 corresponds to a standard sample size estimation (two-tailed Welch's *t*-test, $\alpha = 0.05$, effect sizes as in Fig. 1b,c, power = 80%). The black diamond indicates the current median sample size. An increase in color intensity signifies an increase in power. As n_{prior} increases, n_{total} decreases until a plateau is reached. Subsequent increases in n_{prior} will result in increased prospective power. **b**, Application of historical controls to the experimental dataset RELACS. Posterior distributions of each group and of their mean differences. The test is significant if 0 (continuous line) is outside of the 95% confidence interval (dashed lines) of the means' difference distribution. From the top (Supplementary Table 1), analysis without prior provides the same result as Welch's *t*-test (top); if n_{con} is decreased, the study becomes underpowered (middle); but this can be rescued if a prior from (unrelated) published literature is introduced (bottom). **c**, Prospective power when using historical controls with *index* = 0.3 (weighted at 30% in the analysis, that is, $n_{prior} = 0.3 \times n_{con}$ of other studies within the same meta-analysis ('data A' from Supplementary Fig. 1) but maintaining current resources (n_{total} kept the same; $n_{con} = n_{total} \times 3^{-1}$ as recommended rule of thumb) shown as a histogram. Gray density plots represent the current prospective power as in Fig. 1c.

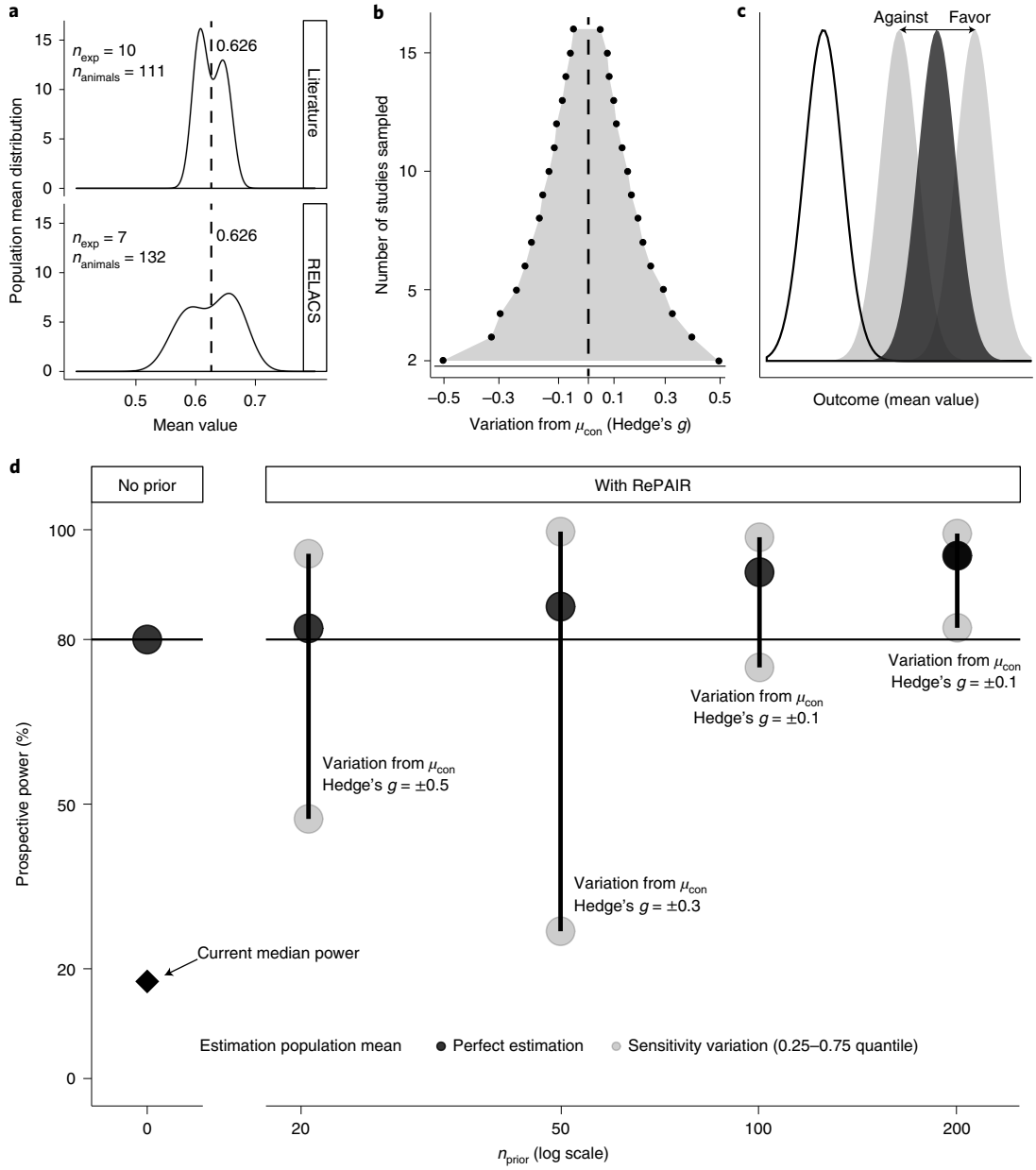

Fig. 3 | Sensitivity simulation. a, Density distribution of control population means with data from prior literature. The dashed line indicates the mean of the control means. n_{exp}, the number of experiments. **b**, Range of variation of estimation of populations means (μ_{con}). Relevant deviations were calculated as the 2.5th and 97.5th percentile interval (gray area) of estimated sampled means (10,000 times, 2–16 experiments combined from literature and the RELACS dataset). Once more than ten experiments were used, the variation (Hedge's $g = 0.1$) became negligible. **c**, Schematic representation of how the variation of estimated population means can be both in favor of and against the hypothesis being scientifically investigated. White, distribution of the experimental group; gray, distribution of the control group. **d**, Prospective power (Hedge's $g = 0.5$, equal variances) with historical controls is higher than that of current practices (black diamond), despite variations in population mean estimation (black line). Because n_{total} is not consistent, due to the increasing n_{prior} values, prospective power can only be interpreted vertically (black lines). Each vertical line displays how prospective power changes if the estimated prior mean is a perfect estimation of the population (dark dot) or if the mean deviates from it in favor of (top, light dot) or against (bottom, light dot) the investigated hypothesis. The light dots correspond to rounded values of the 2.5th and 97.5th interval calculated from **b**. The exact variation for each percentile interval is noted in the figure. The progression of the dark dots is an alternative visualization of the increase in color intensity in Fig. 2a.

Sensitivity analyses are essential[16] when priors are specified. To facilitate such analyses, we included the option to perform two types of sensitivity analyses in RePAIR: (1) the leave-one-out sensitivity analysis, to check whether any prior study has substantial influence on the final result and (2) a sensitivity analysis on the indices by selecting lower or higher indices than those chosen for the analysis. Using the 'leave-one-out' sensitivity, one can assess the impact of each specific experiment on the final analysis. Here, prior parameters are calculated k times for each $k-1$ experiment added; if three prior experiments (A, B, C) were added, three sensitivity analyses will be conducted (A and B, B and C, C and A). To perform the indices' sensitivity analysis, users have to specify the *index* as a range. The average of the range is used for the main analysis, whereas the lower and higher boundaries of the range are used for the sensitivity analyses. In RePAIR, parameters for sensitivity analyses are automatically calculated when specifying the prior.

The resulting file can then be re-uploaded when analyzing data from the new experiment, and sensitivity analyses will be automatically conducted.

Discussion

