

# Identifying optimized decision criteria and experimental designs by simulating preclinical experiments in silico

## 1 Theoretical Background

Preclinical research is essential for identifying promising interventions and to generate robust evidence to support translation to humans. The experiments in a preclinical setting are conducted in two different operating modes [1]. Early-stage preclinical experiments are *exploratory* with the aim to generate hypotheses. These are tested at a later stage under more strict conditions in *confirmatory* mode [2]. At early stages of treatment discovery, preclinical research also plays the role of a gatekeeper – straining out promising interventions and denying access to further investigation to non-effective ones. It is vital, however, that decision criteria are lenient enough to allow also uncertain effects to pass. If the filters applied at this stage are too strict, chances are that potentially effective candidates are falsely eliminated.

Animal models are designed to enable an informed decision to engage in a clinical trial, so one would expect that they predict efficacy and safety in humans, at least to a certain extent. However, decades of failed clinical trials prove otherwise. A report of clinical development success rates from 2006 through to 2015 documents low success rates for Phase II trials, with only 30.7 % of candidate drugs progressing to Phase III [3]. For specific medical conditions like amyotrophic lateral sclerosis (ALS), Alzheimer’s disease (AD), and stroke, decades of research have yielded surprisingly few advances with regard to effective treatments [4–8].

Preclinical research has increasingly come under scrutiny with many reports pointing out low study quality, lack of scientific rigor and high levels of irreproducibility [9]. Among the issues raised are (selective reporting of results [10], insufficient reporting on measures to reduce bias [11],) low sample sizes [12, 13] and concomitant increased false positive and false negative rates [14]. These shortcomings threaten the replicability of preclinical results

and hamper translation of promising findings to humans [15, 16]. Additionally, effect size estimates from low power studies carry a lot of uncertainty and are prone to type M errors. That is, effect magnitudes are larger than the true unknown effect [17, 18].

Along the trajectory, we have to make sure that at a given stage (which implies that we are in a specific operating mode), our strategy is aligned with the goal that we have at this stage. At early stages, we want to find possibly working interventions. In statistical terms this means that we aim for a low Type II error. We do not want to miss effects. That means that our tests should be sensitive. This comes at the cost of specificity. We allow for false positives at this stage. Dirnagl (2016) [4]

\*In exploratory mode, we allow for more false positives. This happens naturally, as we operate with low prior odds that our hypothesis is true.

- focus in exploration should rather be on proper descriptive statistics, including measures of variance and confidence intervals.
- 

[Here you need to focus more. You need to start with the actual problem. The question you want to answer is: How to switch from exploratory to confirmatory mode. No need to go into the shortcomings, etc. In the paragraph above I would also take up the exploratory confirmatory narrative from the first paragraph and find literature that describes how to do this. Ullis paper is pretty explicit that Type I and Type II error plays a major role here. Build on this and also look whether other literature makes recommendations for this. Describe these, but make clear that there is no satisfying solution for this important problem. We will provide suggestion here how to solve this problem. I would definitively remove many of the papers describing the other shortcomings, but focus, focus, focus... We can talk about this soon. ]

The stark contrast between the importance of preclinical research and its shortcomings demands a change to design preclinical experiments that yield robust evidence to guide decision making towards translation. Here, we consider how to do this by simulating a preclinical research trajectory. We understand such a trajectory as a series of experiments, successive in time, generating evidence to support the decision to carry an intervention

forward to clinical testing. Whereas these experiments may include different types of studies, we focus on studies examining the efficacy of a given intervention. The trajectory, for the purpose of this study, comprises exploratory studies as the first stage and within-lab replications as the second stage and we consider the state and quality of evidence generated throughout the first two stages only. Specifically, we examine the effect of different decision criteria and design choices on transition rates from the exploratory to the replication stage, as well as success rates after within-lab replication. **maybe remove** (When moving closer towards a decision regarding translation to humans, more stages are needed (for example, between-lab replications, multicenter studies, inclusion of positive and negative controls))

