MS Nature Methods

Abstract

Brief Communications begin with a brief unreferenced abstract (3 sentences, no more than 70 words),

Main

Preclinical research is essential for identifying promising interventions and to generate robust evidence to support translation to humans. Experiments in a preclinical setting are conducted in two different operating modes.[1] Early-stage preclinical experiments are *exploratory* with the aim to discover potentially effective interventions and generate hypotheses. These are tested at a later stage under more strict conditions in *confirmatory* mode.[2]

Various publications have called for a distinction between exploration and confirmation in preclinical animal research.[1–3] Mogil & Macleod[4] took up this sentiment and proposed that an initial claim must be independently confirmed in order to be published. This suggestion emphasizes confirmation of initial findings. However, the question remains which exploratory results should be contested during confirmation. There is currently no guidance regarding criteria that should be applied to make this decision.

Commonly, the *p*-value is used to make a decision whether or not to further investigate a claim. Given that prior probabilities at this stage of discovery are low and sample sizes are small,[5,6] the *p*-value might prematurely discard potentially promising interventions.[7]

Exploration requires sensitive tests to detect rare and possibly small effects.[7,8] As more sensitive criteria invite more false positive results, confirmation must aim at reducing false positives to ensure that only true effects are carried forward to clinical testing. To increase power and safeguard reliability of results, sample sizes must be increased when switching from exploratory to confirmatory mode. However, ethical, time, and budget constraints prevent preclinical researchers from increasing their sample size and thus complicate the detection and precise estimation of effects. In an effort to prevent false negative results in exploration and reduce false positives during confirmation, it is necessary to devise strategies to move from exploration to confirmation that meet these complementary goals.

Here, we consider how to move from exploratory to confirmatory mode while balancing the number of animals against the likelihood of false negative and false positive outcomes.

To this end, we simulated two preclinical research trajectories comprising an exploratory study and a first confirmatory study (Figure 1?).

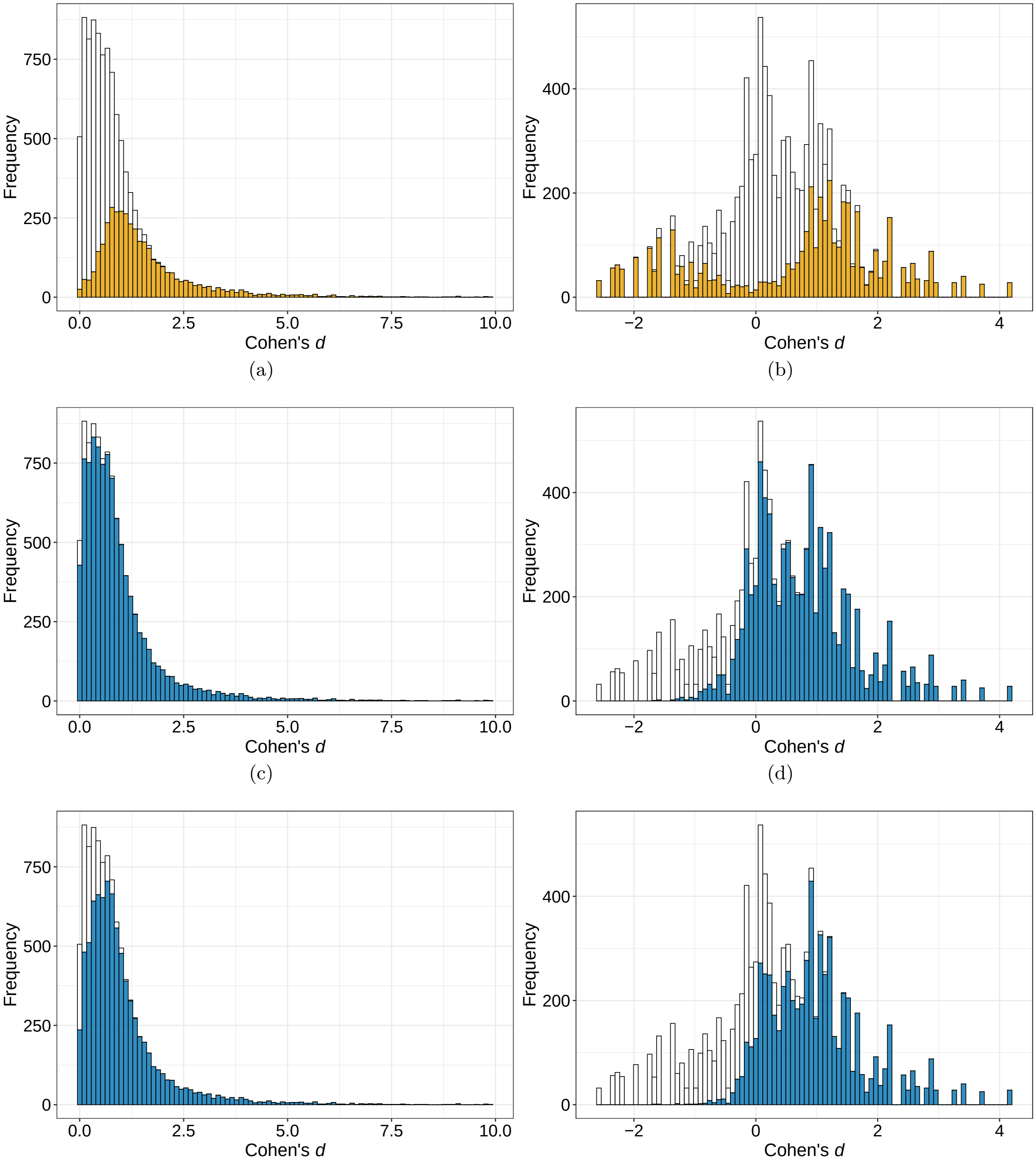
We based our simulations on two empirical effect size distributions reflecting an *optimistic* (high pre-study odds) and a *pessimistic* (low pre-study odds) scenario. After an initial exploratory study, a first decision identified experiments that should move from exploratory to confirmatory mode. One trajectory (standard) employed the conventional significance threshold ($\alpha = .05$) for this decision. The second trajectory (SESOI) used a more lenient threshold based on an *a priori* determined smallest effect size of interest (SESOI).