There is a growing awareness of the reproducibility issue in animal experiments. Study preregistration and the introduction of more rigorous guidelines (for example, PREPARE for planning of animal experiments and ARRIVE[24] for their reporting) can only partially address this issue. We here describe the (lack of) statistical power in animal studies and explain how the use of Bayesian priors can provide a potential solution. As previously suggested by others (for example, refs. [14,17–19]), this statistical method uses historical data to limit the number of animals necessary to perform well-powered research or to reach higher statistical power with the same number of animals as currently used in experiments. We delineated how to best apply Bayesian priors in the context of animal research and created RePAIR, a user interface to ease the implementation of this approach. This approach can substantially increase prospective power without increasing the total number of animals used. It can be an extremely powerful tool, if correctly implemented and interpreted.

Animal experiments have low statistical power. The statistical power of animal experiments is much lower than commonly assumed a priori. Although our approach was not conservative, we estimated that, at best, 12.5% of a large sample of rodent studies were sufficiently powered (that is, prospective power was larger than 80%). This estimate is a best-case scenario, as it is not yet adjusted for any subsequent multiple testing, experimental bias, *P* hacking and/or fishing, selective reporting, etc.

One may wonder why our estimate of sufficiently powered experiments is so low. A technical limitation of our approach is that it considers a range of effect sizes found in literature and not a minimum effect size of 'biological significance'. Although valuable, the minimum effect size criterion is seldom used in power calculations. We therefore consider our estimate reliable. Besides this technical limitation, several observations can explain why prospective study power is much lower than the commonly assumed 80%. One explanation is that effect sizes are often estimated optimistically in power calculations, as they are based on earlier findings that are liable to (publication) bias[25]. A second explanation is that rodent experiments are frequently exploratory in nature[26], and many scientists opt to use a debatable 'standard' of six to ten animals per group. Indeed, the effect size frequently assumed in rodent literature (Hedge's $g=1.4$) is much larger than the range of effect sizes that is commonly observed (Hedge's $g=(0.2, 0.9)$). Effect sizes in certain subfields may be more toward the lower (for example, behavioral phenotyping[9]) or higher (for example, molecular studies[20]) end of this distribution. Still, this discrepancy between assumed and observed effect sizes contributes to the power problem and reproducibility crisis in animal research in a major way.

Limitations and recommendations for the reuse of historical data. The use of historical control data as here proposed requires the researcher to select experiments and to specify weights via the *index*. This selection is naturally subjective and thus can be criticized as introducing bias into an experiment[27]. In the next paragraphs, we discuss how subjectivity might impact an experiment using historical controls, and how these limitations are pragmatically addressed in our methodology. Next, we discuss why using historical controls is a valid approach, despite its subjectivity. Finally, we provide practical recommendations for the reuse of historical control data in new experiments.

When selecting previous experiments, a possible risk is that their cumulative distribution is very different from that of the new experiment's control group (prior-data conflict)[18]. The prior distribution may then push the control group more toward the experimental group (causing a decrease in power) or further away from it (causing an increase in power); in other words, it can introduce a bias in the posterior distribution, that is, the distribution obtained from combining prior and new (control) data. The posterior distribution of the control group may then not be a good estimate of the control population, thereby directly impacting (negatively or positively) the power of the study. Previous reports suggested several ways to mitigate this problem. Some suggested disregarding the prior altogether, although this would cause a reduction in study power. Others suggested redistributing the weights of the prior studies based on their relative discrepancies[18,28]. However, we argue that prior-data conflict cannot be adequately addressed in this way. Thus, these solutions are based on the assumption of a correct evaluation of prior-data conflict. This means that a new experiment was planned with a prior control group and that the data of the new experiment was already collected. The evaluation of prior-data conflict then consists of judging whether the prior control and the new control actually belong to the same population. As the approach presented in this paper is aimed at reducing the number of animals in the new control group as much as possible, the new control group will not be sufficiently large to correctly estimate the new control population and therefore cannot be compared to the prior control population. In other words, we cannot disregard a wealth of previous information based on data from a handful of new animals.