For the purpose of this project, we understand replications as studies whose outcome is taken as “diagnostic evidence about a claim from prior research” [19]. **[As you know there was some backlash in the elife paper on this so we need to be more specific here. You also introduce replication here. Wouldn't it in this context be enough to refer to the switch between exploratory and confirmatory mode. You are also looking in your simulation at a direct replication without changes to design despite the sample size. You are in our framework thus not concerned with validity but reliability of a finding]** In the framework of preclinical research trajectories, replications are considered part of a confirmation process. Given that a series of experiments is needed to confirm a hypothesis about a directional relationship (compound X causes reaction Y), replications incrementally solidify evidence supporting (or refuting) an initial claim. **[I think we need replication as a term, but perhaps not in such a detailed manner]**

When planning a replication study, the sample size has to be determined *a priori*. This is usually done using an effect size estimate from prior research. This can be taken from the literature or a pilot study. In both cases, the effect size estimate used is most likely inflated due to reasons outlined above. **[readers will have forgotten about these reasons already, please also cite literature here that shows an inflated effect size]** Employing an overestimated effect size for sample size calculation given a desired power level (for example, 0.8) and significance threshold (for example, .05), will yield a sample size that is too low to reliably detect effects. This is especially true if true underlying effects are small. When statistical power is low, non-replications denote false negatives under the assumption that an underlying effect exists.

An insufficiently powered replication is not suited to support or refute a previously made claim, as success and failure might be accounted for simply by inference errors.

Above and beyond the concerns mentioned, preclinical animal research raises challenging ethical concerns. Therefore, the practice oriented 3R principles [20] seek to reduce harm and constrain the use of animals for scientific purposes to a minimum. Current animal welfare regulations undermine the scientific value of animal research, as they fail to integrate principles that ensure high quality of conducted studies [21]. From an ethical point of view, this is problematic in two ways: animals are sacrificed for inconclusive research and translation to patients is impeded. Preclinical animal research needs to balance animal welfare considerations and scientific benefit. Efforts to promote and implement the 3R's must not come at the cost of scientific value. [This comes pretty late, also the beginning of the next paragraph could be higher up. If researchers balance exploratory and confirmatory research they need to consider ethical, time, and budget constraints.]

Ethical, time, and budget constraints prevent preclinical researchers from increasing their sample size and thus complicate the detection and precise estimation of effects. This study investigates how to design and conduct preclinical animal experiments that balance the number of animals tested against the likelihood of false negative and false positive outcomes. Our approach strives to provide a systemic perspective that does not evaluate outcomes of single experiments but evidence that is generated throughout a preclinical research trajectory. Further, we explore conditions under which replication in preclinical settings is most sensible and feasible given prior evidence and practical constraints.

## 2 Methods

### 2.1 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 1). 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 stage [if you use this lingo, it should match the one in the intro]. 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 1). 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.

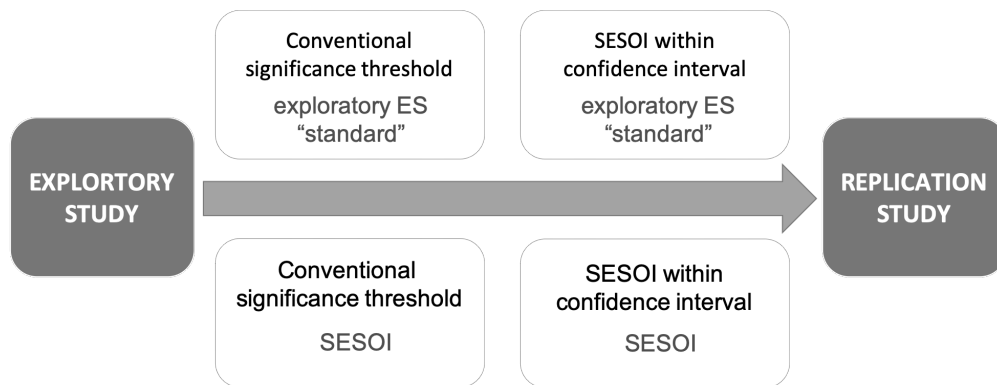


Figure 1: A preclinical research trajectory from the exploratory stage to a within-lab replication. The four panels along the arrow display four possible combinations of decision criteria and approaches to calculate the sample size for replication that can be employed throughout the trajectory.

