Increasing discovery rates in preclinical research through optimised statistical decision criteria

Abstract

The success of scientific discovery in preclinical drug development is based on the different roles of exploration and confirmation in this process. Via simulations, we show that our systemic approach - based on a smallest effect size of interest - increases discovery rates compared to a p-value based approach while keeping animal numbers low. Based on our findings, we argue for a reconsideration of planning and conducting preclinical experiments.

Main

Preclinical research is essential for identifying promising therapeutic interventions and to generate robust evidence to support their translation to humans. To achieve these goals, experiments 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]

This sequential approach putatively increases the likelihood that generated evidence is robust and potentially reduces translational failures.[3–5] But which exploratory results are promising and worthwhile to confirm? Here we compare two approaches to select and conduct confirmatory studies with regard to their success in discovering true effects. In one approach, decisions for selecting a study are based on the p-value, the ubiquitous criterion in the biomedical publication record. An alternative approach focuses on an *a priori* defined smallest effect size of interest (SESOI), similar to a minimally clinically important difference in clinical trials.[6] We show through simulations that the latter approach increases discovery rates indexed by an improved positive predictive value.[7]

The effectiveness of any approach is based on the different roles exploration and confirmation fulfill in scientific discovery. Via sensitive tests exploration detects potentially relevant effects among many hypotheses that often have a low prior probability of being true. The sensitivity of a test gives the probability of correctly identifying true effects (also known as the true positive rate or power). 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 subsequent testing. To complicate matters, ethical, time, and budget constraints limit degrees of freedom in experimental design. Consequently, to prevent false negative results in exploration and to reduce false positives during confirmation, it is necessary to devise strategies to optimize these complementary goals when advancing from exploration to confirmation. Specifically, the likelihood of false negative and false positive outcomes needs to be balanced with the goal to minimize the number of animals needed.

To this end, we simulated two preclinical research trajectories each comprising an exploratory study and a subsequent confirmatory study (Fig. 1a). We based our simulations on a published effect size distribution derived from empirical research[8] (Fig. 1b). After an initial exploratory study using a two-sided two-sample Welch's t-test, a decision criterion selected experiments to advance 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* defined smallest effect size of interest. For this, we estimated the effect size of each exploratory study and its 95 % confidence interval (CI). We examined whether this CI covered our SESOI (Hedges' g of 0.5 and 1.0, respectively). Statistically, this method is similar to a non-inferiority test with our SESOI as equivalence threshold. Importantly, we did not consider significance, thus, with this method significant and non-significant results were carried forward to confirmation.

As expected, more experiments transitioned to the confirmatory phase in the SESOI trajectory compared to the Standard trajectory (SESOI 0.5: 83.32 %, SESOI 1.0: 74.85 %, Standard: 32.72 %; Fig. 1c–e). Out of the 10000 sampled effect sizes 132 were zero. Thus, in the Standard trajectory 6596 effect sizes > 0 were missed and the smallest effect size detected was 0.9, whereas in the SESOI trajectory 1536 and 2383 positive effect sizes were falsely eliminated given a SESOI of 0.5 and 1.0, respectively. This demonstrates that the significance threshold risks to discard potentially meaningful effects at early stages of discovery. In contrast, the smallest effect sizes selected for confirmation in the SESOI trajectory were 0.0001 (SESOI = 0.5) and 0.16 (SESOI = 1.0) when a 95 % CI was used. This decision criterion thus reduced false negatives compared to the conventional significance threshold, but at the same time made more confirmatory studies necessary.

In a subsequent step, we estimated the sample size for a confirmatory study. In the Standard trajectory this was done via a power analysis using the initial exploratory effect size. The SESOI trajectory used the pre-defined SESOI that was employed as decision criterion earlier. Importantly, we set the power to .50 to detect an effect of the size of our SESOI (for more details see Methods section).