Although we cannot adequately check whether the prior control group is reasonable (that is, there is no prior-data conflict) after we conduct the new experiment, we can evaluate whether prior control groups are potentially incompatible while we plan the new experiment. Prior controls can be from one's own lab, from other labs or a combination of the two. Using information from multiple laboratories can be beneficial. If each laboratory is a single population[7], the overall population can be addressed as a population of populations. As a consequence, results based on information from multiple laboratories should be more generalizable. However, using information from multiple laboratories can also be a major source of variation in the prior distribution, because variation within a laboratory is likely smaller than variations between laboratories[5,13]. An experimenter can check whether one's own prior control data differ largely from prior control data selected from literature or whether a particular experiment stands out. This evaluation must occur on a case-to-case basis with careful assessment and justification, ideally while planning the experiment. When building a prior, the experimenter can visually compare the distributions of datasets from the selected previous experiments of their own or others and assess (for example) whether their own prior control data is too different from that of others or whether there is an 'odd-one-out' dataset that drives the prior control distribution. The experimenter can then choose to exclude the odd-one-out dataset or to not use prior control data from other laboratories at all if they are too different from their own prior control data. In both circumstances, the experimenter may nonetheless wish to review the potential origin of the differences, for example, by comparing experimental design between studies. To facilitate the process of assessing the compatibility of prior control data, the RePAIR app provides a visualization tool; this will aid the experimenter in the process of selecting prior experiments and determining their *index*.

Besides selecting previous experiments subjectively, in our methodology, the experimenter also specifies their weight (*index*) subjectively. To avoid subjectivity, one may wonder whether it is necessary to use weights or whether they could be derived from a calculation. The use of weights is in agreement with the common view that past information needs to be somewhat downweighed because experiments are rarely identical[17]. Several methods (for example, refs. [17,18,28,29]) were developed to overcome subjectivity in defining

the weights by analytically deriving them based on the discrepancy between historical and new data. These methods are appealing and definitely pragmatic for clinical sciences. However, we argue that these methods are not appropriate for animal studies. The argument is similar to the one used in the previous paragraphs to deprecate the assessment prior-data conflict in animal experiments: if prior controls are used to reduce sample size as much as possible in the new control group, it cannot be assumed that the new control group (likely based on a small number of animals) will provide a good estimation of the new control population. A correct estimation of the new control population is necessary to evaluate the discrepancy between prior and new control groups. As a consequence, methodologies that analytically derive weights based on this discrepancy cannot be used in the context of animal experiments, where the goal is to reduce sample size as much as possible. Therefore, weights are necessary and need to be specified by the experimenter. In our methodology, we use the 'equal-but-discounted' method based on power priors, as suggested by Ibrahim and Chen[22]. Briefly, by setting a certain discount or weight (for example, *index* = 0.5), the sample size is reduced (for example, from ten to five). Scientists themselves (by expert elicitation, an accepted practice in Bayesian statistics[30]) can therefore decide to what extent they value earlier data. Although subjective, even conservative (low) *indices* can be beneficial.

One could argue that the subjective selection of previous experiments and related *indices* is susceptible to gaming and offers yet another 'degree of freedom' when performing analyses. This concern is valid, especially for research fields for which little 'past evidence' exists. Until optimal population parameters are known, specification of a prior is subject to variation. At the same time, it is impossible to pre-define how many high-quality studies are necessary for estimating an optimal parameter. We recommend preregistering prior experiments and their *indices* on suitable platforms such the Open Science Framework (https://osf.io/), https://preclinicaltrials.eu/[31] or the Experimental Design Assistant[32]. During preregistration, scientists should define the prior experiments and related *indices* and should also describe the rationale behind the choice of experiments and planned sensitivity analyses. Furthermore, scientific societies can facilitate the process of defining reliable priors, for example, by establishing expert panels. This could eventually result in an 'atlas' of common control priors in animal research. As the number and quality of experiments increases, more precise estimates of the parameters of the control population can be obtained, and, consequently, the subjectivity in selecting experiments and *indices* will decrease.

Despite the above-mentioned limitations, the use of historical controls is desirable and valid. It is desirable because it offers the possibility of increasing statistical power, thereby improving the reliability of animal research. It is valid because it is a translation in statistical terms of assumptions already used in daily research practice. New experiments are usually planned based on information obtained in previous studies. Even though variations between strains and labs clearly exist[7,12,33,34], researchers have similar expectations about how a control group 'should respond'. Indeed, if this expectation is not met, a researcher would likely not trust the data and conclude that the experiment 'did not work' or 'needs to be better optimized'. In this context, an advantage of rodent studies is that they are relatively well controlled and often employ 'standard' tests used in many labs. For example, if the plasma concentration of a hormone normally varies from 60 to 100 $\mu g\,ml^{-1}$ in control animals, an experimenter would rightfully question the validity of data from control animals that show a range between 5 and 10 $\mu g\,ml^{-1}$. Translating the above into statistical terms, researchers assume that control animals always belong to the same overall population. This warrants the formal statistical use of priors to supplement control group data.