### 2.1.1 Empirical effect size distributions

Simulations were based on empirical effect size distributions from the recently published literature [22, 23]. 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 and cover vastly different amounts of data [the last is not clear, what does amount of data mean].

The distribution of effect sizes extracted from Szucs & Ioannidis (2017) [22] 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 (Figure 2a). 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) [23] 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 (Figure 2b). 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]

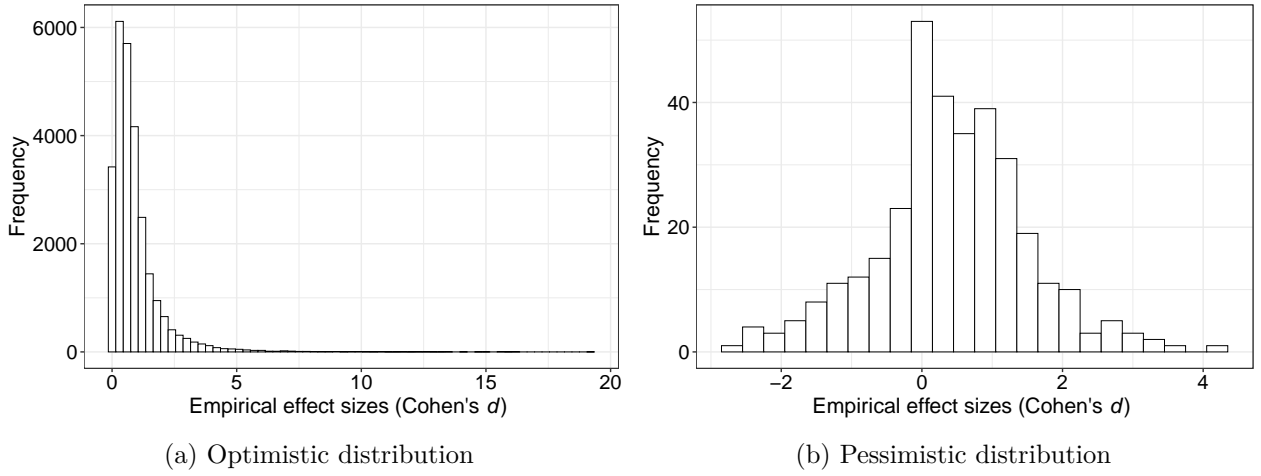


Figure 2: Empirical effect size distributions. Note that in (a) 13 values  $> 20$  were removed in order to display the distribution.

### 2.1.2 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 [13]. [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.

### 2.1.3 Decision criteria to proceed to replication

The first decision criterion employs 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  $t$ -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 [24]. 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. [Perhaps we need to enrich Figure 1 with more detail, we should brainstorm on a whiteboard for this]

#### 2.1.4 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,  $\alpha$  was set to .05. In order to calculate the sample size for replication given power and  $\alpha$ , 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).

### 2.1.5 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 ( $\alpha = .05$ ).

## 2.2 Trajectories

We tested a set of four possible combinations of decision criteria and approaches to calculate the sample size for a replication study that can be employed along the trajectory (Figure 1).

In the results section, we compare only the two trajectories that are the extremes of the spectrum of trajectories we have modeled: The first trajectory uses the conventional significance threshold as decision criterion and employs a standard sample size calculation using initial effect size estimate. The second trajectory applies a SESOI as decision criterion to move from the exploratory to the replication stage, and to calculate the sample size for the replication. In the following, we refer to the first trajectory as T1 and to the second as T2 [I think we need a good name for the trajectories as this will help readers keep track]. We have stored data, results, and figures of all four trajectories in an online repository (insert URL here).