In the standard trajectory, in the optimistic scenario, 38.71 percent of experiments met the criterion $p \leq .05$. In the pessimistic scenario, 32.02 percent of experiments had a p-value $\leq .05$. A closer look at the effect sizes of the experiments that proceeded to replication reveals that the conventional significance threshold is a conservative filter. Many effect sizes smaller than 1 but larger than 0 are eliminated and are not further investigated (Figure 1a–b). This demonstrates that the conventional significance threshold is not useful to screen for potentially meaningful effects at the early stages of drug discovery even if pre-study odds are high.

In the SESOI trajectory, we estimated the exploratory effect size and 95 percent confidence interval (CI) around that estimate. We examined whether the CI covered our SESOI (0.5 and 1.0, respectively). If this was the case for an experiment, it advanced to confirmatory mode. Importantly, we did not consider the p-value additionally. Applying this decision criterion resulted in 85.81 and 81.45 percent of experiments moving to confirmation in case of the optimistic distribution and 65.46 and 61.65 percent for the pessimistic distribution based on an SESOI of 0.5 and 1.0, respectively. Compared to the conventional significance threshold, the range of effect size estimates that proceeded to confirmation shifted and included less extreme values, as well as more values closer to zero (Figure 1c–f). This decision criterion meets the goal of keeping the false negative rate low better than the conventional significance threshold.

In a second step, we calculated the sample size for a first confirmatory study. In the standard trajectory this was done via a standard power analysis using the initial exploratory effect size. The SESOI trajectory used again a pre-defined smallest effect size of interest (SESOI). Note that all sample sizes reported are the number of animals needed in *each* group (control and intervention).

In the standard trajectory, this resulted in a mean number of 7.2 (SD = 3.67) animals in the