The choices involved in building the prior distribution must be considered when interpreting the results, for which sensitivity analysis remains essential[16]. Choosing prior studies and the related *indices* is similar to selecting literature for a new experiment. We recommend considering the quality of the study as well as design variations that likely impacted the results. For example, researchers may select previous experimental data obtained from only a specific sex (for example, females) if the outcome is sex specific (for example, ovulation) or from both sexes if it is not expected to be sex specific[35]. Similarly, blinding and randomization may be chosen as inclusion and exclusion criteria or might be used to define the *index*. The *index* is specified for each study separately. As a rule of thumb, previous reports attributed a large weight value (0.9) to studies that belonged to the same meta-analysis and lower weight values (0.7–0.8) to studies that did not[18]. We suggest a more conservative stand: large weight values (0.8–1) could be applied to repeated experiments from the same lab (for example, different batches), medium weight values (0.4–0.8) to experiments that (could) belong to the same meta-analysis and small weight values (0.1–0.4) to experiments from other sources. We also recommend specifying a range for the *index* and conducting sensitivity analyses. RePAIR has in-built features to support each step of the process, from visualization of distribution of prior experiments to automatic sensitivity analyses.

If sufficient prior information is available, it is theoretically possible to decrease n_{con} to as low as two (to still be able to calculate a standard deviation). However, this is not advisable, because randomization would be difficult. As a rule of thumb, we recommend that control animals comprise at least one-third of the total number of animals in a new experiment. Even though sample size can be no longer reduced, prior information can still be beneficial because it will increase statistical power above 80%.

Finally, if sufficient prior information is not available, priors should not be used; in this case, the researcher should perform an appropriately powered experiment, even if this means that a sample size of (well) over 20 animals per group is required.

Concluding remarks. The reuse of historical control data in animal experiments can be an extremely powerful tool to increase statistical power and the reliability of animal studies, if correctly implemented and interpreted. Although here discussed in relation to *t*-tests, the same approach can be used in more complex experimental designs (for example, 2×2 ANOVAs), in which multiple groups could then be considered as 'controls'. It is a feasible solution to reduce and replace animal use for those research questions for which good alternatives to animal testing are not yet available.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-020-00792-3.

Received: 24 February 2020; Accepted: 23 December 2020;
Published online: 18 February 2021

References

1. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
2. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
3. Smaldino, P. E. & McElreath, R. The natural selection of bad science. *R. Soc. Open Sci.* **3**, 160384 (2016).
4. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).

5. Crabbe, J. C., Wahlsten, D. & Dudek, B. C. Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**, 1670–1672 (1999).
6. 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.* **10**, 712–713 (2011).
7. Voelkl, B. et al. Reproducibility of animal research in light of biological variation. *Nat. Rev. Neurosci.* **21**, 384–393 (2020).
8. Macleod, M. & Mohan, S. Reproducibility and rigor in animal-based research. *ILAR J.* **60**, 17–23 (2019).
9. Bonapersona, V. et al. The behavioral phenotype of early life adversity: a 3-level meta-analysis of rodent studies. *Neurosci. Biobehav. Rev.* **102**, 299–307 (2019).
10. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).
11. Richter, S. H., Garner, J. P. & Würbel, H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nat. Methods* **6**, 257–261 (2009).
12. Karp, N. A. Reproducible preclinical research—is embracing variability the answer? *PLoS Biol.* **16**, e2005413 (2018).
13. Richter, S. H. et al. Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. *PLoS ONE* **6**, e16461 (2011).
14. Kramer, M. & Font, E. Reducing sample size in experiments with animals: historical controls and related strategies. *Biol. Rev. Philos. Soc.* **92**, 431–445 (2017).
15. Brakenhoff, T., Roes, K. & Nikolakopoulos, S. Bayesian sample size re-estimation using power priors. *Stat. Methods Med. Res.* **28**, 1664–1675 (2018).
16. Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation* **1** (Wiley, 2004).
17. Galwey, N. W. Supplementation of a clinical trial by historical control data: is the prospect of dynamic borrowing an illusion? *Stat. Med.* **36**, 899–916 (2017).
18. Mutsvari, T., Tytgat, D. & Walley, R. Addressing potential prior-data conflict when using informative priors in proof-of-concept studies. *Pharm. Stat.* **15**, 28–36 (2016).
19. Walley, R. et al. Using Bayesian analysis in repeated preclinical in vivo studies for a more effective use of animals. *Pharm. Stat.* **15**, 277–285 (2016).
20. Nord, C. L., Valton, V., Wood, J. & Roiser, J. P. Power-up: a reanalysis of 'power failure' in neuroscience using mixture modeling. *J. Neurosci.* **37**, 8051–8061 (2017).
21. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, 1977).
22. Ibrahim, J. G. & Chen, M. H. Power prior distributions for regression models. *Stat. Sci.* **15**, 46–60 (2000).
23. Rice, C. J., Sandman, C. A., Lenjavi, M. R. & Baram, T. Z. A novel mouse model for acute and long-lasting consequences of early life stress. *Endocrinology* **149**, 4892–4900 (2008).
24. Percie du Sert, N. et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. *PLoS Biol.* **18**, e3000410 (2020).
25. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
26. Rubin, E. J. & Fortune, S. M. Misunderstanding the goals of animal research. *BMJ* **360**, 29321149 (2018).
27. Gelman, A. Objections to Bayesian statistics. *Bayesian Anal.* **3**, 445–450 (2008).
28. Viele, K. et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm. Stat.* **13**, 41–54 (2014).
29. Neuenschwander, B., Capkun-Niggli, G., Branson, M. & Spiegelhalter, D. J. Summarizing historical information on controls in clinical trials. *Clin. Trials* **7**, 5–18 (2010).
30. O'Hagan, A. Expert knowledge elicitation: subjective but scientific. *Am. Stat.* **73**, 69–81 (2019).
31. van der Naald, M., Wenker, S., Doevendans, P. A., Wever, K. E. & Chamuleau, S. A. J. Publication rate in preclinical research: a plea for preregistration. *BMJ Open Sci.* **4**, e100051 (2020).
32. Du Sert, N. P. et al. The experimental design assistant. *Nat. Methods* **14**, 1024–1025 (2017).
33. Crabbe, J. C. & Phillips, T. J. Mother nature meets mother nurture. *Nat. Neurosci.* **6**, 440–442 (2003).
34. Kafkafi, N. et al. Addressing reproducibility in single-laboratory phenotyping experiments. *Nat. Methods* **14**, 462–464 (2017).
35. Shansky, R. M. Are hormones a 'female problem' for animal research? *Science* **364**, 825–826 (2019).