T1 constitutes our reference, as we consider it to be closest to current practice. T1 and T2 are compared 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.

## 2.3 Outcome variables

### 2.3.1 Percentage of experiments proceeding to replication stage

This outcome serves as an indication of how strict or lenient the filters are that we used to select effect sizes that we deem relevant for further examination (i.e. to confirm them in a within-lab replication study).

### 2.3.2 Number of animals needed in the replication

[something missing here?]

### 2.3.3 Positive predictive value across trajectory

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 [25]. To calculate the positive predictive value, one (ideally) needs the pre-study odds (prevalence), as well as the sensitivity and specificity of the test (measurement). Given that we drew effect sizes from empirical effect size distributions, we know the pre-study odds of an effect of 0.5 and 1.0. Drawing from empirical effect size distributions, enables the direct comparison between pre-study odds and the positive predictive value across the trajectory. If, as outlined in the introduction, throughout the preclinical research trajectory evidence for an initial claim is strengthened, we would observe an increased positive predictive value compared to pre-study odds. [it will be important to have supplementary material for low SESOI as well.]

### 2.3.4 False positive rate across trajectory

The false positive rate (FPR) is the proportion of tests that detected a positive result when there was no true effect present. In our simulation, a false positive corresponds to a significant finding given an underlying true effect smaller than 0.5 or 1.0, respectively. [We need to motivate this. In a standard approach there is no SESOI. So every effect that is significant will be considered meaningful. The truth is, however, that studies powered for

a certain effect in the standard paradigm, so every effect below this threshold should be considered a false positive. This is a point to discuss!]

### 2.3.5 False negative rate across trajectory

The false negative rate (FNR) is the proportion of tests that missed a true positive relationship. In our simulation, this corresponds to non-significant results given that the underlying true effect was greater or equal to 0.5 or 1.0, respectively. [or again the effect the study powered for]

### 2.3.6 Effect size precision

text

## 2.4 Robustness checks

As a robustness check, we ran additional simulations using seven and 15 EUs in each group of the exploratory study, respectively. We also used additional effect sizes as SESOI (0.3 and 0.7). To compare the SESOI decision criterion at the first stage to a more lenient criterion based on statistical significance, we also simulated trajectories using  $\alpha = .1$  as significance threshold. We have stored data, results, and figures of the robustness checks in an online repository (insert URL here).

## 3 Results

### 3.1 Percentage of experiments proceeding to replication stage

#### 3.1.1 Conventional significance threshold

For this decision criterion, we extracted the  $p$ -value from the two-sided two-sample  $t$ -test we conducted at the exploratory stage, and compared it to the significance level we chose as a cut off ( $\alpha = .05$ ). The decision to proceed to replication was solely based on the  $p$ -value. Effect size estimates were not considered in this step. In case of our “optimistic” scenario

based on the empirical effect size distribution reported by Szucs & Ioannidis (2017) [22], 38.71 percent of experiments met the criterion  $p \leq .05$ . In our “pessimistic” scenario based on the empirical effect size distribution reported by Carneiro et al. (2018)[23], 32.02 percent of experiments had a  $p$ -value  $\leq .05$ . We removed effect sizes that were smaller than zero, as we reasoned that only experiments showing effects in the direction treatment > control would be further investigated. Thus the 38.71 and 32.02 percent, respectively, do include only effect sizes larger than zero. A closer look at the effect sizes of the experiments that proceeded to replication reveals that the conventional significance threshold is a conservative filter. There aren’t many effect sizes around zero, but only larger effect sizes (in both directions before removal of negative effect sizes). #figure? This has consequences for the sample size calculation as well as the PPV. [This sounds already a bit like disussion]

### 3.1.2 SESOI within the 95 percent CI of the exploratory effect size estimate

In another approach, after conducting a two-sided two-sample  $t$ -test at the exploratory stage, we estimated the effect size and 95 percent 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 the replication stage. 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 replication 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 replication shifted and included less extreme values, and more values closer to zero. #figure?