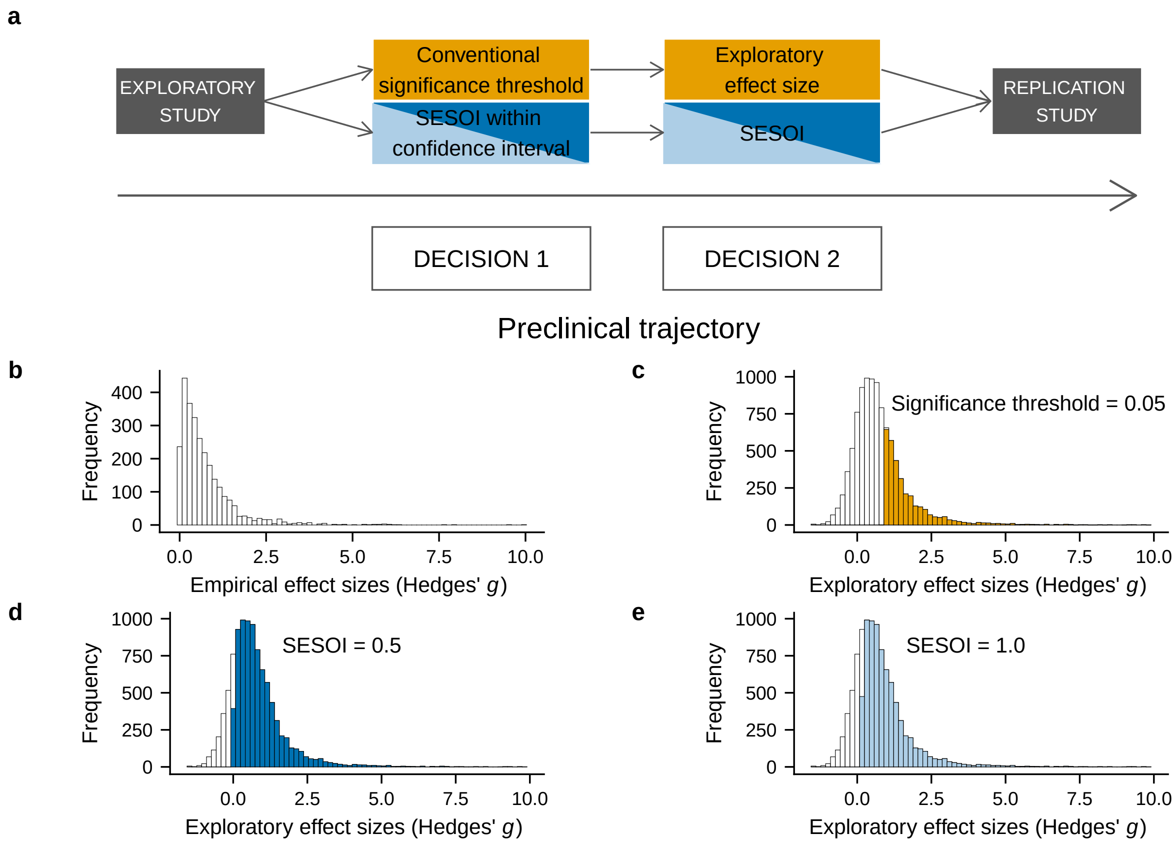

Figure 1: **Preclinical research trajectory and transition rates.** **a**, Along the trajectory, two decisions have to be made. DECISION 1: Which experiments should move from exploration to confirmation? DECISION 2: How should the sample size for a confirmatory study (i.e. within-lab replication) be estimated? The yellow and blue panels display the decision criteria and approaches to estimate the sample size along the two tajectories that were compared in this study (yellow = Standard; blue = SESOI). **b**, Distribution of empirical effect sizes ($n = 2729$) extracted from Bonapersona *et al.*, 2021. **c-e**, Exploratory effect sizes ($n = 9958$). The shaded bars show the effect sizes that were detected using one of the two decision criteria. **c**, Yellow shaded bars indicate those effect sizes that were identified for confirmation using the conventional significance threshold ($\alpha = .05$). **d**, Dark blue shaded bars indicate effect sizes that were selected using a SESOI of 0.5. **e**, Light blue shaded bars indicate effect sizes that were selected using a SESOI of 1.0. Note that in panels **b-e** values > 10 were removed in order to display the distribution.

In the Standard trajectory, this resulted in small sample sizes (Standard: n_{rep} $(M \pm SD) =$ 8.08 ± 4.05; Fig. 2a). In the SESOI trajectory, the number of animals varied with the SESOI that was chosen (SESOI 0.5: $n_{rep} = 23$, SESOI 1.0: $n_{rep} = 7$; Fig. 2a). The small numbers in the Standard trajectory reflect the large effect sizes that passed on to confirmation and formed the basis for confirmatory sample size estimation (Fig. 1c).

To estimate discovery success, we calculated the positive predictive value (PPV), false positive rate (FPR), and false negative rate (FNR) across both trajectories. The PPV of a study is the post-study probability that a positive finding which is based on statistical significance reflects a true effect.[7] The PPV is calculated from the prior probability (or prevalence), as well as the sensitivity and specificity of the test. In our study, the prior probability of an effect of a given size (Hedges' g of 0.5 and 1.0, respectively) was calculated from the empirical effect size distribution. For example, if we wanted to know the prior probability of an effect Hedges' $g = 0.5$, we divided the number of effects ≥ 0.5 by the total number of effects in the distribution. For an effect Hedges' $g = 0.5$, the prior probability was 0.5, for Hedges' $g = 1.0$, the prior probability was 0.24. If evidence for an initial claim is strengthened throughout the preclinical research trajectory, we would observe an increased PPV compared to the prior probability.

Employing a SESOI at both stages along the decision-making process elevates the PPV above the prior probability (SESOI 0.5: 0.64, SESOI 1.0: 0.34). In comparison, across the Standard trajectory, the PPV drops below the prior probability (Standard 0.5: 0.3, Standard 1.0: 0.19; Fig. 2b).

The FPR was consistently higher in the SESOI trajectory compared to the Standard trajectory (SESOI 0.5: 0.16, SESOI 1.0: 0.15, Standard 0.5: 0.01, Standard 1.0: 0.06; Fig. 2c).

Whereas in the Standard trajectory the FNR was consistently close to 20 percent, it was considerably lower in the SESOI trajectory for smaller SESOI (SESOI 0.5: 0.13, SESOI 1.0: 0.2, Standard 0.5: 0.18, Standard 1.0: 0.18; Fig. 2d).

Overall, our simulation shows that current practice reflected by the Standard trajectory does not meet the complementary goals of exploration and confirmation along a preclinical research trajectory. In the Standard trajectory, the switch from exploratory to confirmatory mode eliminates numerous potentially meaningful effects. Thus, in the Standard trajectory, the prior probability of an effect of 0.5 and 1.0, respectively, was considerably decreased in the sample of studies that transitioned to confirmation (Hedges' $g = 0.5$: 0.29, Hedges' $g = 1.0$: 0.2.). This, as well as lower sensitivity and specificity across the trajectory resulted in low discovery rates as indexed by the

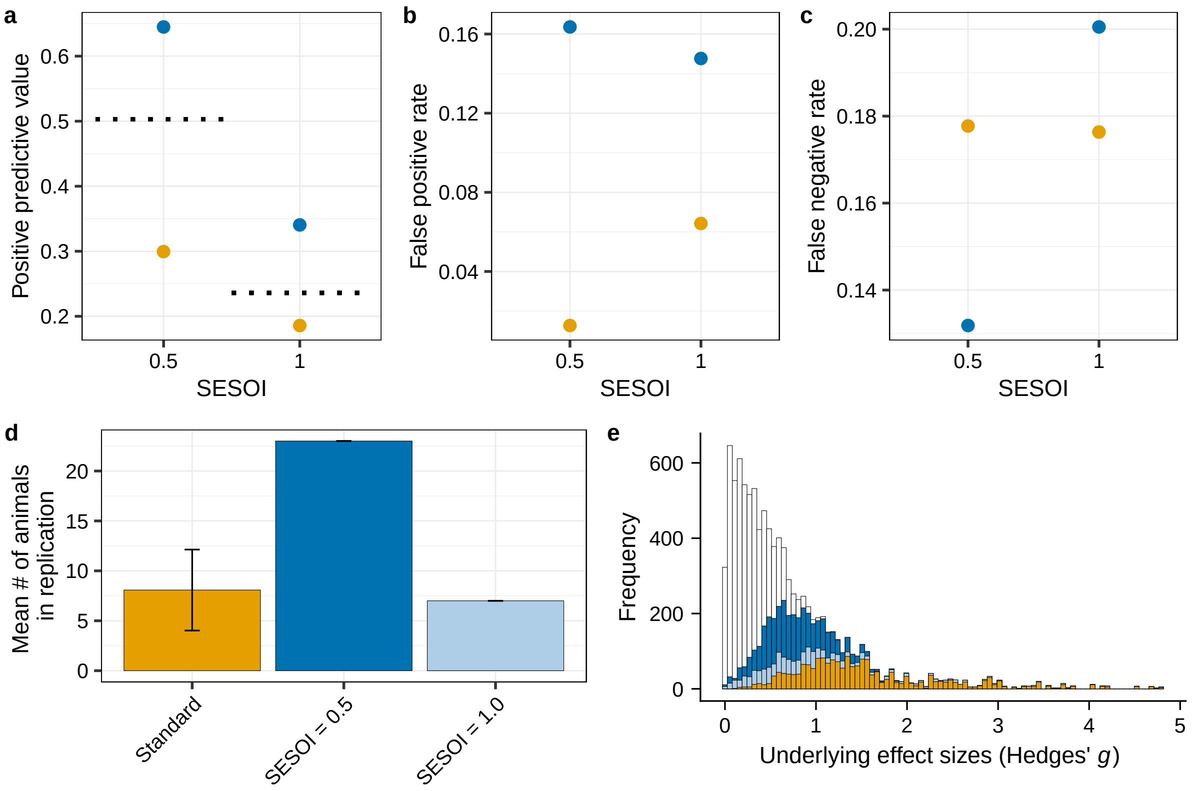

Figure 2: **Outcomes across trajectory, sample sizes in confirmation, and significant effect sizes after replication.** **a**, The positive predictive value across the trajectory is consistently above the prior probability (dotted lines) only in the SESOI trajectory. **b**, The false positive rate across the trajectory is consistently lower in the Standard trajectory. **c**, The false negative rate across the trajectory varies as a function of trajectory and SESOI. **d**, Number of animals needed in the confirmatory study. In the Standard trajectory, sample sizes are low ($n_{rep} = 8.08 \pm 4.05$), 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. Using a SESOI of 0.5 results in $n_{rep} = 23$, whereas a SESOI of 1.0 achieves a sample size comparable to those in the Standard trajectory ($n_{rep} = 7$). All sample sizes displayed and reported in the text are the number of animals needed in *each* group (control and intervention). **e**, The histogram displays the distribution of sampled effect sizes ($n = 9875$, values > 5 were removed for better display), that constituted our underlying true effect sizes in the simulation. Shaded bars indicate which underlying effect sizes were detected in the two-stage process. Dark blue shaded bars represent effect sizes that were detected across the SESOI trajectory employing a SESOI of 0.5. Light blue bars represent effect sizes that were detected employing a SESOI of 1.0. The median of eliminated effect sizes is 0.26 (SESOI = 0.5) and 0.35 (SESOI = 1.0) in the SESOI trajectory. Yellow shaded bars indicate effect sizes that were detected across the Standard trajectory. The median of eliminated effect sizes is 0.39 in the Standard trajectory. Using the Standard trajectory for detection is not efficient as it overlooks numerous potentially meaningful effects.

PPV. In the SESOI trajectory, the prior probability did not appreciably drop from exploration to confirmation (Hedges' $g = 0.5$: 0.49, Hedges' $g = 1.0$: 0.23), resulting in increased discovery rates across the trajectory. A more lenient significance threshold ($\alpha = 0.1$) did not change the direction of this difference (Fig. S8–S9), neither did a slight decrease/increase of the initial sample size (Fig. S3–S5). Also, a combination of the decisions and sample size calculations favored the SESOI trajectory (Fig. S7). This holds even when the SESOI was as small as 0.1. As a robustness test, we also tested a more pessimistic published effect size distribution and results were preserved (Fig. S10–S11). In light of our findings, we advise a reconsideration of planning and conducting experiments in preclinical research.

Whereas the SESOI trajectory achieves a higher PPV compared to the standard trajectory, and its FNR is in line with the canonical type 2 error rate of 20 percent, the FPR is high. However, in the framework of preclinical research trajectories, confirmation is an incremental process in which experiments are gradually refined to eliminate false positives and increase reliability. Our results underscore the need for a systematic estimation of true effects. We therefore propose to proceed following these steps when embarking on an investigation: 1. Choose a SESOI. The rationale for the choice of a SESOI should be transparently reported and can be based on various considerations, like feasibility given practical constraints or meta-analytic effect size estimates in your field.[9] 2. Conduct an exploratory study and apply the pre-specified SESOI as decision criterion to decide whether to move forward to confirmation. Be aware that the width of your CI is a function of the sample size, i.e., the less animals you use, the wider the CI. The CI, in which the SESOI should fall, needs potentially be adjusted based on the reliability of the initial study. This has to be done also *a priori*. 3. If the decision points to a transition to confirmatory mode, use the SESOI to calculate the confirmatory sample size with a power of .50, ideally using a one-sided test. 4. Conduct a confirmatory study. Here, we chose significance as a measure of confirmation success. However, other methods such as the estimation of effect sizes and their CI or the sceptical p-value[10] might be more informative.

Our results are dependent on underlying assumptions represented by the empirical effect size distribution from Bonapersona *et al.* (2021). If effect sizes in a field considerably deviate from this distribution, following our recommendation might be dispensable, as the issue of low sensitivity to detect effects does not arise. Our method specifically addresses preclinical research areas that need to weigh the number of animals against the likelihood of inference errors. This is the case in fields where effect estimates are small or medium.

The method we present here is an easily applicable alternative to current practice in preclinical animal research. Optimizing decision criteria and sample size calculations by employing a SESOI increases chances to detect true effects while keeping the number of animals at a minimum.

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 confirmatory study (within-lab replication; Fig. 1a). 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 power analysis based on the exploratory effect size), as outlined in detail below.

Empirical effect size distribution. Simulations were based on an empirical effect size distribution from the recently published literature.[8] From this distribution, we were able to calculate the prior probability of a certain alternative hypothesis (H_1) which we defined as an effect of a given size. For example, if we wanted to know the prior probability of an effect Hedges' $g = 0.5$, we divided the number of effects ≥ 0.5 by the total number of effects in the distribution.

The distribution of effect sizes extracted from Bonapersona *et al.*, (2021)[8] contains 2738 effect sizes from the fields neuroscience and metabolism. All effect sizes are calculated as the standardized difference in means (Hedges' g). Effect size estimates range from 0 to 24.61, and have a median of 0.5.

Exploratory mode. From the empirical effect size distribution, 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 (EUs) each drawn from a normal distribution. We chose a number of ten EUs based on reported sample sizes in preclinical studies.[8,11] Our simulated design mimics a comparison between two groups where one group receives an intervention and the other serves as a control group. The difference between the intervention and control group in each study data set was determined by the effect size that was drawn from the empirical effect size distribution. In that way, we ensured that our simulated data was comparable to data in the published literature in the respective fields (neuroscience and metabolism). In the simulated exploratory study, the study data sets are compared using a two-sided two-sample Welch's t-test. From these exploratory study results, we extracted p-values, exploratory effect size estimates and their 95 % confidence intervals (CI). We then employed two different criteria based on the p-value or 95 % CI, respectively, to decide whether to continue to confirmatory mode.

Decision criteria to proceed to confirmation. The first decision criterion employed the conventional significance threshold ($\alpha = .05$) to decide whether to replicate an exploratory study. If a p-value extracted from a two-sided two-sample Welch's t-test was $\leq .05$, this study proceeded to confirmation. If not, the trajectory was 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 % CI around the exploratory effect size estimate covers this SESOI. A SESOI is the effect size that the researcher based their domain knowledge and given practical constraints considers biologically and clinically meaningful.[9] Our proposed method is similar to a non-inferiority test. In our case the null hypothesis stated that the exploratory study identified an effect that is inferior to our pre-defined SESOI. If this was not the case for an exploratory study, i.e., if the CI around the exploratory effect size covered our SESOI, non-inferiority was established and the study was carried forward to confirmatory mode. Importantly however, we did not consider significance in this approach. In our simulation, we used Hedges' $g = 0.5$ and Hedges' $g = 1.0$ 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 and to allow a broader range of effect sizes to pass on to be further investigated.

Both decision criteria identified effect sizes < 0 to transition to confirmation. This would mean that an experiment that showed an effect in favor of the control would be replicated. As we think this is an unrealistic scenario, negative effect sizes were not carried forward to confirmation in the simulation.

Approaches to determine sample size for replication. Once the decision to continue to confirmatory mode has been made, we employed two different approaches to determine the sample size for the confirmatory study. After the exploratory study, only effect sizes that showed an effect in favor of the treatment (> 0) were considered for further investigation. Thus, for the confirmatory study, a one-sided two-sample Welch's t-test was performed. In the standard trajectory, the desired power level for replication was set to .80, α was set to .05. To determine the sample size given power and α, we used the exploratory effect size estimate. In the SESOI trajectory, we employed the same SESOI used as decision criterion earlier to calculate the confirmatory sample size. Our SESOI was set such that the confirmatory study has a power of .50 to detect an effect of this size. This power level was chosen to ensure that the likelihood of a false positive finding below the threshold determined by our SESOI is negligible. The aim during confirmation is to weed out false positives. Additionally, 0.5 is only the *smallest* effect size we are interested in, meaning all other effect sizes of interest (i.e., effect sizes > 0.5) have a higher chance of being detected. This increase in power is larger with a power level of 50 % compared to a power level of 80 % as illustrated by the power curves in Fig. S2.

Confirmatory mode. For each of the studies that met the decision criterion after the exploratory study (either $p \leq .05$ or SESOI within the 95 % CI of the exploratory effect size estimate), a confirmatory study was performed. The number of 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 confirmatory study was performed as a one-sided two-sample Welch's t-test, where the number of animals in each group was determined by the approach to determine the sample size. For a confirmatory study to be considered "successful", the p-value had to be below the conventional significance threshold ($\alpha = .05$).

Outcome variables. We compared the two trajectories (Standard and SESOI) regarding the transition rates from exploration to confirmation, number of animals needed in the confirmatory study, and positive predictive value (PPV), false positive rate (FPR), and false negative rate (FNR) across the trajectory. Importantly, the PPV was calculated from the known prior probability (given by the empirical effect size distribution), the sensitivity (true positive rate), and the specificity (true negative rate). To make the rates comparable for the Standard and SESOI trajectory, we did not specify a non-existent effect as zero, but as an effect that is smaller than our SESOI (0.5 and 1.0, respectively). A true positive thus is a significant result that has an underlying effect of at least 0.5 (or 1.0, if our SESOI was 1.0). If a significant result represents an underlying effect smaller than 0.5 (or 1.0), this is categorized as a false positive (likewise for true and false negatives). This is a departure from the classical definition in the null hypothesis significance testing framework, where the benchmark against which is tested is zero.

Additional trajectories. In addition to the two trajectories compared in the main text, we also performed all simulations and analyses with the two crossed-over trajectories where each decision criterion is combined with the respective other approach to determine the sample size for confirmation. This resulted in the two trajectories SESOI-Standard (SESOI as decision criterion to move from exploration to confirmation + sample size estimation using the exploratory effect size estimate) and Standard-SESOI (Conventional significance threshold as decision criterion + SESOI for sample size estimation). Sample sizes in confirmation as well as outcomes across the trajectories are shown in Fig. S6 and Fig. S7.

Robustness checks. As robustness checks, we simulated all trajectories with an initial sample size (n_{init}) of 5, 7 and 15 animals per group (in addition to 10 as described in the main part). We further performed all simulations and analyses with additional SESOI 0.1, 0.3, and 0.7. Outcomes are displayed in Fig. S3–S5. We also varied the significance threshold employed as decision criterion after exploration in the standard trajectory. The outcomes resulting from a more lenient significance threshold of $\alpha = 0.1$ are displayed in Fig. S8–S9. For the SESOI trajectory, we additionally varied the confidence interval around the exploratory effect size estimate used to determine whether to transition to confirmation. Transition rates are displayed in Fig. S1. Lastly, for the SESOI trajectory, we calculated the sample size for confirmation setting the power to .80 to investigate whether this would have beneficial effects on the positive predictive value. Outcomes are shown in Fig. S2. We ran the whole simulation also with a more pessimistic published effect size distribution.[12] Outcomes are displayed in Fig. S10–11.

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. Drude, N. I., Gamboa, L. M., Danziger, M., Dirnagl, U. & Toelch, U. Science forum: Improving preclinical studies through replications. *eLife* **10**, e62101 (2021).
6. Chuang-Stein, C., Kirby, S., Hirsch, I. & Atkinson, G. The role of the minimum clinically important difference and its impact on designing a trial. *Pharmaceutical statistics* **10**, 250–256 (2011).
7. Ioannidis, J. P. Why most published research findings are false. *PLoS medicine* **2**, e124 (2005).
8. Bonapersona, V., Hoijtink, H., Sarabdjitsingh, R. & Joëls, M. Increasing the statistical power of animal experiments with historical control data. *Nature Neuroscience* 1–8 (2021).
9. Lakens, D. Sample size justification. (2021).
10. Held, L. A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **183**, 431–448 (2020).
11. Howells, D. W., Sena, E. S. & Macleod, M. R. Bringing rigour to translational medicine. *Nature Reviews Neurology* **10**, 37 (2014).
12. 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).