RELACS Consortium

M. Abbinck[4], T. Z. Baram[5,6], J. L. Bolton[5], V. Bonapersona[1], J. Bordes[7], M. Joëls[1,3,13], J. Knop[1], A. Korosi[4], H. J. Krugers[4], J. T. Li[8,9], E. F. G. Naninck[4], K. Reemst[4], S. R. Ruigrok[4], R. A. Sarabdjitsingh[1,13], M. V. Schmidt[7], E. H. L. Umeoka[4,10], C. D. Walker[11], X. D. Wang[12] and K. Y. Yam[4]

[4]Swammerdam Institute for Life Sciences, SILS-CNS, University of Amsterdam, Amsterdam, The Netherlands. [5]Department of Anatomy/Neurobiology, University of California Irvine, Irvine, CA, USA. [6]Department of Pediatrics, University of California Irvine, Irvine, CA, USA. [7]Department of Stress Neurobiology and Neurogenetics, Max Planck Institute of Psychiatry, Munich, Germany. [8]National Clinical Research Center for Mental Disorders, Peking University, Beijing, China. [9]Key Laboratory of Mental Health, Peking University, Beijing, China. [10]Faculty of Medicine, University Center Unicerrado, Goiatuba, Brazil. [11]Department of Psychiatry, McGill University, Montreal, Quebec, Canada. [12]Department of Neurobiology, Zhejiang University School of Medicine, Hangzhou, China.

Methods

General information. Every effort was made to minimize bias; for example, data gathering and analysis were performed blindly, multiple experts were consulted for sensitive information (inclusion and exclusion criteria), and studies' characteristics were prospectively defined. This study was developed after a preliminary analysis of study power and estimation of sample sizes, conducted on a meta-analytic dataset developed previously by our own lab[36]. Part of this data was also used in this publication. Although no ex-ante protocol was preregistered, each component of this study was thoroughly planned in advance, unless otherwise stated in each individual section. For data, code and other information about the project, see https://osf.io/wvs7m/. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Statistics. To compare control and experimental groups, we used two-tailed Welch's independent samples *t*-test ($\alpha = 0.05$). We chose Welch's test instead of Student's *t*-test, because Welch's test does not assume equal variances between groups. Data distribution was assumed to be normal, but this was not formally tested. Given the small sample sizes of animal experiments, it is also likely that normality tests were underpowered. Bayesian analyses are explained in detail in the following sections.

Evaluation of studies for the systematic review was performed in a random order. Briefly, each study was given a pseudo-random number generated in R. This number was then used for the ordering and assessment of publications. For the case study, presence of randomization was an inclusion criteria. However, we do not have information on how randomization was conducted by the single independent laboratories.

Throughout the study, every effort was made to limit selection and confirmation biases. Inclusion and exclusion criteria for the systematic review were defined before starting the review. The choices of distributions and ranges throughout the analysis (for example, estimation of effect sizes, sensitivity variation range) were performed once the data were already collected but before any data visualization. For the definition of prior information and the definition of inclusion and exclusion criteria for the case study, the researcher (V.B.) did not have access to the effect sizes but did have access to meta information from the study (for example, characteristics of the ELA model).

All analyses were conducted with R (version 4.0.0) in the RStudio environment on a macOS Mojave (version 10.14.6). The following R packages were central to this study: (1) tidyverse[37] (version 1.3.0) for general data handling, (2) shiny[38] (version 1.5.0) for the RePAIR web-based tool and (3) MESS[39] (version 0.5.6) for power calculations. The case study power calculation was also confirmed with G*Power[40] (version 3.1.9.2).

Estimation of study power. Because real effect sizes are not known, estimating statistical power of animal research is equivocal. A common approach is to calculate achieved statistical power from meta-analyses identified with a systematic literature search (Supplementary Notes 1 and 2 and Supplementary Table 2).

The achieved power is the probability of rejecting the null hypothesis (that is, no difference between the control and experimental group) with the observed sample sizes. Here, this was retrospectively calculated for each set of summary statistics extracted from the systematic literature search ('data B' from Supplementary Fig. 1). Although data may have come from complex experimental designs, we assumed it always belonged to two independent groups (two-tailed Welch's *t*-test, $\alpha = 0.05$, sample size and Hedge's *g* of 'data B' from Supplementary Fig. 1). This retrospective power calculation is a biased estimation of prospective study power, because the larger the *P* value observed in a study, the smaller its achieved power[41]. We replicated previous reports[2,3] that used meta-analysis to estimate real effect sizes. This retrospective power calculation was not part of the original study protocol and was subsequently added.