## 3.2 Number of animals needed in the replication

In trajectory T1, sample sizes were calculated using the standard approach. We extracted the effect size estimate from the exploratory study and used it to calculate the sample size for the replication study. This resulted in a mean number of 7.2 (SD = 3.67) animals in the “optimistic”, and 7.04 (SD = 3.47) in the “pessimistic” scenario. In trajectory T2, 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 3). Note that the sample sizes reported are the number of animals needed in each group (control and intervention).

[give some description of what we see in the Figure, speaking captions!]

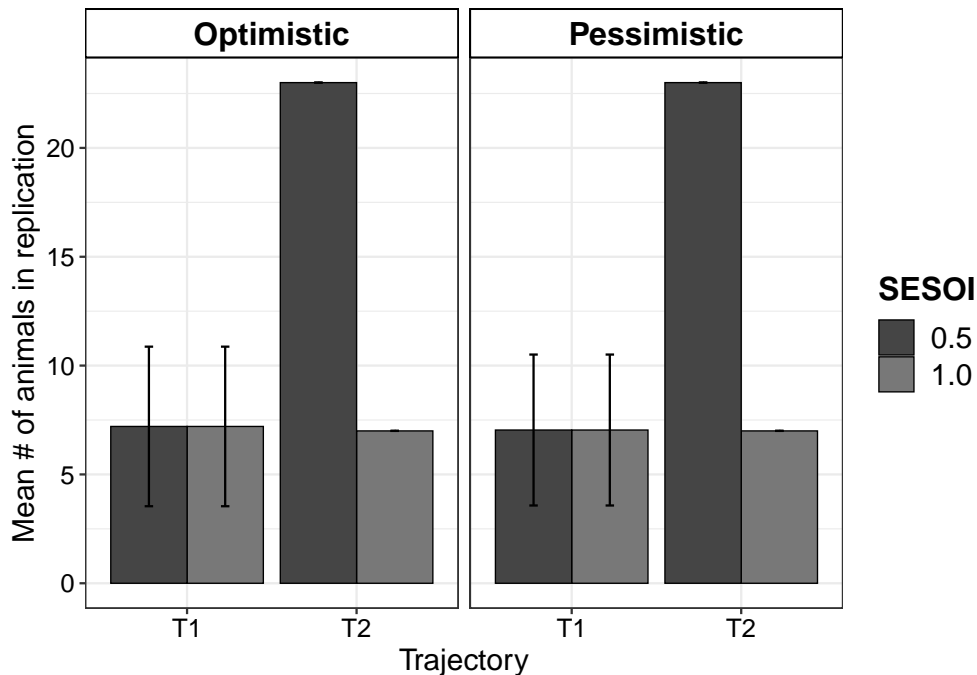


Figure 3: Number of animals needed in replication. Error bars represent standard deviations. Note that in case of trajectories using a SESOI to calculate sample size for replication the number of animals is fixed.

### 3.3 Positive predictive value across trajectory

In our study, the pre-study odds were determined by the empirical effect size distributions and varied with the SESOI. 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 trajectory T1, the PPV drops below pre-study odds in both scenarios (Figure 4). After the within-lab replication study, 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 trajectory T2, employing a SESOI at both stages along the

decision-making process elevates the PPV above pre-study odds. Given our “optimistic” empirical distribution, the PPV is 0.73 and 0.44 for SESOI of 0.5 and 1.0, respectively. In the “pessimistic” scenario, the PPV is 0.53 and 0.33