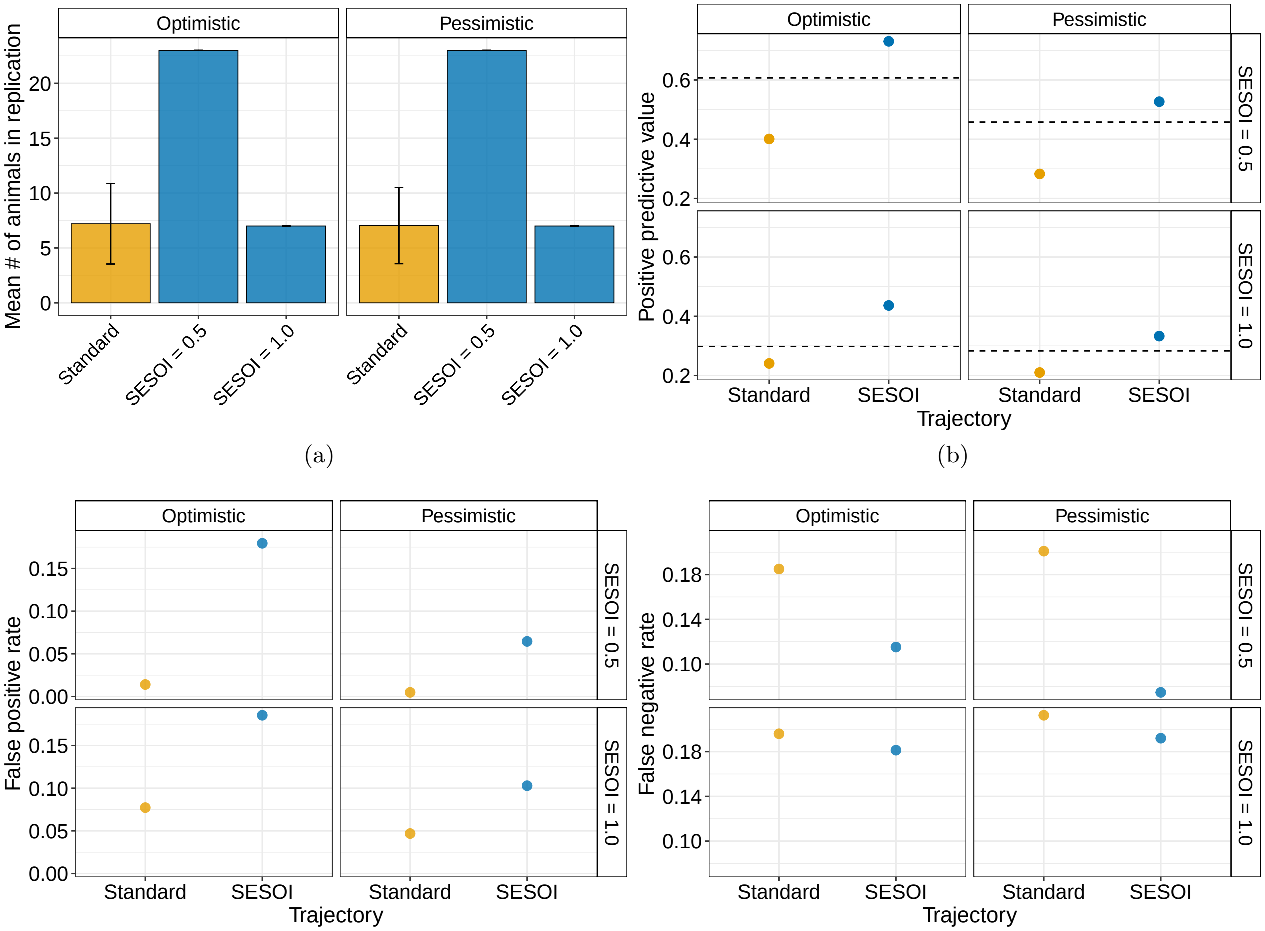
(a) (b) (c) (d) (e) (f)

Figure 1: Samples (n = 10000) drawn from the two empirical effect size distributions (left: optimistic, right: pessimistic). The panels show the underlying effect sizes that were detected using one of the two decision criteria. Orange shaded bars (panels (a) and (b)) indicate those effect sizes that were identified for replication using the conventional significance threshold (p = .05). Blue shaded bars indicate effect sizes that were selected using a SESOI of 0.5 (panels (c) and (d)) or 1.0 (panels (e) and (f)), respectively. Note that in (a), (c), and (e) 16 values > 10 were removed in order to display the distribution.

optimistic, and 7.04 (SD = 3.47) in the pessimistic scenario. These small numbers reflect the large effect sizes that passed on to confirmation and were the basis for sample size calculation (Supplementary Figure 2a–b).

In the SESOI trajectory, the number of animals varied with the SESOI that was chosen. For an SESOI of 1.0, 7 animals were needed in the replication in both the optimistic and pessimistic scenario. If the SESOI was 0.5, animal numbers increased to 23 (Figure **??**). Note that the sample sizes reported are the number of animals needed in *each* group (control and intervention).