An alternative approach is to estimate a common prospective study power, thereby partially overcoming the limitations of achieved power calculations. As an experimental design, we assumed two independent groups (two-tailed Welch's *t*-test, $\alpha = 0.05$), while sample sizes were gathered from our systematic search ('data A' from Supplementary Fig. 1). Importantly, only the two largest groups reported in each paper were extracted, assuming that at least the comparison of these two groups was sufficiently powered, while all other experiments may have been control experiments. For effect sizes, we aimed to estimate a plausible range, rather than a single value, to mimic scenarios of researchers initiating a new study.

To estimate a plausible range of effect sizes in preclinical literature, we calculated the 25th, 50th and 75th percentiles of absolute values of Hedge's *g* and defined them as small, medium and large effect sizes, respectively (based on 'data B' from Supplementary Fig. 1). Blinded to the results, we chose the 25–75% interval instead of the 95% confidence interval, to avoid extreme values. Extremely small effect sizes may not be biologically relevant and are confounded by null effects, while extremely large values may lead to interpretation issues and are confounded by overestimations due to biases. We confirmed (see code at https://osf.io/wvs7m/) that these values were replicable by applying the same methodology to a separate dataset[2,20].

Within this framework, prospective power is the probability of rejecting the null hypothesis if the effect size is equal to a small, medium or large value. A simple experimental design was assumed (*t*-test), while sample sizes and effect sizes were estimated from literature. As a consequence, this approach for calculating prospective power portrays a plausible scenario that a new researcher may expect.

Simulation study on the relationship between prior information, sample size and statistical power. The mathematical derivation of the algorithm for prior distributions[42] is described in detail in the Supplementary Note 3. In our study, priors were built based on conjugate distributions, meaning distributions that, when multiplied by the likelihood function, would create a posterior distribution, which summarizes information from previous and current studies with respect to the mean of the control group. The posterior distribution was from the same family as the prior distribution. We therefore chose the prior distribution for the mean in the control and experimental groups to be normal and for the variance to have an inverse χ^2 distribution. Although modern computing power is reducing the need for conjugacy[16], we preferred this method because of its solid mathematical foundation, and the assumption of normality seemed appropriate, as it is frequently used in preclinical sciences.

Of note, informative priors (namely, priors based on previous experiments) were applied only to the control group. The mean and the variance of the experimental group also have a prior and a posterior distribution. However, the prior distribution of the experimental group is 'uninformative', meaning that it will not have an impact on the results. Therefore, the posterior distributions that describe the mean and variance of the experimental group in our approach depended only on the information from the current experiment.

We performed a simulation study to evaluate the extent to which a prior could reduce the number of animals necessary and how this would influence study power. The more informative a prior was for the mean in the control group, the more influence it will have on the conclusions of the experiment. Mean and variance of data in the control group were kept identical in all conditions ($\mu_{con} = 0$, $\sigma^2_{con} = 1$); therefore, the influence of the prior was dependent only on its varying sample size, n_{prior}. Supplementary Table 3 summarizes all factors varied in the simulation. For each combination of factors, 10,000 datasets were sampled from the corresponding population.

First, we calculated how many animals ($n_{total} = n_{con} + n_{exp}$) one would need to perform experiments with the determined characteristics, given a standard sample size calculation ($n_{prior} = 0$). This was later confirmed by G*Power[40]. The calculation assumed a balanced design, meaning $n_{con} = n_{exp}$. Second, n_{con} was decreased by adding n_{prior} while keeping n_{exp} the same. Because it would be illogical for n_{con} to become negative when $n_{prior} > n_{con}$, n_{con} is minimally 2, which is the lowest possible sample size to compute a standard deviation. The total number of animals used is then

$$n_{total} = n_{exp} + n_{con}$$

$$n_{con} = n_{exp} - n_{prior}$$

$$n_{prior} = \sum_{p=1}^{p} n_p \times index_p$$

where the number of animals in the control group (n_{con}) is diminished by the effective number of prior animals (n_{prior}), meaning the sum of the animals in each experiment used to define the prior (n_p), multiplied by the respective weight ($index_p$). The *index* is a value between 0 and 1. An *index* of 0.3 means that the information in the prior study at hand will only be weighted for 30% in the analysis. In the simulation, we set $index = 1$, and we assumed that the prior is a perfect estimation of the population, although this issue was further addressed with a sensitivity simulation study (Case study). For analyses, researchers may opt to vary this value, depending on the degree of similarity of the prior experiments to the current study. For more information about this topic, see 'expert elicitation' in ref. [30].

Case study. For validation and as an example, we applied Bayesian priors, as described in the previous sections, to an experimental dataset. Here, the prior for the control group was specified from unrelated literature, while the prior for the experimental group was uninformative.

A well-powered dataset investigating a real and reproducible difference between two groups was required. We defined an effect as 'real' and 'reproducible' as one that persists in a high-quality, well-powered meta-analysis. These criteria were met by the effects of ELA on memory after non-stressful learning, as identified by a recent meta-analysis of literature previously conducted by our own lab[9]. From this study, an effect size of Hedge's $g = 0.4$ was estimated to describe the difference in performance on the object-in-location memory task between control animals and animals that experienced ELA with the limited bedding and nesting model[23]. Considering a Welch's two-tailed independent means *t*-test and an α value of 0.05, 200 animals would be required to achieve a power of 80%.