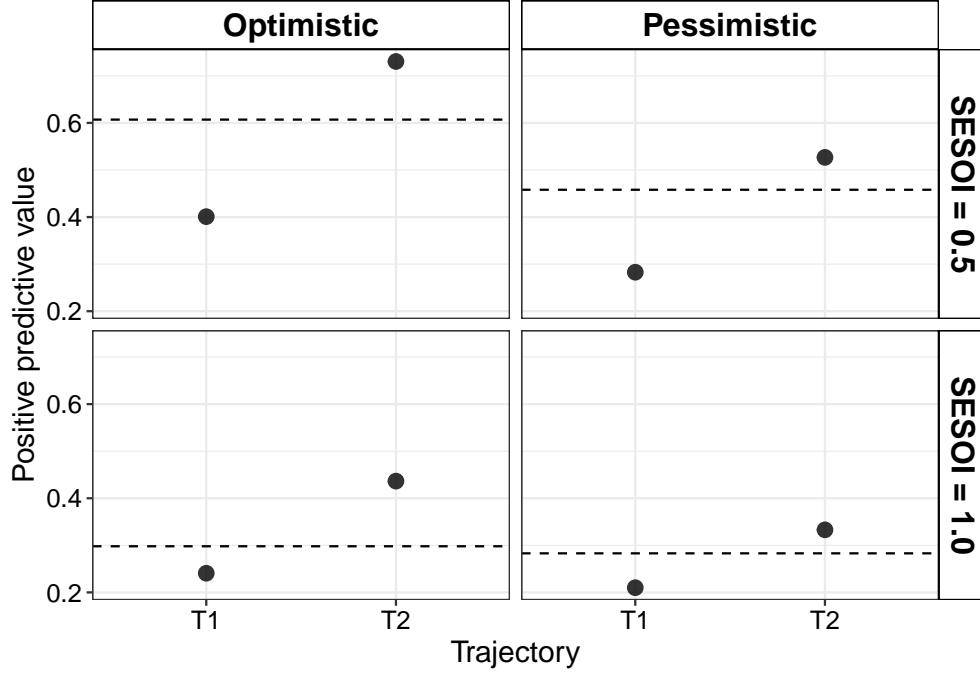


Figure 4: Positive predictive value across trajectory. Dashed lines indicate pre-study odds based on empirical effect size distributions.

### 3.4 False positive rate across trajectory

Across trajectory T1, 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 trajectory T2, 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.

### 3.5 False negative rate across trajectory

Across trajectory T1, 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

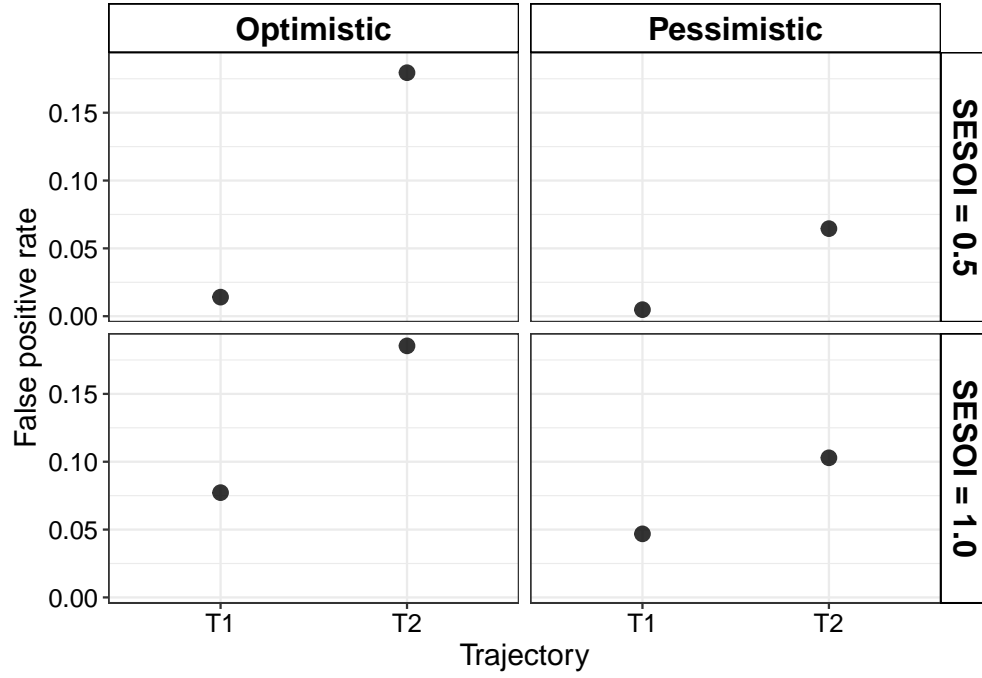


Figure 5: False positive rate across trajectory.

0.21 for SESOI of 0.5 and 1.0, respectively. Across trajectory T2, 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

### 3.6 Effect size precision

text

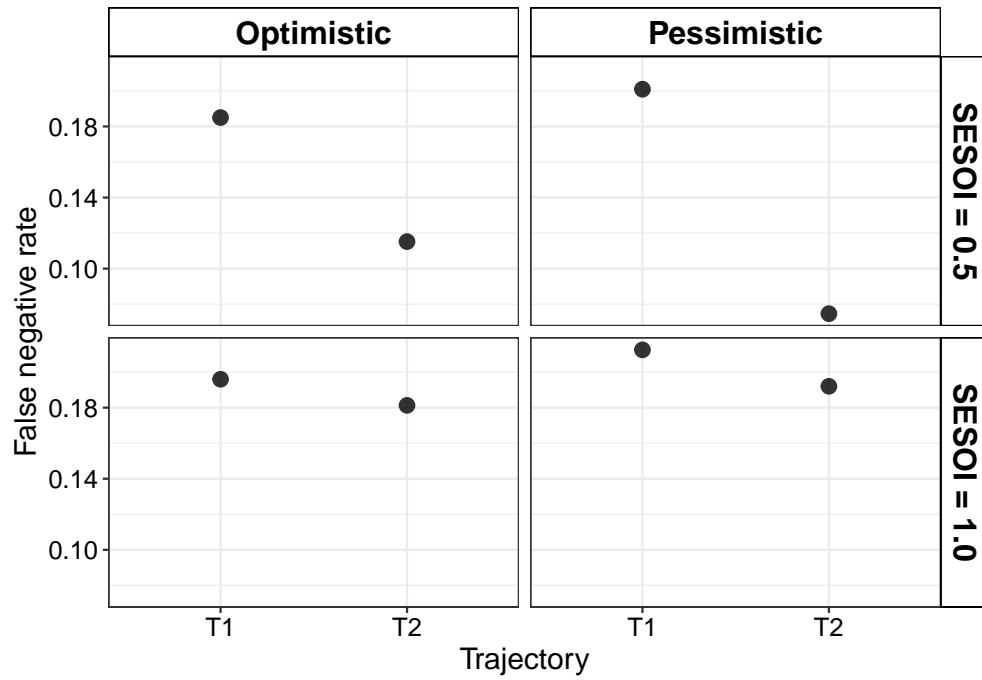
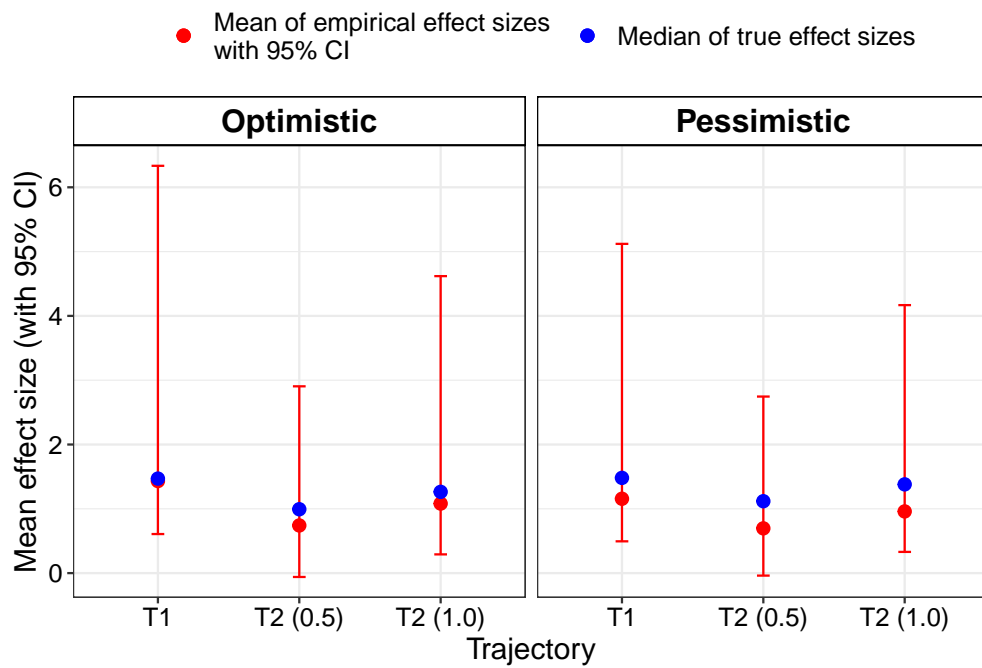


Figure 6: False negative rate across trajectory.



## 4 Discussion

text



## 5 References

1. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS biology*. 2014;12:e1001863.
2. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012;490:187–91.
3. Thomas DW, Burns J, Audette J, Carroll A, Dow-Hygelund C, Hay M. Clinical development success rates 2006–2015. *BIO Industry Analysis*. 2016;1:16.
4. Dirnagl U. Thomas willis lecture: Is translational stroke research broken, and if so, how can we fix it? *Stroke*. 2016;47:2148–53.
5. Mullane K, Williams M. Preclinical models of alzheimer’s disease: Relevance and translational validity. *Current protocols in pharmacology*. 2019;84:e57.
6. Perrin S. Preclinical research: Make mouse studies work. *Nature News*. 2014;507:423.
7. Petrov D, Mansfield C, Moussy A, Hermine O. ALS clinical trials review: 20 years of failure. Are we any closer to registering a new treatment? *Frontiers in aging neuroscience*. 2017;9:68.
8. Van Den Berg LH, Sorenson E, Gronseth G, Macklin EA, Andrews J, Baloh RH, et al. Revised airleie house consensus guidelines for design and implementation of als clinical trials. *Neurology*. 2019;92:e1610–23.
9. Freedman LP, Inglese J. The increasing urgency for standards in basic biologic research. *Cancer research*. 2014;74:4024–9.
10. Sena ES, Van Der Worp HB, Bath PM, Howells DW, Macleod MR. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS biology*. 2010;8:e1000344.
11. Vogt L, Reichlin TS, Nathues C, Würbel H. Authorization of animal experiments is based on confidence rather than evidence of scientific rigor. *PLoS biology*. 2016;14:e2000598.
12. Macleod MR, Van Der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of nxy-059 in experimental focal cerebral ischaemia is confounded

by study quality. *Stroke*. 2008;39:2824–9.

13. Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. *Nature Reviews Neurology*. 2014;10:37.

14. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*. 2013;14:365.

15. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature*. 2012;483:531.

16. Prinz F, Schlange T, Asadullah K. Believe it or not: How much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*. 2011;10:712.

17. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society open science*. 2014;1:140216.

18. Gelman A, Carlin J. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*. 2014;9:641–51.

19. Nosek BA, Errington TM. What is replication? *PLoS Biology*. 2020;18:e3000691.

20. Russell WMS, Burch RL, Hume CW. The principles of humane experimental technique. 1959;238.

21. Strech D, Dirnagl U. 3Rs missing: Animal research without scientific value is unethical. *BMJ Open Science*. 2019;3:e000048.

22. Szucs D, Ioannidis JP. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology*. 2017;15:e2000797.

23. Carneiro CF, Moulin TC, Macleod MR, Amaral OB. Effect size and statistical power in the rodent fear conditioning literature—a systematic review. *PloS one*. 2018;13:e0196258.

24. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*. 2018;1:259–69.

25. Ioannidis JP. Why most published research findings are false. *PLoS medicine*. 2005;2:e124.