We further calculated the positive predictive value (PPV), false positive rate (FPR), and false negative rate (FNR) across both trajectories. The positive predictive value (PPV) of a study is the post-study probability that a positive finding which is based on statistical significance reflects a true effect.[9] The PPV is calculated from the pre-study odds, as well as the sensitivity and specificity of the test. In our study, pre-study odds of an effect of a given size (0.5 and 1.0, respectively) were determined by the empirical effect size distributions. If evidence for an initial claim is strengthened throughout the preclinical research trajectory, we would observe an increased PPV compared to pre-study odds. In the optimistic scenario, the pre-study odds were 0.61 and 0.3 for SESOI of 0.5 and 1.0, respectively. In the pessimistic scenario pre-study odds were 0.46 and 0.28, respectively. Across the standard trajectory, the PPV drops below pre-study odds in both scenarios (Figure **??**). After the within-lab replication, the PPV is 0.4 and 0.24 in the optimistic scenario for SESOI of 0.5 and 1.0, respectively. In the pessimistic scenario, the PPV is 0.28 and 0.21. In the SESOI trajectory, employing a SESOI at both stages along the decision-making process elevates the PPV above pre-study odds. Given a SESOI of 0.5 and 1.0, respectively, the PPV is 0.73 and 0.44 in the optimistic scenario, and 0.53 and 0.33 in the pessimistic scenario. Across the standard trajectory, given the “optimistic” scenario, the FPR was 0.01 and 0.08 for SESOI of 0.5 and 1.0, respectively. In the “pessimistic” scenario, the FPR was 0.005 and 0.05 for SESOI of 0.5 and 1.0, respectively. Across the SESOI trajectory, the FPR increased to 0.18 and 0.19 in the “optimistic” scenario for SESOI of 0.5 and 1.0. Given the “pessimistic” scenario, the FPR was 0.06 and 0.1 for SESOI set to 0.5 and 1.0. Across the standard trajectory, given the “optimistic” scenario, the FNR was 0.19 and 0.2 for SESOI set to 0.5 and 1.0, respectively. In the “pessimistic” scenario, the FNR was 0.2 and 0.21 for SESOI of 0.5 and 1.0, respectively. Across the SESOI trajectory, the FNR decreased to 0.12 and 0.18 in the “optimistic” scenario for SESOI set to 0.5 and 1.0. Given the “pessimistic” scenario, the FNR was 0.07 and 0.19 for SESOI of 0.5 and 1.0



(a)

(b)

(c)

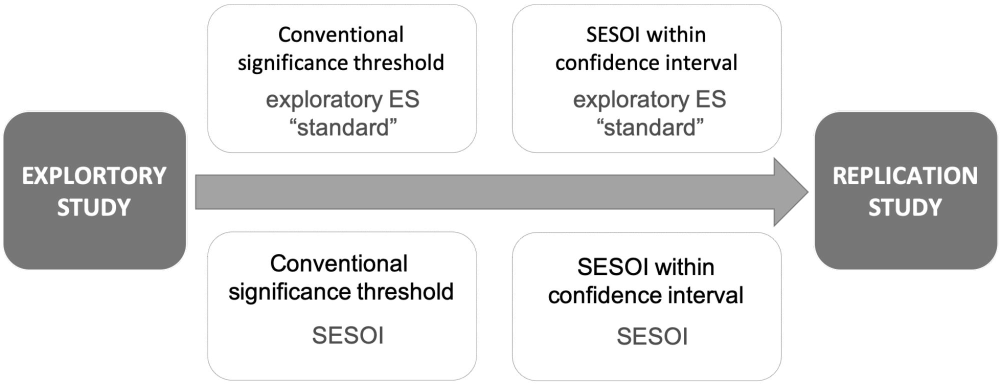
(d)

Figure 2: (a) Number of animals needed in the first confirmatory study. In the standard trajectory, sample sizes are low, as they are based on large exploratory effect sizes. Error bars represent standard deviations. In case of trajectories using a SESOI, the number of animals is fixed. (b) Positive predictive value across trajectory. Dashed lines indicate pre-study odds based on empirical effect size distributions.

Methods

Simulation. We explored different approaches to perform preclinical animal experiments via simulations. To this end, we modeled a simplified preclinical research trajectory from the exploratory stage to the results of a within-lab replication study (Figure 3). Along the trajectory, there are different ways to increase the probability of not missing potentially meaningful effects. After an initial exploratory study, a first decision identifies experiments for replication. In our simulation, we employed two different decision criteria that indicate when one should move from the exploratory to confirmatory mode. If a decision has been made to replicate an initial study, we applied two approaches to determine the sample size for a replication study (smallest effect size of interest (SESOI) and standard power analysis), as outlined in detail below.

We explored different approaches to perform preclinical animal experiments via simulations. To this end, we modeled a simplified preclinical research trajectory from the exploratory stage to the results of a within-lab replication study (Figure 3 make new figure and include two trajectories in MS and all 4 in Supplement). Along the trajectory, there are different ways to increase the probability of not missing potentially meaningful effects. After an initial exploratory study, a first decision identifies experiments for replication. In our simulation, we employed two different decision criteria that indicate when one should move from the exploratory stage to the replication [if you use this lingo, it should match the one in the intro] stage. If a decision has been made to replicate an initial study, we applied two approaches to determine the sample size for a replication study (smallest effect size of interest (SESOI) and standard power analysis), as outlined in detail below.



EXPLORTORY STUDY
Conventional significance threshold
exploratory ES "standard"
SESOI within confidence interval
exploratory ES "standard"
Conventional significance threshold
SESOI
SESOI within confidence interval
SESOI
REPLICATION STUDY

Figure 3: A preclinical research trajectory from the exploratory stage to a within-lab replication. The four panels along the arrow display four possible combinations of decision criteria and approaches to calculate the sample size for replication that can be employed throughout the trajectory.

*Empirical effect size distributions.*Simulations were based on empirical effect size distributions from the recently published literature.[10,11] This enabled us to determine the prior probability (pre-study odds) of a certain alternative hypothesis (H_1) which we defined as an effect of a given size (e.g. a Cohen's d of 0.5). The two distributions reflect different research fields.

The distribution of effect sizes extracted from Szucs & Ioannidis (2017)[10] contains 26841 effect sizes from the cognitive neuroscience and psychology literature published between January 2011 and August 2014. All effect sizes are calculated as the standardized difference in means (Cohen's d). Effect size estimates range from 0 to 298.5, and have a median of 0.65. As the pre-study odds of a medium effect of 0.5 are rather large (0.61), we will refer to this distribution as "optimistic". We acknowledge that the effect sizes were mainly extracted from human studies. However, in large parts the distribution is in agreement with effect sizes reported to be typical of (some areas of) preclinical research (find good ref here [This could be infectious diseases, usually antibiotics are all or nothing effects]).

As our study is concerned specifically with preclinical research, we chose a second distribution of empirical effect sizes to represent one field of the preclinical realm. Carneiro et al.'s (2018)[11] study systematically examined effect sizes in the rodent fear conditioning literature. Effect sizes were extracted from 410 experiments published in PubMed in 2013. The publication included a data file containing all extracted effect sizes. After removing missing values, the data set consisted of 336 effect sizes, again, calculated as Cohen's *d*. The effect sizes range from -2.6 to 4.14, and have a median of 0.38. The prior probability of observing an effect of 0.5 is 0.46. We will therefore refer to this distribution as "pessimistic". [in the limitations section you could cite the Bioarxiv paper and compare the two distributions to support your choice and that it may not be overly pessimistic. Particular in fields where there was no progress like neurodegenerative diseases]

*Exploratory stage.*From each of the two distributions, we drew 10000 samples of effect sizes from which we created 10000 study data sets. Each data set comprised data of two groups consisting of ten experimental units each drawn from a normal distribution. We chose a number of ten EUs based on reported sample sizes in preclinical studies.[6] [For the limitation section we need to argue that different initial sample sizes like 7 or 15 will not change results dramatically]. Our simulated design mimics a comparison between two groups where one receives an intervention and the other functions as a control group. The study data sets are compared using a two-sided two-sample *t*-test. From these exploratory study results, we extracted the *p*-values and 95 percent confidence intervals (CI). We then employed two different criteria based on the *p*-value or 95 percent CI, respectively, to decide whether to continue to a replication.

*Decision criteria to proceed to replication.*The first decision criterion employs the conventional significance threshold (α = .05) to decide whether to replicate an exploratory study. If a *p*-value extracted from a two-sided two-sample *t*-test is $\leq$ .05, this study will proceed to the replication stage. If not, the trajectory is terminated after the exploratory study. We chose this decision criterion as our reference, as this is what we consider to be current practice.

As an alternative to this approach, we propose to set a smallest effect size of interest (SESOI) and examine whether the 95 percent CI around the exploratory effect size estimate covers this SESOI. A SESOI is the effect size that the researcher based their knowledge of the literature in their respective field (domain knowledge) and given practical constraints considers biologically and clinically meaningful.[12] In our simulation, we used 0.5 and 1.0 as SESOI. This approach emphasizes the importance of effect sizes rather than statistical significance to evaluate an intervention's effect. Further, we expected this approach to be more lenient than statistical significance (at least if the significance threshold was set at $\alpha = .05$) and to allow a broader range of effect sizes to pass on to be further investigated.

Approaches to determine sample size for replication. Once the decision to continue to replication has been made, we employed two different approaches to determine the sample size for the replication study. After having conducted an exploratory study, we have an estimate of the direction of the effect. Only effect sizes that showed an effect that favors the treatment over the control group were considered for further investigation. Thus, for the replication study, a one-sided two-sample *t*-test was performed. The desired power level for replication was set to .80, α was set to .05. In order to calculate the sample size for replication given power and α, an effect size estimate is required. In one approach, we used the exploratory effect size estimate to compute the replication sample size. In an alternative approach, we employed the same SESOI used as decision criterion earlier. In statistical terms, our SESOI was set such that the replication study would have a power of .50 to detect an effect of this size (# explain in more detail and find good ref to motivate this: Lakens? [This is mainly to ensure that the likelihood of a type I error below this threshold is negligible. The goal is to reduce Type I error in this second phase.]). Consequently, in the first approach, the replication sample size was dependent on the outcome of the exploratory study, whereas using a SESOI always yielded the same sample size regardless of the exploratory effect size (e.g. 23 EUs in each group for a SESOI of 0.5).

Replication stage. For each of the studies that met the decision criterion after the exploratory study (either $p \leq .05$ or SESOI within the 95 percent CI of the exploratory effect size estimate), a replication study was performed. The number of replication studies conducted varied with the decision criterion used and, in case of the criterion employing a SESOI, also with the SESOI (0.5 and 1.0). A replication study was performed as a one-sided two-sample *t*-test, where the number of animals in each group was determined by the approach to calculate the sample size. For a replication be considered "successful", the *p*-value had to be below the conventional significance threshold (α = .05).

Trajectories. We compared the two trajectories (Standard and SESOI) regarding the number of experiments proceeding to the replication stage, number of animals needed in the replication, and positive predictive value across the trajectory. Secondary outcomes are the false positive rate, false negative rate, and effect size precision. Outcome variables are outlined in more detail in the following section. The standard trajectory constitutes our reference, as we consider it to be closest to current practice. We have stored data, results, and figures of all four trajectories in an online repository (insert URL here).

References

1. Kimmelman, J., Mogil, J. S. & Dirnagl, U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS biology* **12**, e1001863 (2014).
2. Landis, S. C. *et al.* A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490**, 187–191 (2012).
3. Dirnagl, U. Thomas willis lecture: Is translational stroke research broken, and if so, how can we fix it? *Stroke* **47**, 2148–2153 (2016).
4. Mogil, J. S. & Macleod, M. R. No publication without confirmation. *Nature News* **542**, 409 (2017).
5. Macleod, M. R. *et al.* Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* **39**, 2824–2829 (2008).
6. Howells, D. W., Sena, E. S. & Macleod, M. R. Bringing rigour to translational medicine. *Nature Reviews Neurology* **10**, 37 (2014).
7. Dirnagl, U. Resolving the tension between exploration and confirmation in preclinical biomedical research. *Good Research Practice in Non-Clinical Pharmacology and Biomedicine* 71–79 (2020).
8. Bonapersona, V., Hoijtink, H., Sarabdjitsingh, R., Joels, M. & others. RePAIR: A power solution to animal experimentation. *BioRxiv* 864652 (2019).
9. Ioannidis, J. P. Why most published research findings are false. *PLoS medicine* **2**, e124 (2005).
10. Szucs, D. & Ioannidis, J. P. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology* **15**, e2000797 (2017).
11. Carneiro, C. F., Moulin, T. C., Macleod, M. R. & Amaral, O. B. Effect size and statistical power in the rodent fear conditioning literature–a systematic review. *PloS one* **13**, e0196258 (2018).
12. Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science* **1**, 259–269 (2018).