Due to the paucity of power of animal studies, it is not surprising that we were unable to identify any study on this experimental outcome using (at least) 200

animals. Even though no single laboratory uses such sample sizes, the required power could be attained by combining data from multiple laboratories. To this end, we created RELACS, a unique rodent consortium constituted by several laboratories around the globe studying ELA.

We identified relevant authors from a recent systematic search by our lab[9], as well as through our network (Supplementary Note 4). The consortium was prospectively founded and ultimately included seven independent experiments that met the specified criteria for this particular study. We calculated, for each experiment (that is, an independent set of animals), a measure of discrimination (discrimination ratio) as the ratio between the time spent in the novel location divided by the total exploration time, meaning the sum of the time spent in the novel and the familiar location ($discrimination\ ratio = \frac{time_{novel}}{time_{novel}+time_{familiar}}$). When analyzed independently, *a P* value <0.05 was reached in only two of seven experiments, which is in agreement with the low power of preclinical studies. By combining the seven experiments, we reached a sample size of 275 animals, distributed as $n_{con}=132$ and $n_{ELA}=143$. The effect size (Hedge's $g=0.37$) calculated in the RELACS dataset was similar to the one estimated from literature (Hedge's $g=0.4$). We concluded that this dataset meets the required criteria to validate RePAIR: it describes a reproducible effect as shown by the meta-analysis, and it is sufficiently powered, as the sample size was larger than the expected 200.

Of note, aggregating data from multiple laboratories in such a way would normally be unadvisable, as it does not meet the criteria of an individual participant data meta-analysis. However, we used this approach here because our intent was to 'mimic' a well-powered experiment, which was otherwise unavailable in the literature.

To specify a prior from unrelated studies, one of us (V.B.) selected relevant literature to mimic planning an experiment with the same characteristics (Supplementary Table 4) as the RELACS dataset, that is, investigating memory after non-stressful learning with the object-in-location task in adult (aged 9–41 weeks; median, 18 weeks) male mice. The researcher was requested to select eight publications that she would use to set up her study, while focusing on the control and not the experimental group. The selected publications did not belong to the ELA field and were not used elsewhere in this manuscript. Furthermore, for each study, the researcher defined a similarity *index*, a number between 0 and 1 that would express how similar the control group from each literature study was to that in the experiment that she was planning to perform (1, identical or equal). Two publications reported the same outcome on two separate groups of animals. Both experiments were considered, albeit with a lower *index*. The process was overseen by a senior researcher (R.A.S.).

As the experimental dataset and prior specification were identified as described above, we had all the elements to validate that the Bayesian approach would reach, with fewer animals, the same conclusions as current practices. First, we performed a Welch's independent samples *t*-test (two-tailed, $\alpha=0.05$) on the RELACS dataset to replicate the result that control and ELA groups differed in discrimination (*P* value <0.05) in the object-in-location task. We then performed the same test but with fewer animals in the control group and an informative prior. Several tests (Supplementary Table 1) were conducted as controls.

Although V.B. selected the prior while blinded to the results of the RELACS dataset, prior specification had some degree of subjectivity; that is, another researcher may choose different publications on which to base their study. To experimentally quantify relevant variation in article selection, we simulated many different priors by picking at random $10{,}000\times k$ experiments ($k=2$–16) from the 17 experiments identified in total (ten from V.B.'s literature selection and seven from the RELACS dataset). Variation in article selection for each *k* was calculated as the 2.5th and 97.5th percentiles to avoid extreme values. Changes in Hedge's *g* between 0.1 and 0.5 could appropriately describe the variation across *k*, and ten articles were here sufficient for a stable estimation of the population parameters. Of note, the sampling occurred from a finite population, where 17 experiments represent the reference value of the estimated variations. As a consequence, the intervals may be underestimated.

With this experimentally derived estimation of population mean's variation, we conducted a sensitivity simulation study to investigate how the variation of the prior control population mean affected prospective study power. Of note, this variation can act both in favor of or against the hypothesis under experimental investigation, depending on whether the prior control population mean moves toward or away from the population mean of the experimental group. Despite this limitation, we preferred this approach of experimentally deriving variation values over using a canonical variation of Hedge's $g=0.1$.

We preferred to use the number of animals rather than the number of experiments in the sensitivity simulation, to remain consistent with the first power simulation study. The relationship between the number of sampled experiments and the number of animals is not straightforward. For example, one can achieve $n_{prior}=20$ with just one experiment or two (for example, each with $n=10$) or three (for example, $n=9$, $n=6$ and $n=5$). To transform the variations due to experimental selection to the variations linked to sample sizes, we identified—across the $k\times10{,}000$ sampled estimations of means—sample sizes of animals roughly equivalent to 20, 50, 100 and 200 (n_{prior} in our sensitivity simulation). In these subgroups, we calculated the 2.5th and 97.5th percentiles and visually validated their consistency. These values were used in the sensitivity simulation study to vary prior control population means (between 0 and ±0.5 Hedge's *g*, depending on n_{prior}). All factors of the sensitivity simulation were kept identical to those of the previous simulation study (Supplementary Table 3).

Estimating how prior control information can impact statistical power with the current total number of animals used. We estimated the increase in prospective power if the Bayesian prior methodology would be used in new animal experiments with the resources currently available. We considered each study identified within each meta-analysis ('data A' from Supplementary Fig. 1) as a new experiment, for which n_{total} was kept the same, but animals were redistributed in favor of the experimental group ($n_{exp}=2\times n_{con}$, according to our rule of thumb). The controls from all other studies within the same meta-analysis were then considered as priors. In other words, n_{prior} was calculated from the cumulative n_{con} of all other papers included within the same meta-analysis. This cumulative n_{prior} was then multiplied by the similarity *index* = 0.3, meaning that the degree of similarity from the control groups of studies included in the meta-analysis was valued at 30%. In this circumstance, the value of 0.3 is arbitrary. To evaluate how the similarity *index* affects power, we also calculated prospective power with a similarity *index* of 1.

Prospective power was calculated in the case of a two-tailed Welch's independent means *t*-test, for the plausible range of effect sizes previously identified (Estimation of study power), when considering an α value of 0.05. Because we adopted the same methodology and the same data, the immediate potential impact can be assessed by comparing the prospective power of previously performed experiments without prior information to that of experiments with prior information.

Lastly, we created the web user interface RePAIR to facilitate the implementation of Bayesian prior methodology to improve statistical power in animal experimentation. The supporting code is also freely available (https://osf.io/wvs7m/).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of the current study can be downloaded from https://osf.io/wvs7m/.

Code availability

All code used in this manuscript is available at https://osf.io/wvs7m/.

References

36. Bonapersona, V., Joëls, M. & Sarabdjitsingh, R. A. Effects of early life stress on biochemical indicators of the dopaminergic system: a 3 level meta-analysis of rodent studies. *Neurosci. Biobehav. Rev.* **95**, 1–16 (2018).
37. Wickham, H. et al. Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
38. Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. shiny: Web Application Framework for R. *R Package version 1.5.0* https://CRAN.R-project.org/package=shiny (2020).
39. Ekstrøm, C. T. MESS: Miscellaneous Esoteric Statistical Scripts. *R package version 0.5.6* https://CRAN.R-project.org/package=MESS (2019).
40. Faul, F., Erdfelder, E., Lang, A. G. & Buchner, A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 (2007).
41. Lenth, R. V. Some practical guidelines for effective sample size determination. *Am. Stat.* **55**, 187–193 (2001).
42. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis*. (Chapman & Hall, 1995).

Acknowledgements

We thank J. Knop and M. Sep for helpful discussions and R. de Kloet for critically reviewing the manuscript. This work was supported by the Consortium of Individual Development (CID), which is funded by the Gravitation program of the Dutch Ministry of Education, Culture and Science and the Netherlands Organization for Scientific Research (NWO grant no. 024.001.003).

Author contributions

V.B. contributed to conceptualization, data curation, analysis, investigation, methodology, software, visualization and writing the manuscript; H.H. contributed to conceptualization, analysis, methodology, supervision and reviewing and editing the manuscript; members of the RELACS consortium provided the data; R.A.S. contributed to conceptualization, project administration, supervision and editing and reviewing the manuscript; M.J. contributed to conceptualization, funding acquisition, project administration, supervision and writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-020-00792-3.

Correspondence and requests for materials should be addressed to V.B.

Peer review information *Nature Neuroscience* thanks Stanley Lazic, Malcolm MacLeod, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature research

Corresponding author(s): Valeria Bonapersona

Last updated by author(s): Dec 1, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data used belonged to three categories: systematic review, simulation studies, and gathered primary data from other laboratories. Systematic review data was collected and is stored as two separate .csv files. The data of simulation studies was not stored, but can be retrieved by running the related code. Primary data gathered from other laboratories has been processed and stored in a unique .csv file. The data is available at https://osf.io/wvs7m/ .
Data analysis	All analysis is available at https://osf.io/wvs7m/. For the analysis, we used R (version 4.0.0) in the R studio environment on a macOS Mojave (version 10.14.6). The code for the RePAIR tool is available directly at the GitHub repository. We used the following R packages: tidyverse (version 1.3.0), shiny (version 1.5.0) and MESS (0.5.6). Part of the analysis (power calculations) were separately confirmed with G*Power (version 3.1.9.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that supports the findings of this study are openly available and can be downloaded at https://osf.io/wvs7m/ . By running the code, one can directly recreate the (unedited) images of this publication.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This section is applicable only for the case study data. We performed a power calculation considering a Welch two sided independent means t-test, with an alpha level of 0.05. Based on a meta-analysis, we estimated the expected effect size to be Hedge's G = 0.4. When considering these parameters, a total of 200 animals is required to achieve a power of 80%. We gathered this number of animals by aggregating datasets of 7 laboratories. A total of 275 animals was included, 132 controls and 143 experimental (early life adversity as limited nesting and bedding model) animals. The sample size of the dataset was therefore sufficient according to our power calculation.
Data exclusions	The data of the case study derives from a consortium, of which the purpose was larger than that of this specific case study. Blinded to the results, we specified a list of inclusion / exclusion criteria. For a full list of these criteria and the related explanations, please see Supplementary Table 4 in the Supplementary Information.
Replication	In this study, no new experimental data was generated, therefore replication of experiments was not required. For simulation studies and Bayesian analysis, we used a 10000x sampling strategy, and we discuss the results in terms of distributions.
Randomization	Evaluation of studies for the systematic review was performed in a random order. Briefly, each study was given a pseudo random number generated in R. This number was then used for the ordering and assessment of publications. For the case-study, presence of randomization was an inclusion criteria. However, we do not have information on how randomization was conducted by the single independent laboratories.
Blinding	Inclusion and exclusion criteria for the systematic review were defined before starting the review. The choice of distribution and ranges throughout the analysis (e.g. estimation of effect sizes, sensitivity variation range) was performed once the data was already collected, but prior to any data visualization. For the definition of prior information and the definition of inclusion/exclusion criteria for the case-study, the researcher (VB) did not have access to the effect sizes, but did have access to meta information of the study (e.g. characteristics of the ELA model).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging