PLOS BIOLOGY

META-RESEARCH ARTICLE

Reproducibility of preclinical animal research improves with heterogeneity of study samples

Bernhard Voelkl[1], Lucile Vogt[1], Emily S. Sena[2], Hanno Würbel[1]*

1 Division of Animal Welfare, VPH Institute, Vetsuisse Faculty, University of Bern, Bern, Switzerland, **2** Centre for Clinical Brain Sciences, Chancellors Building, University of Edinburgh, Edinburgh, United Kingdom

* hanno.wuerbel@vetsuisse.unibe.ch

OPEN ACCESS

Citation: Voelkl B, Vogt L, Sena ES, Würbel H (2018) Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biol 16(2): e2003693. https://doi.org/10.1371/journal.pbio.2003693

Academic Editor: Eric-Jan Wagenmakers, University of Amsterdam, Netherlands

Received: July 20, 2017

Accepted: January 19, 2018

Published: February 22, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: European Research Council ERC (grant number 322576). Swiss Food Safety and Veterinary Office FSVO (grant number 2.13.01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Single-laboratory studies conducted under highly standardized conditions are the gold standard in preclinical animal research. Using simulations based on 440 preclinical studies across 13 different interventions in animal models of stroke, myocardial infarction, and breast cancer, we compared the accuracy of effect size estimates between single-laboratory and multi-laboratory study designs. Single-laboratory studies generally failed to predict effect size accurately, and larger sample sizes rendered effect size estimates even less accurate. By contrast, multi-laboratory designs including as few as 2 to 4 laboratories increased coverage probability by up to 42 percentage points without a need for larger sample sizes. These findings demonstrate that within-study standardization is a major cause of poor reproducibility. More representative study samples are required to improve the external validity and reproducibility of preclinical animal research and to prevent wasting animals and resources for inconclusive research.

Author summary

Preclinical animal research is mostly based on studies conducted in a single laboratory and under highly standardized conditions. This entails the risk that the study results may only be valid under the specific conditions of the test laboratory, which may explain the poor reproducibility of preclinical animal research. To test this hypothesis, we used simulations based on 440 preclinical studies across 13 different interventions in animal models of stroke, myocardial infarction, and breast cancer and compared the reproducibility of results between single-laboratory and multi-laboratory studies. To simulate multi-laboratory studies, we combined data from multiple studies, as if several collaborating laboratories had conducted them in parallel. We found that single-laboratory studies produced large variation between study results. By contrast, multi-laboratory studies including as few as 2 to 4 laboratories produced much more consistent results, thereby increasing reproducibility without a need for larger sample sizes. Our findings demonstrate that excessive standardization is a source of poor reproducibility because it ignores biologically meaningful variation. We conclude that multi-laboratory studies—and potentially other ways of creating more heterogeneous study samples—provide an effective means of

Abbreviations: BMSC, bone marrow stem cell; CAMARADES, Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies; CI_{95}, 95% confidence interval; CSC, cardiac stem cell; DC, derived cell; DOR, diagnostic odds ratio; EMEA, European Agency for the Evaluation of Medicinal Products; FDA, Food and Drug Administration; FNR, false negative rate; FPR, false positive rate; G × E, gene by environment; IL1-RA, interleukin 1 receptor antagonist; MSC, mesenchymal stem cell; pc, coverage probability; RLM, Random Lab Model; T × L, treatment by laboratory; TMZ, temozolomide; tPA, tissue plasminogen activator.

improving the reproducibility of study results, which is crucial to prevent wasting animals and resources for inconclusive research.

Introduction

Reproducibility of results from preclinical animal research is alarmingly low, and various threats to reproducibility have been proposed, including a lack of scientific rigor, low statistical power, analytical flexibility, and publication bias [1–8]. All of these biases undermine the scientific validity of findings published in the scientific literature; however, empirical evidence demonstrating a causal link between any of these aspects and poor reproducibility in preclinical research is critically lacking. Moreover, an important aspect that has been almost completely overlooked so far is the rigorous standardization of animal experiments. Importantly, while all other sources of poor reproducibility mentioned above represent violations of good laboratory practice, standardization is considered good laboratory practice. Therefore, both genetic standardization (animals) and environmental standardization (housing and husbandry) are explicitly recommended by laboratory animal science textbooks [9] and are taught in laboratory animal science courses as a means to guarantee both precision and reproducibility. However, standardization renders study populations more homogenous and the results more specific to the specific standardized study conditions. Therefore, contrary to the common belief that standardization guarantees reproducibility (e.g., [9]), both theoretical [10–12] and empirical [13–17] evidence indicate that rigorous standardization may generate spurious results that are idiosyncratic to the specific standardized conditions under which they were obtained, thereby causing poor reproducibility. This is because the response of an animal to an experimental treatment (e.g., a drug) often depends on the phenotypic state of the animal, which is a product of the genotype and the environmental conditions. Therefore, phenotypic plasticity caused by gene-by-environment (G × E) interactions determines the range of variation (reaction norm) of an animal's response [18]. Instead of incorporating such natural biological variation in the experimental design, laboratory animal scientists consider this variation as a nuisance, which they aim to eliminate through rigorous standardization of both genotype and environmental conditions [9]. However, because laboratories differ in many environmental factors that affect the animals' phenotype (e.g., noise, odors, microbiota, or personnel [13,19]), animals will always differ between laboratories due to G × E interactions, and the variation of phenotypes between laboratories is generally much larger than the variation within laboratories. This implies that whenever a study is replicated in a different laboratory, a distinct sample of phenotypes will be tested. Therefore, instead of indicating that a study was biased or underpowered, a failure to reproduce its results might rather indicate that the replication study was testing animals of a different phenotype [12,16]. Nevertheless, rigorously standardized single-laboratory studies continue to be the gold standard approach to animal research from basic exploratory research to late-phase preclinical testing.

A landmark study that brought this problem to the attention of the scientific community for the first time was a multi-laboratory study by Crabbe and colleagues [13] investigating the confounding effects of the laboratory environment and G × E interactions on behavioral strain differences in mice. Despite rigorous standardization of housing conditions and study protocols across 3 laboratories, systematic differences were found between laboratories, as well as significant interactions between genotype and laboratory. The most direct way to account for such between-laboratory variation is the use of multi-laboratory study designs. Such study designs are common in medical research, especially for Phase III clinical trials [20], and

increasingly also in psychological research [21,22]. While clinical multicenter studies are often motivated by the need to recruit large samples, their potential for detecting confounding effects has been recognized by the research community [23–25]. However, in preclinical animal research, the confounding effect of the laboratory is likely to be much stronger because laboratory standards of housing and care strongly affect the animals' phenotype. Nevertheless, multi-laboratory studies are still very uncommon in preclinical animal research, despite recent initiatives [26,27] promoting their implementation. The aim of this study is, therefore, to assess how the heterogenization of study samples through multi-laboratory study designs affects the outcome of preclinical animal studies, with the hypothesis that it improves the accuracy and reproducibility of the results.

Results

To investigate how multi-laboratory designs alter the outcome and reproducibility of preclinical animal studies, we simulated single-laboratory and multi-laboratory studies based on published data of preclinical research obtained through the Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES) database [28,29]. In a first step, we selected 50 independent studies on the effect of therapeutic hypothermia on infarct volume in rodent models of stroke. In a second step, we replicated the same analysis with 12 further interventions in animal models of stroke, myocardial infarction, and breast cancer. For the sake of clarity, and to reflect the progression of this study, we will first present the analysis of the hypothermia data in full detail, followed by a summary of the analysis of the 12 replicate data sets.

A random-effect meta-analysis of the 50 studies on hypothermia yielded an estimated mean reduction of infarct volume by hypothermia of 47.8% (95% confidence interval [CI_{95}] = 40.6%–55.0%). For the simulation of single-laboratory versus multi-laboratory studies, we took this estimate as our estimate of the "true" effect. The existence of such an effect is corroborated by the efficacy of hypothermia in clinical settings [30,31]. This conjecture allowed us to compare the performance of different study designs by assessing how often and how accurately the simulated studies predicted that effect. Specifically, we compared effect size estimates and inferential statistics of single-laboratory studies to multi-laboratory studies including 2, 3, or 4 randomly selected laboratories, using the same sample size for all designs (Fig 1).

Given typical sample sizes in early preclinical animal research, we first simulated studies with a sample size of 12 animals per treatment group ($N = 24$). By randomly selecting 1 study and sampling 12 values from a Normal distribution with parameters as reported for the control group, and likewise sampling another 12 values with parameters as reported for the treatment group, we calculated an effect size estimate (mean difference) and a corresponding CI_{95} (Fig 1). Repeating this procedure 10^5 times, we found that, of such simulated single-laboratory studies, the CI_{95} captured the true effect size (i.e., the summary effect size of the meta-analysis) in only 47.9% of the cases (coverage probability [pc] = 0.48), and inferential tests failed to find a significant effect in 17.6% of the cases (false negative rate [FNR] = 0.18). Therefore, although the studies were sufficiently powered (>0.8) to detect a treatment effect, single-laboratory studies failed to predict the true effect size accurately in more than half of the cases.

To simulate multi-laboratory designs, 2, 3, or 4 different studies were randomly drawn from the pool of 50 studies, and proportionate numbers of sample values for both control and treatment group were generated to run these multi-laboratory studies with the same overall sample size as the single-laboratory studies (Fig 1). For the 2-laboratory design, pc increased to 0.73, for the 3-lab design to 0.83, and for the 4-laboratory design to 0.87, while the FNR decreased to 0.14, 0.13, and 0.13, respectively. The increase in pc with increasing numbers of

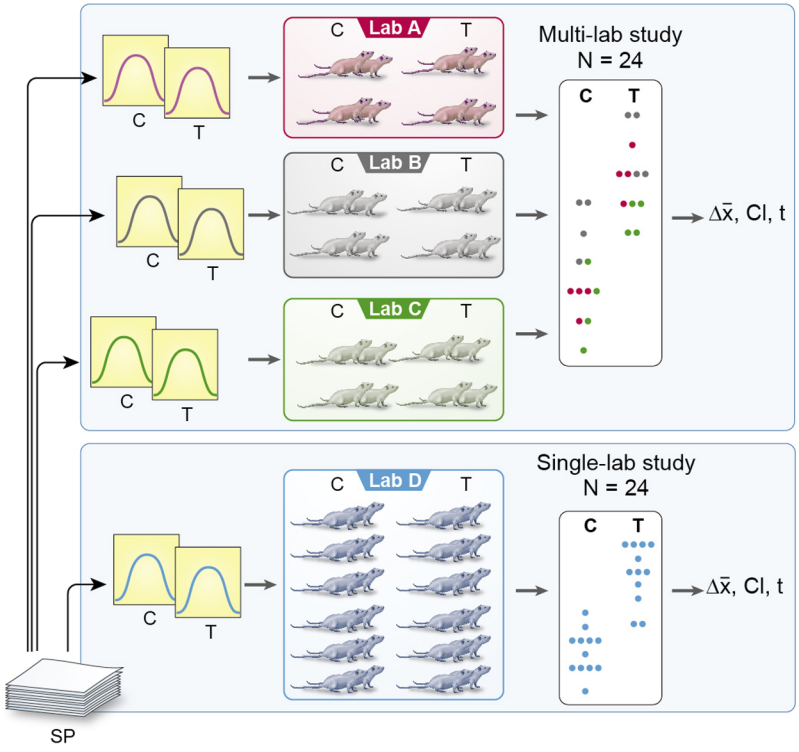

Fig 1. Sampling scheme for simulated single-lab and multi-lab studies. For a single-lab study, 1 original study is randomly selected from the study pool, and response values for control and treatment groups are generated by sampling from a Normal distribution with parameters as reported in the original study. For the multi-lab study, several original studies are selected, and values are sampled proportionate from the corresponding distributions. C, control group; SP, study pool; T, treatment group.

https://doi.org/10.1371/journal.pbio.2003693.g001

laboratories is a result of increased accuracy and reduced variation between effect size estimates. These findings are illustrated in Fig 2A, showing exemplary forest plots based on 15 randomly selected simulations for each study design. As illustrated by the first panel, effect size estimates of single-laboratory studies varied substantially, ranging from detrimental effects of hypothermia on infarct volume (effect size <0) to the complete abolition of infarct through hypothermia (effect size ≈ 1). By contrast, multi-laboratory studies including 4 laboratories produced effect size estimates very close to the true effect. The decrease of between-study variation in effect size estimates with increasing number of laboratories per study is illustrated by the width of the summary confidence interval (shaded area), which reflects the reproducibility of the results of the sampled studies.

We repeated this analysis with a smaller ($N = 12$) and a larger ($N = 48$) overall sample size to cover a range of sample sizes commonly encountered in in-vivo research. This range would comprise 7,339 (84%) of the 8,746 preclinical studies in the CAMARADES database. For $N = 12$, we only investigated the 1-, 2-, and 3-laboratory conditions but not the 4-laboratory condition because 12 animals cannot be distributed evenly over 4 laboratories and 2 experimental conditions. For all 3 sample sizes, we found an increase in pc with increasing number of participating laboratories (Fig 2B). Plotting pc against the mean width of the CI_{95} (Fig 2C) shows that the increase in pc was associated with an increase in the width of the CI_{95} estimates, yet the trade-off was reduced with increasing sample size (indicated by the steeper slopes for larger sample sizes in Fig 2C). In line with this, increasing the number of participating laboratories affected the FNR, depending on sample size. Whereas for larger sample sizes ($N = 24$

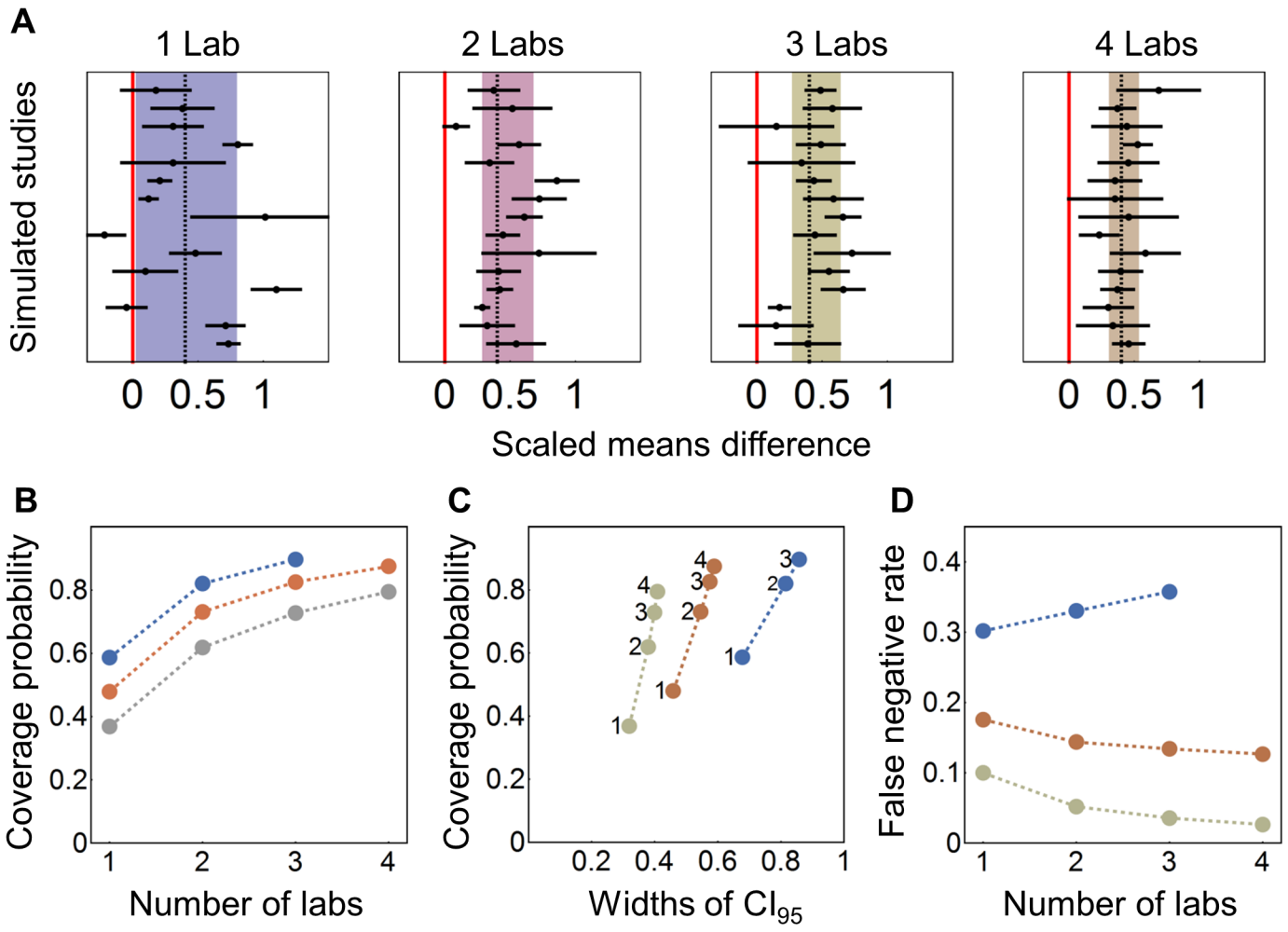

Fig 2. Results of resampling from studies on hypothermia in rodent models of stroke. (A) Forest plot of 15 randomly selected simulated studies for the 1-, 2-, 3-, and 4-lab scenario and $N = 24$; dashed line: estimated true effect; shaded area: 95% CI for the effect size estimate based on the sampled studies. The red line indicates a null effect (effect size of 0). (B) pc plotted against the number of participating laboratories for $N = 12$ (blue), $N = 24$ (orange), and $N = 48$ (grey). (C) pc plotted against the average width of the 95% CI. (D) False negative rate plotted against number of laboratories. pc, coverage probability.

https://doi.org/10.1371/journal.pbio.2003693.g002

and $N = 48$) the FNR decreased with increasing number of laboratories, this trend was reversed for $N = 12$, with the FNR increasing from 0.30 for 1 laboratory to 0.36 for 3 participating laboratories (Fig 2D). Fig 2D suggests that a divide exists somewhere near FNR of 0.2: when sample sizes were large enough for the single-laboratory design to achieve an FNR of 0.2 (reflecting statistical power of 0.8), multi-laboratory designs reduced the FNR further. In contrast, when statistical power of the single-laboratory design was below 0.8, multi-laboratory designs can lead to an increase of the FNR.

To determine whether these findings generalize across experimental treatments, we replicated this simulation study based on data for a further 12 interventions in animal models of stroke, myocardial infarction, and breast cancer ($N = 20$–58 studies per intervention; Table A in S1 Text). In all cases, we found an increase in pc with increasing number of participating laboratories (Fig 3). We also replicated the finding that the FNR generally decreases in multi-laboratory designs when statistical power is high but may increase when statistical power is

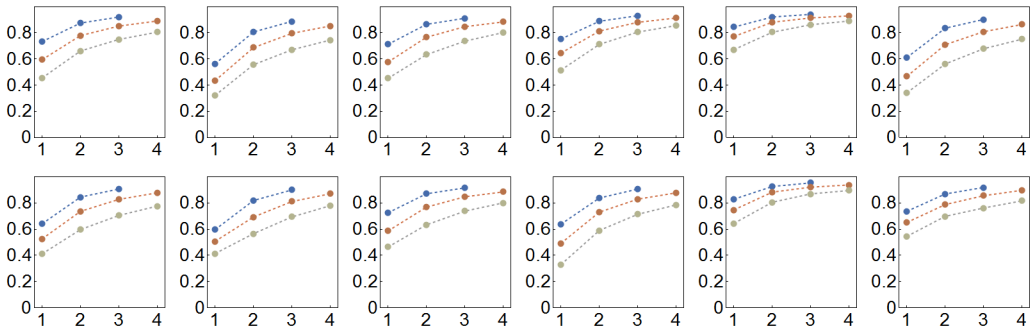

Fig 3. Coverage probability plotted against the number of participating laboratories for $N = 12$ (blue), $N = 24$ (orange), and $N = 48$ (grey) for simulated studies for 12 additional intervention studies of mouse models of stroke, myocardial infarction, and breast cancer. First row: tPA, trastuzumab, FK506, rosiglitazone 2, IL-1RA, cardiosphere DC; second row: estradiol, human MSC, MK-80, TMZ, c-kit CSC, rat BMSC (see Table A in S1 Text for details). BMSC, bone marrow stem cell; CSC, cardiac stem cell; DC, derived cell; IL1-RA, interleukin 1 receptor antagonist; MSC, mesenchymal stem cell; TMZ, temozolomide; tPA, tissue plasminogen activator.

https://doi.org/10.1371/journal.pbio.2003693.g003

low, though the exact level of statistical power above which FNR decreases in multi-laboratory designs may vary (Fig D in S1 Text).

Because it is common practice to interpret effect sizes conditional on statistical significance (for a critique of this, see, e.g., [3,32]), we calculated the proportion of studies reporting a "statistically significant" and "accurate" effect size estimate with a CI covering the true effect but not 0, p_{sa} (see Fig 4A for definition), which can be regarded as a measure of external validity in an ideal world without publication bias. As shown in Fig 4B, the external validity in terms of the proportion of statistically significant and accurate effect size estimates (p_{sa}) increased substantially in almost all cases (Fig 4B). Increasing the number of participating laboratories introduced heterogeneity and increased the total variance. In the absence of such effects, multi-laboratory designs would not substantially alter effect size estimates and statistical inference. However, heterogeneity among laboratories was large in all 13 data sets (median $I^2 = 85\%$,

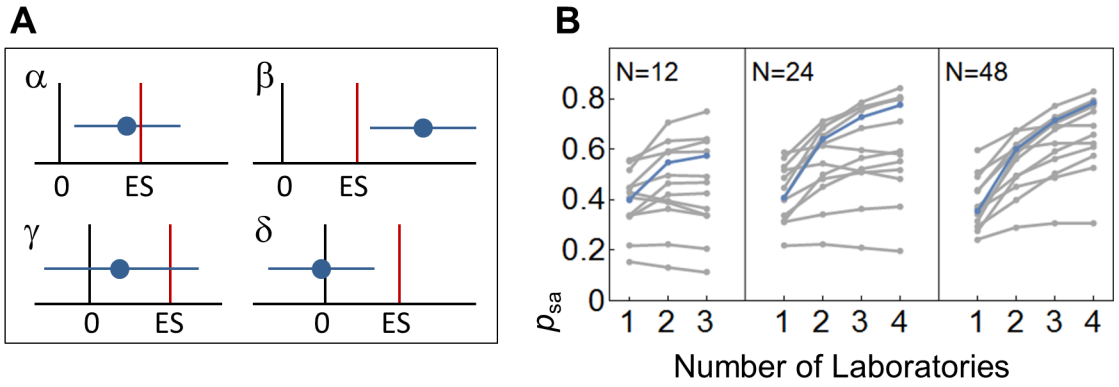

Fig 4. Proportion of studies reporting significant and accurate effects. (A) Schematic of study outcomes. A study reporting both ES estimates and inferential significant statements can lead to 1 of 4 outcomes. (α) The reported CI for the ES estimate (horizontal blue line) includes the true ES, and the CI is not including 0, suggesting the existence of an effect; (β) the CI covers neither 0 nor the true ES, suggesting the existence of an effect, though its magnitude is either over- or underestimated; (γ) the CI covers the true effect but also 0—in this case "no significant" effect would be reported, and the ES estimate would be ignored or treated as nonrelevant (which is often the case in underpowered studies); (δ) the CI includes 0 but not the true ES, leading again to a "nonsignificant" result. Based on this, we can calculate the ratio of studies accurately estimating the true ES as $p_{sa} = \alpha / (\alpha + \beta + \gamma + \delta)$. (B) p_{sa} based on 10^5 simulated samples for the hypothermia treatment of stroke (blue) and 12 further interventions (grey) for total sample sizes of $N = 12$, 24, and 48 subjects and $k = 1$–4 laboratories. ES, effect size.

https://doi.org/10.1371/journal.pbio.2003693.g004

Table 1. Definitions of key terms used in this manuscript.

Key term	Definition
Reproducibility	The similarity of outcomes between replicate studies. This can be measured, e.g., by the CI_{95} of the mean effect size estimates of a sample of replicate studies (depicted by the shaded area in Fig 2A).
FNR	False negative rate: proportion of true positives that yield negative test outcomes. FNR = false negative ÷ (true positive + false negative).
FPR	False positive rate: proportion of true negatives that yield positive test outcomes. FPR = false positive ÷ (false positive + true negative).
DOR	Diagnostic odds ratio: ratio of the odds of the test being positive in the case of a true positive relative to the odds of the test being positive in the case of a true negative. DOR = (true positive ÷ false positive) ÷ (false negative ÷ true negative).
pc	Coverage probability: the probability with which the CI_{95} of an effect size estimate includes the true effect size.
p_{sa}	The proportion of studies reporting both a significant effect for $\alpha = 0.05$ and a CI_{95} for the effect size estimate that includes the true effect.
I^2	I^2 is a descriptive statistic of the ratio of excess dispersion to total dispersion. $I^2 = (Q - df \div Q) \times 100\%$, where Q is the weighted sum of squares of study effect sizes and df gives the degrees of freedom.

Abbreviations: CI_{95}, 95% confidence interval; DOR, diagnostic odds ratio; FNR, false negative rate; FPR, false positive rate; pc, coverage probability.

https://doi.org/10.1371/journal.pbio.2003693.t001

range: 42%–97%, where I^2 is the ratio of excess dispersion to total dispersion; Table 1, Table B in S1 Text). In fact, taking a reaction norm perspective on animal traits, such environment-dependent differences and the resulting interactions are expected to be ubiquitous [12,16].

Discussion

Using simulated sampling, we compared the outcomes of single- and multi-laboratory studies, using the same overall number of animals, in terms of their accuracy of effect size estimates (pc) and FNR. For these simulations, we chose to use a large sample of published data from preclinical studies to guarantee that the results reflect real-life conditions. We found that pc increased substantially with the number of participating laboratories, without causing a need for larger sample sizes. This demonstrates that using more representative study samples through multi-laboratory designs improves the external validity and reproducibility of preclinical animal research.

Although higher pc and greater external validity come at the cost of higher uncertainty (i.e., wider CIs), this simply reflects the true uncertainty that exists when certain sources of variation are either unknown or unavoidable, which is usually the case in animal research. Of course, we cannot exclude some bias among the study samples used for our simulation approach (due to, e.g., lack of scientific rigor, publication bias). Both lack of scientific rigor [33] and publication bias [34] have been found to inflate summary effect sizes in meta-analyses. However, although this remains to be examined further, there is currently no evidence to suggest that accounting for such risks of bias would reduce the variation among replicate studies, thereby invalidating our findings. Rather, our results suggest that eliminating these and other risks of bias (e.g., low statistical power, analytical flexibility) is not sufficient to guarantee reproducibility; the results will remain idiosyncratic to the specific laboratory conditions unless these conditions are varied. Importantly, we see that increasing sample size to increase statistical power does not help but makes things even worse: it produces results that are more precise (smaller CIs) but less accurate (decreased pc) and therefore less reproducible. Relying

on more representative study samples to improve the accuracy of effect size estimates may therefore be a critical step on the way out of the current reproducibility crisis.

Taken together, our results indicate that multi-laboratory designs—and possibly other means of systematic heterogenization of study samples—will increase the accuracy of results and decrease inference errors, as long as the studies are sufficiently powered. As a consequence of this, results will gain external validity and therefore be more likely to be reproducible. Importantly, these improvements require neither many participating laboratories nor larger sample sizes. In fact, the greatest improvement in pc was observed between single-laboratory studies and studies involving 2 laboratories. As a rule of thumb, we suggest that multi-laboratory designs can improve inference and accuracy of effect size estimates, whenever sample size is large enough to achieve statistical power of at least 0.8 for a 1-way ANOVA design (i.e., a single-lab study). This suggestion is based on the finding that the trade-off between increased pc and increased uncertainty (the width of the CIs) with increasing numbers of laboratories may result in an increased FNR, which may override the positive effect of increased pc when sample size is too small.

The effects that we show here are consistent with findings reported by IntHout and colleagues [35], who compared inference errors of a single highly powered study to those of several low-powered studies, combined in a random-effects meta-analysis. These authors showed that even low levels of heterogeneity can lead to increased false positive rates (FPRs) of single-laboratory studies, while meta-analyses based on even just 2 randomly selected studies lead to notably reduced FPRs. Comparing the effect of meta-analyses comprising 2 or 3 studies, IntHout and colleagues found that the largest reduction in the FPR was observed when moving from the interpretation of 1 to 2 studies, while meta-analyses with 3 studies performed very similarly to those with only 2 studies [35]. This, too, is in line with our findings that the largest increase in pc is found when contrasting single-laboratory studies with a multi-laboratory study involving 2 participating laboratories. Furthermore, it mirrors recommendations issued by the Food and Drug Administration (FDA) [36] and the European Agency for the Evaluation of Medicinal Products (EMEA) [37] to replicate studies at least once ($N = 2$).

Besides known differences between the studies included in our analysis, such as the species or strain of animals (i.e., genotype) or reported differences in animal husbandry and experimental procedures, sources of variation included also many unknown and unknowable differences, such as the influence of the experimenter [38,39] or the microbiome [40], as well as subtle differences in visual, olfactory, and auditory stimulation. All those factors might affect treatment effects. Multi-laboratory designs are ideal to account for all of these sources of between-laboratory variation and should therefore replace standardized single-laboratory studies as the gold standard for late-phase preclinical trials [27].

However, logistic limitations may render multi-laboratory studies unsuitable for earlier, more basic types of research. One approach that was recently proposed is to statistically account for between-laboratory variation in single-laboratory studies by including a Treatment by Laboratory (T × L) interaction term as a random factor in the analysis [16]. This "Random Lab Model" (RLM) approach generates an adjusted yardstick against which treatment effects are tested in single-laboratory studies. A recent analysis of multi-laboratory data sets indicated that T × L adjustment can reduce spurious results and improve reproducibility considerably without losing much statistical power [16]. Compared with simply lowering the *p*-value of statistical significance across the board to, e.g., 0.005 as proposed by others [41,42], T × L adjustment is more specific because it takes the true heterogeneity among different laboratories into account. However, the RLM approach depends on reliable estimates of T × L interaction, which for most animal studies are not readily available. Whether the strength of

this interaction can at least be roughly estimated for specific research fields as proposed by Kafkafi and colleagues [16] remains to be tested empirically.

Because multi-laboratory studies are logistically demanding and may not be appropriate for more basic or exploratory studies, and because statistical approaches may be worrisome because of questionable assumptions, an alternative approach would be to systematically heterogenize experimental conditions, thereby mimicking multi-laboratory studies within single-laboratory studies [15]. For example, Karp and colleagues [17] found considerable phenotypic variation between different batches of knockout mice tested successively in the same laboratory. Therefore, batch heterogenization might be a useful starting point for within-lab heterogenization. A proof-of-concept study demonstrated that heterogenization based on age and housing condition of mice can improve the reproducibility of results [14], but an experimental test indicated that such simple forms of heterogenization may not be effective enough to account for the large variation between replicate studies in different laboratories [43]. In the present study, the heterogeneity among the studies used for the simulations comprised both environmental differences between laboratories and genetic differences between the different strains or—in some cases—different species used. The heterogeneity found here may, therefore, be larger than in a planned multi-laboratory study based on a specific strain of animals and harmonized environmental conditions. However, as shown here, such variation is real in preclinical research, and the evidence base generated by meta-analysis commonly includes such variation. An important future goal will therefore be to find practicable ways to mimic between-laboratory variation within single-laboratory studies using controlled, systematic variation of relevant genetic and environmental variables.

Standardization is often promoted also for ethical reasons because standardization reduces variation in experimental results, and therefore fewer animals are needed per experiment to achieve a desired level of statistical power. Using as few animals as possible for animal research is an important goal of the 3Rs principles [44]. However, our findings show that reducing animals per experiment through standardization may be short sighted because it means trading animals against the external validity and reproducibility of experimental results. Poor external validity and poor reproducibility question the benefit of the research in the harm-benefit analysis of animal experiments, which could mean that although fewer animals may be used in a standardized experiment, they may be wasted for inconclusive research [45]. As a consequence, more replicate experiments may be needed—and therefore overall more animals—to answer a given research question conclusively, which is clearly at odds with the 3Rs principles.

Materials and methods

Data acquisition and simulated sampling

Parameter estimates for the simulations were extracted from the CAMARADES [27,46] database, based on a list of a priori inclusion and exclusion criteria (Fig A in S1 Text, Table A in S1 Text). All included studies were of a 1-way ANOVA design, reporting mean estimates for a control group and a treatment group along with standard deviations, but they differed in several aspects of study protocol, including species or strain of animals, experimental procedure, and outcome assessment. We therefore scaled the reported parameter values for each study by dividing them by the mean estimate for the control group of that study. In order to simulate a single-laboratory study, we randomly selected 1 study from the study pool and sampled 6, 12, or 24 values from a Normal distribution with according parameter values for the control group and another 6, 12, or 24 values from a Normal distribution with according parameter values for the treatment group. For multi-laboratory studies with k laboratories, we randomly selected k studies from the study pool and sampled $1/k$ of values from the distributions of each

respective study. For each simulated study, we calculated the mean difference as effect size estimate and performed a 1-way ANOVA for the 1-laboratory case and a fixed-effect 2-way ANOVA with treatment and laboratory as main effects and $\alpha = 0.05$ for inference for the multi-laboratory setting. An extended discussion of alternative approaches for analyzing data of multi-lab studies (pooled *t* tests, mixed-effect linear models) is given below. The ANOVA outcome allowed us to estimate the FNR—i.e., the proportion of cases in which the F-ratio test of the ANOVA did not indicate a significant difference between groups. To assess the FPR (i.e., the proportion of cases in which the F-ratio test of the ANOVA did indicate a significant difference between groups, even though there was none), we ran a second set of simulations in which again we randomly selected original studies, but for which parameter values for both control and treatment group were drawn from the same Normal distribution with mean and standard deviation being set to the mean of the reported values for treatment and control group. The FPR of the 2-way ANOVA stayed relatively close to 0.05 under all conditions (Fig B in S1 Text, Fig C in S1 Text), corroborating the suitability of the test. As a consequence, changes in the diagnostic odds ratio (DOR; Fig B in S1 Text, Fig C in S1 Text) were mainly driven by the FNR. Simulations were first run in R 3.2.2. by LV and independently replicated by BV using Mathematica 10.1. (Wolfram Research, www.wolfram.com; see S1 and S2 Text for pseudocode and program code). Reported numbers and figures are based on the simulations run in Mathematica. Random-effect meta-analyses on the original data sets were carried out using the R package metafor1.9–9 [47] with restricted maximum likelihood estimators.

Analysis of multi-laboratory studies

The design of multi-lab studies presented in this analysis is a 2-way ANOVA design with one factor being the treatment with 2 levels—treatment or control—and the other factor being the laboratory at which subjects were housed and tested. The interaction term was not included. In the case of a single-lab study, this simplifies to a 1-way ANOVA design. Different analysis schemes have been used in the past, and battles about the appropriate analysis have been fought elsewhere [25,48,49]. The statistical analysis of multi-laboratory studies is not the topic of this manuscript, and we deliberately abstained from discussing this issue in the main text (but see S1 Text for an extended discussion). For didactical clarity, we have chosen a fixed-effect ANOVA, though with respect to our focus, the same outcomes would be retrieved if we simply performed a *t* test on the pooled (and scaled) data—as it was sometimes done in the past [49,50, but see 32,51 for a critique]—or if we treated laboratory as a random factor in a linear mixed-effect model as it is more recently advocated [23,51–53] (Fig D in S1 Text).

Supporting information

S1 Text. Supporting information, including inclusion/exclusion criteria for data sets, data set summaries, supporting discussion of the inference method, supporting data, and annotated code.
(PDF)

S2 Text. Mathematica code for simulated sampling.
(PDF)

S1 Code. Mathematica notebook with code for simulated sampling.
(NB)

S1 Data. Data extracted from the CAMARADES database and used in this study.
CAMARADES, Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies.
(CSV)

Acknowledgments

We thank Thomas Reichlin for his contributions during the early phase of this manuscript and Malcolm Macleod, Georgia Salanti, Yoav Benjamini, Jarrod Hadfield, Raghavendra Gadagkar, and Marcel A.L.M. van Assen for constructive comments on earlier versions of this paper.

Author Contributions

Conceptualization: Bernhard Voelkl, Lucile Vogt, Emily S. Sena, Hanno Würbel.

Data curation: Bernhard Voelkl, Lucile Vogt, Emily S. Sena.

Formal analysis: Bernhard Voelkl, Lucile Vogt.

Funding acquisition: Hanno Würbel.

Investigation: Bernhard Voelkl, Lucile Vogt, Hanno Würbel.

Methodology: Bernhard Voelkl.

Resources: Emily S. Sena.

Supervision: Bernhard Voelkl, Emily S. Sena, Hanno Würbel.

Validation: Bernhard Voelkl.

Visualization: Bernhard Voelkl.

Writing – original draft: Bernhard Voelkl, Lucile Vogt, Hanno Würbel.

Writing – review & editing: Bernhard Voelkl, Hanno Würbel.

References

1. Loken E, Gelman A. Measurement error and the replication crisis. Science. 2017; 355: 584–585. https://doi.org/10.1126/science.aal3618 PMID: 28183939
2. Freedman L, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol. 2015; 13: e1002165. https://doi.org/10.1371/journal.pbio.1002165 PMID: 26057340
3. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2: 696–701.
4. Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. Meta-research: Evaluation and improvement of research methods and practices. PLoS Biol. 2015; 13: e1002264. https://doi.org/10.1371/journal.pbio.1002264 PMID: 26431313
5. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, et al. A manifesto for reproducible science. Nat Hum Behav. 2017; 1: 0021.
6. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012; 483: 531–533. https://doi.org/10.1038/483531a PMID: 22460880
7. 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. 2011; 10: 712. https://doi.org/10.1038/nrd3439-c1 PMID: 21892149
8. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med. 2015; 8: 341ps12–341ps12.
9. Beynen AC, Gärtner K, van Zutphen LFM. Standardization of animal experimentation. In: Zutphen LFM, Baumans V, Beynen AC, editors. Principles of laboratory animal science. 2nd ed. Amsterdam: Elsevier Ltd; 2003. pp. 103–110.
10. Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd; 1935.

11. Würbel H. Behaviour and the standardization fallacy. Nat Genet. 2000; 26: 263. https://doi.org/10.1038/81541 PMID: 11062457
12. Voelkl B, Würbel H. Reproducibility crisis: Are we ignoring reaction norms? Trends Pharmacol Sci. 2016; 37: 509–510. https://doi.org/10.1016/j.tips.2016.05.003 PMID: 27211784
13. Crabbe JC, Wahlsten D, Dudek BC, Sibilia M, Wagner EF. Genetics of mouse behavior: Interactions with laboratory environment. Science. 1999; 284: 1670–1672. PMID: 10356397
14. Richter SH, Garner JP, Auer C, Kunert J, Würbel H. Systematic variation improves reproducibility of animal experiments. Nat Methods. 2010; 7: 167–168. https://doi.org/10.1038/nmeth0310-167 PMID: 20195246
15. Richter SH, Garner JP, Würbel H. Environmental standardization: Cure or cause of poor reproducibility in animal experiments? Nat Methods. 2009; 6: 257–261. https://doi.org/10.1038/nmeth.1312 PMID: 19333241
16. Kafkafi N, Golani I, Jaljuli I, Morgan H, Sarig T, Würbel H, et al. Addressing reproducibility in single-laboratory phenotyping experiments. Nat Methods. 2016; 14: 462–464.
17. Karp NA, Speak AO, White JK, Adams DJ, Hrabe´ de Angelis M, Hérault Y, Mott RF. Impact of temporal variation on design and analysis of mouse knockout phenotyping studies. PLoS ONE. 2014; 9: e111239. https://doi.org/10.1371/journal.pone.0111239 PMID: 25343444
18. Schoener TW. The newest synthesis: Understanding ecological dynamics. Science. 2011; 331: 426–429.
19. Würbel H. Behavioral phenotyping enhanced—beyond (environmental) standardization. Genes Brain Behav. 2002; 1: 3–8. PMID: 12886944
20. Llovera G, Hofmann K, Roth S, Salas-Pérdomo A, Ferrer-Ferrer M, Perego M, et al. Results of a preclinical randomized controlled multicenter trial (pRCT): Anti-CD49d treatment for acute brain ischemia. Sci Transl Med. 2015; 7: 299ra121–299ra121. https://doi.org/10.1126/scitranslmed.aaa9853 PMID: 26246166
21. Wagenmakers EJ, Beek T, Dijkhoff L, Gronau QF, Acosta A, Adams RB Jr, et al. Registered replication report: Strack, Martin, and Stepper (1988). Persp Psychol Sci. 2016; 11: 917–28.
22. Hagger MS, Chatzisarantis NL, Alberts H, Anggono CO, Batailler C, Birt AR, et al. A multilab preregistered replication of the ego-depletion effect. Persp Psychol Sci. 2016; 11: 546–73.
23. Kahan BC, Morris TP. Analysis of multicentre trials with continuous outcomes: When and how should we account for centre effects? Stat Med. 2013; 32: 1136–1149. https://doi.org/10.1002/sim.5667 PMID: 23112128
24. Cornfield J. Randomization by group: A formal analysis. Am J Epidemiol. 1978; 108: 100–102. PMID: 707470
25. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: An overview. Ann Intern Med. 2001; 135: 112–123. PMID: 11453711
26. Lefer DJ, Bolli R. Development of an NIH consortium for preclinicAl AssESsment of CARdioprotective therapies (CAESAR): A paradigm shift in studies of infarct size limitation. J Cardiovasc Pharmacol Ther. 2011; 16: 332–339. https://doi.org/10.1177/1074248411414155 PMID: 21821536
27. Multi-Part website. www.multi-part.org. Accessed on 10 July 2017.
28. Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Studies (CAMARADES) website. www.camarades.info. Accessed on 30 September 2015.
29. Van Der Worp HB, Sena ES, Donnan GA, Howells DW, Macleod MR. Hypothermia in animal models of acute ischaemic stroke : A systematic review and meta-analysis. Brain. 2007; 130: 3063–3074. https://doi.org/10.1093/brain/awm083 PMID: 17478443
30. Wu T-C, Grotta JC. Hypothermia for acute ischaemic stroke. Lancet Neurol. 2013; 12: 275–284. https://doi.org/10.1016/S1474-4422(13)70013-9 PMID: 23415567
31. Reith J, Jørgensen HS, Pedersen PM, Nakamaya H, Jeppesen LL, Olsen TS, Raaschou HO. Body temperature in acute stroke: Relation to stroke severity, infarct size, mortality, and outcome. Lancet. 1996; 347: 422–425. PMID: 8618482
32. Forstmeier W, Wagenmakers E-J, Parker TH. Detecting and avoiding likely false-positive findings—a practical guide. Biol Rev. 2017; 92: 1941–1968. https://doi.org/10.1111/brv.12315 PMID: 27879038
33. Rooke ED, Vesterinen HM, Sena ES, Egan KJ, Macleod MR. Dopamine agonists in animal models of Parkinson's disease: a systematic review and meta-analysis. Parkinsonism Relat D. 2011; 17: 313–320.
34. 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 Biol. 2010; 8: e1000344. https://doi.org/10.1371/journal.pbio.1000344 PMID: 20361022

35. InHout J, Ioannidis JPA, Borm GF. Obtaining evidence by a single well-powered trial or several modestly powered trials. Stat Meth Med Res. 2012; 25: 538–552.

36. Food and Drug Administration. Guidance for Industry: Providing evidence of effectiveness for human drug and biological products. Maryland: United States Food and Drug Administration;1998.

37. Committee for Proprietary Medicinal Products. Points to consider on application with 1. Meta-analyses 2. One pivotal study. London: European Agency for the Evaluation of Medicinal Products; 2001.

38. Sorge RE, Martin LJ, Isbester KA, Sotocinal SG, Rosen S, Tuttle AH, et al. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat Methods. 2014; 11: 629–632. https://doi.org/10.1038/nmeth.2935 PMID: 24776635

39. Chesler EJ, Wilson SG, Lariviere WR, Rodriguez-Zas SL, Mogil JS. Influences of laboratory environment on behavior. Nat Neurosci. 2012; 5: 1101–1102.

40. Franklin CL, Ericsson AC. Microbiota and reproducibility of rodent models. LabAnimal. 2017; 46: 114–122.

41. Johnson VE. Revised standards for statistical evidence. Proc Natl Acad Sci USA. 2013; 110: 19313–19317. https://doi.org/10.1073/pnas.1313476110 PMID: 24218581

42. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2017; 1: 1.

43. Richter SH, Garner JP, Zipser B, Lewejohann L, Sachser N, Schindler B, et al. Effect of population heterogenization on the reproducibility of mouse behavior : A multi-laboratory study. PLoS ONE. 2011; 6: e16461.

44. Russell WMS, Burch RL. The principles of humane experimental technique. London: Methuen; 1959.

45. Würbel H. More than 3Rs: The importance of scientific validity for harm-benefit analysis of animal research. Lab Anim. (NY). 2017; 46: 164–166.

46. Sena E, van der Worp HB, Howells D, Macleod M. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci. 2007; 30: 433–439. https://doi.org/10.1016/j.tins.2007.06.009 PMID: 17765332

47. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010; 36: 1–48.

48. Cornfield J. Randomization by group: A formal analysis. Am J Epidemiol. 1978; 108: 100–102. PMID: 707470

49. Vierron E, Giraudeau B. Design effect in multicenter studies: Gain or loss of power? BMC Med Res Methodol. 2009; 9: 39. https://doi.org/10.1186/1471-2288-9-39 PMID: 19538744

50. Reitsma A, Chu R, Thorpe J, McDonald S, Thabane L, Hutton E. Accounting for center in the early external cephalic version trials: An empirical comparison of statistical methods to adjust for center in a multicenter trial with binary outcomes. Trials. 2014; 15: 377. https://doi.org/10.1186/1745-6215-15-377 PMID: 25257928

51. Raudenbush SW. Hierarchical linear models as generalizations of certain common experimental design models. In: Edwards L, editor. Applied analysis of variance in behavioral science. New York: Dekker; 1993. pp. 459–496.

52. Moerbeek M, Van Breukelen GJP, Berger MPF. A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. J Clin Epidemiol. 2003; 56: 341–350. PMID: 12767411

53. Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. BMC Med Res Methodol. 2013; 13: 58. https://doi.org/10.1186/1471-2288-13-58 PMID: 23590245

Reproducibility of pre-clinical animal research improves with heterogeneity of study samples

Bernhard Voelkl*, Lucile Vogt*, Emily S. Sena, Hanno Würbel

Supporting Text

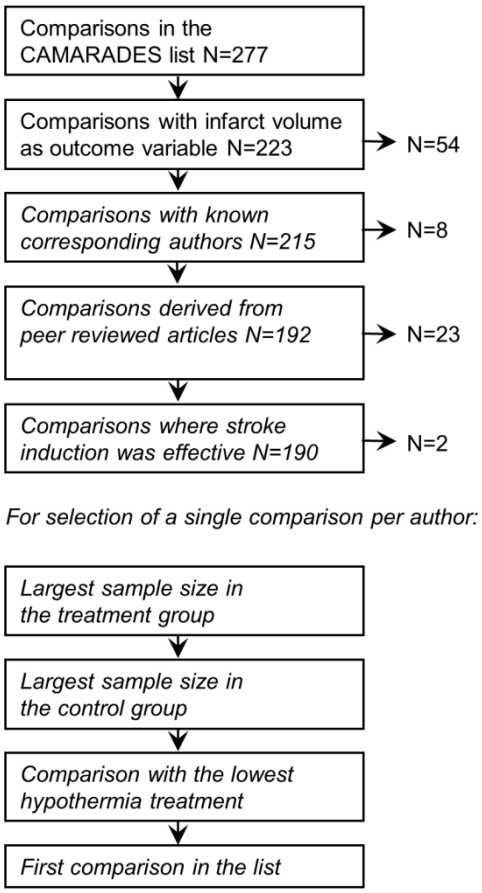

Figure A: Inclusion/exclusion criteria for the data set of hypothermia studies. The CAMARADES database contains a total of 277 comparisons between a control and a hypothermia treated group. We included only comparisons where the outcome variable was infarct volume (N=223) with a known author for correspondence (N=215), derived from a peer-reviewed article (N=192) and for which an infarct was present (i.e. induction of the lesion was successful. N=190). In order to ensure that comparisons were independent (i.e. they came from different laboratories), we classified comparisons by corresponding author and applied an additional set of inclusion criteria to obtain only one comparison per corresponding author. The 190 comparisons were published by 50 different corresponding authors. When more than one comparison was published by the same author, we selected the comparison with the largest sample size in the treatment group. If more than one study had shared maximum sample size, the study with the largest sample size in the control group was selected. When more than one study had the same sample sizes for both groups, we selected the comparison for which the lowest hypothermia temperature was used. Finally, if there were still more than one comparison by corresponding author, we selected the first comparison on the list.

	Intervention	Outcome	Species	restrict	N	Power: Median	Quartiles
0	hypothermia	Infarct volume	Rat	Yes	50	0.95	0.82-0.99
1	tPA	Infarct volume	Rat	Yes	57	0.14	0.12-0.17
2	Trastuzumab	Tumour volume ratio	Mouse	No	58	0.77	0.55-0.87
3	FK506	Infarct volume	Rat	Yes	31	0.95	0.92-0.96
4	Rosiglitazone 2	Infarct volume	Rodent	No	21	0.93	0.83-0.96
5	IL-1RA	Infarct volume	Rodent	No	37	0.42	0.34-0.57
6	Cardiosphere DC	EF (%)	Rodent	Yes	35	0.98	0.97-0.99
7	Estradiol	Infarct volume	Rat	Yes	24	0.57	0.37-0.63
8	Human MSC	Infarct volume	Rat	No	26	0.56	0.56-0.78
9	MK-801	Infarct volume	Rat	Yes	30	0.80	0.64-0.89
10	TMZ	Infarct volume	Rodent	No	26	0.96	0.84-0.99
11	c-kit CSC	EF (%)	Rodent	Yes	20	0.54	0.41-0.68
12	Rat BMSC	Infarct volume	Rat	No	25	0.33	0.24-0.36

Table A: Descriptors and selection criteria for 12 additional replicate study pools. We searched the CAMARADES database for all drugs or treatments for which we found more than 25 contrasts for the same outcome measure. Outcome measures that were collective terms for various measures or tests (spec. 'neuro-behavioural score', 'memory', and 'learning') were not considered. Contrasts where one of the following data was missing were excluded: sample size, mean outcome, and standard deviation for control and treatment group. Studies done with non-rodent species were excluded. If the majority of studies were done with a single species, only studies with the predominating species were included (indicated by "Mouse" and "Rat" in the column *Species*), otherwise all species were included ("Rodent" in column *Species*). If species were excluded this is indicated in the column 'restrict'. From all contrasts which stem from the same publication only one was selected following the following selection rules: 1. The contrast with the largest overall sample size was selected. 2. If more than one study had shared maximum sample size, the study with the larger sample size in the treatment group was selected. 3. If more than one study shared the maximum number of subjects in the treatment group, one study was selected randomly using a random number generator. Only treatments where 20 or more contrasts remained, after applying the exclusion criteria, entered into the final study pool. N: final number of studies after application of inclusion/exclusion criteria. For each study power was estimated for two-sided mean difference tests; expected effect size and standard deviation were based on mean values for each intervention, sample sizes as reported in the studies. Median power and interquartile range are given for each intervention. A full list with the CAMARADES identifiers of the included studies is given in the supplementary data set 'S1_Data.csv'.

	Intervention	N	ES	S.E.	z	*p*	CI_L	CI_U	*Q*	*p*(*Q*)	dev	τ^2	I^2	H^2
0	Hypothermia	50	0.48	0.037	12.99	<0.0001	0.41	0.55	667.5	<0.0001	5.86	0.05	91.4	11.56
1	tPA	57	0.20	0.050	3.40	<0.0001	0.10	0.29	276.6	<0.0001	54.77	0.10	84.8	6.56
2	Trastuzumab	58	0.46	0.046	10.05	<0.0001	0.37	0.55	1071.9	<0.0001	46.30	0.09	94.6	18.37
3	FK 506	31	0.39	0.038	10.31	<0.0001	0.32	0.46	186.5	<0.0001	-8.21	0.03	81.9	5.52
4	Rosiglitazone 2	21	0.47	0.039	12.20	<0.0001	0.40	0.55	84.4	<0.0001	-12.22	0.02	73.0	3.70
5	IL-1RA	37	0.34	0.029	11.91	<0.0001	0.29	0.40	114.3	<0.0001	-23.17	0.01	54.5	2.20
6	Cardiosphere DC	35	-0.52	0.036	-14.50	<0.0001	-0.59	-0.45	248.5	<0.0001	-2.44	0.03	91.9	12.43
7	Estradiol	24	0.36	0.064	5.61	<0.0001	0.23	0.48	127.4	<0.0001	14.34	0.07	84.7	6.54
8	Human MSC	26	0.25	0.037	6.56	<0.0001	0.17	0.32	414.4	<0.0001	-9.93	0.03	97.2	35.91
9	MK 801	30	0.36	0.038	9.41	<0.0001	0.28	0.43	116.5	<0.0001	-6.30	0.03	80.7	5.19
10	TMZ	26	0.51	0.044	11.72	0.0109	0.43	0.60	176.3	<0.0001	-0.64	0.03	86.1	7.21
11	c-kit CSC	20	-0.30	0.028	-10.79	<0.0001	-0.36	-0.25	36.7	0.0078	-15.89	0.01	41.7	1.71
12	Rat BMSC	25	0.19	0.048	3.94	<0.0001	0.10	0.29	955.9	<0.0001	-3.34	0.05	94.6	18.41

Table B: Results of random effects meta-analyses with REML estimators using the R-package *metafor* 1.9-9. N: number of studies in the final study pool, ES: effect size estimate, S.E.: standard error of the effect size estimate, *Q*: *Q*-statistic for homogeneity of effect sizes, τ^2: between-study variance, dev: deviance, I^2: fraction of total heterogeneity divided by total variability, H^2: fraction of total variability divided by sampling variability. Meta-analyses were performed after scaling.

Supporting Text:

For estimating the true effect for a treatment we employed fixed effect meta-analyses. Here we will briefly discuss the reasoning of this choice. Historically, clinical multi-centre studies have been analysed in different ways: by simply pooling the data from different centres or by treating centre as a fixed or random variable. Pooling the data from different laboratories and performing a t-test or F-test on the pooled data is a problematic approach, because it clearly violates the assumption of independence of the data. This problem has been discussed at length [S1-S3] and the pitfalls of pseudo-replication by ignoring statistical dependencies are extensively treated in almost all textbooks on experimental design and statistics. We do not recommend this approach, yet we have to note that comparing the diagnostic odds ratios for pooled t-tests, 2-way ANOVAs and mixed-effect models, we see—for the 13 interventions analysed in this study—only marginal differences in the performance of the test methods (Fig. D). Having agreed on accounting for statistical dependencies, which arise from testing multiple animals in the same laboratory, we face the decision whether to treat lab-membership as a fixed or random factor. With respect to this question there seems to be no overall agreement [S4-S6]. We would argue that, conceptually, laboratory is clearly a random factor, as the laboratories, which participated in the multi-lab study, are a random sample from the set of all existing—or potentially existing—laboratories. Arguably, it is not a true random sample (amongst other reasons, national laws and regulations and regional research cultures create spatial correlations), but this might be an issue that cannot be resolved. More importantly, we have to note that this random factor will only have a very limited number of levels—from 2 to 4 in our simulations and perhaps, under rare conditions, up to 5 or 6 for large multi-lab studies. For practical and organizational reasons multi-lab studies in pre-clinical research with even higher numbers of participating laboratories seem rather unrealistic and we are not aware of any attempts at achieving that. With only a few levels of the random factor, the estimation of the hyper-parameters might be rather poor [S7]. Several authors suggested rules-of-thumb for a minimum number of levels for treating a factor as random. These rules-of-thumb typically suggest between 5 and 15 levels as a minimum. This is clearly more than the number of laboratories in a multi-lab study and, following this line of reasoning, one should better treat the factor 'laboratory' as a fixed factor. On the other side, Gellman and Hill [S7] have argued that, even in the extreme case of only two levels, the mixed-effect model does not perform worse than a fixed effect model and, therefore, it might even be appropriate in cases, where there are only very few levels of the random variable. Apart from this issue, there is the question how many degrees of freedom one should attribute to the random factor. While some authors suggest that this number can at least be approximated [S8-S10], others disagree and recommend forgoing the reporting of *p*-values and inferential hypothesis testing [S11]. The question of degrees of freedom is relevant because the estimation of the ratio p_{sa} requires repeated hypothesis testing. Therefore, we didn't want to dismiss this potential problem light-heartedly. However, we must again note that for the examined range of sample sizes and number of participating laboratories, inference based on random effect models leads to very similar results as inference based on fixed effect models (Fig. D), suggesting that both approaches can be equally feasible.

Supporting References:

S1. Johnson VE. Revised standards for statistical evidence. Proc. Natl. Acad. Sci. U. S. A. 2013; 110: 19313–19317.

S2. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nat. Hum. Behav. 2017; 1: 1.

S3. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS. Generalized linear mixed models: A practical guide for ecology and evolution. Trends Ecol. Evol. 2009; 24: 127–135.

S4. Hector A, Bett T, Hautier Y, Isbell F, Kery M, Reich PB, et al. BUGS in the analysis of biodiversity experiments: Species richness and composition are of similar importance for grassland productivity. PLoS One. 2011; 6: e17434.

S5. Bennington CC, Thayne WV. Use and misuse of mixed model analysis of variance in ecological studies. Ecology. 2016; 75: 717–722.

S6. Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, et al. A brief conceptual tutorial of multilevel analysis in social epidemiology: Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. J. Epidemiol. Community Health. 2006; 60: 290–297.

S7. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press; 2007.

S8. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics, 1946; 2: 110-114.

S9. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. 1997; 53: 983–997.

S10. Welch BL. Note on some criticisms made by Sir Ronald Fisher. J. R. Stat. Soc. Ser. B. 1956; 18: 297–302.

S11. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 2015; 67: 1–48.

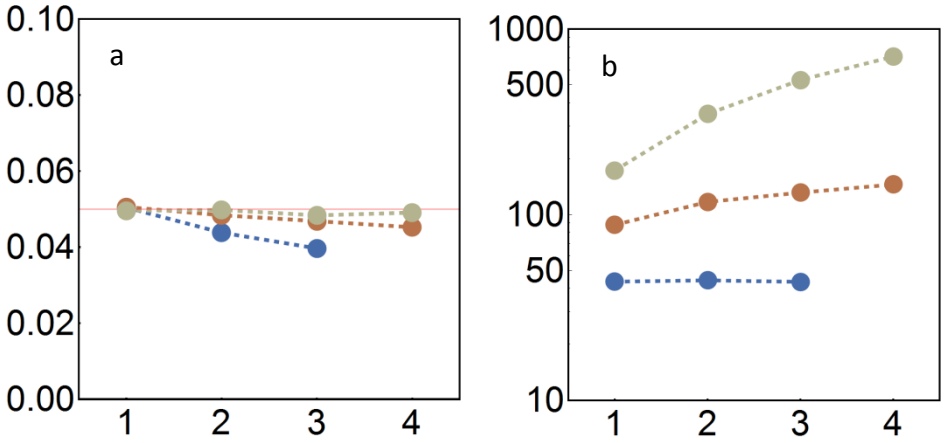

Figure B. (a) False positive rate and (b) diagnostic odds ratio (DOR) based on 10^5 simulated samples based on 50 studies on hypothermia treatment of stroke for overall sample sizes of N=12 (blue), N=24 (orange) and N=48 (grey) animal subjects and k= 1 to 4 participating laboratories. Inference was based on fixed effects 2-way ANOVA (Y= Treatment + Lab). The thin red line in panel (a) indicates the 0.05 probability threshold. The diagnostic odds ratio is the ratio of the positive likelihood ratio and the negative likelihood ratio, i.e. DOR= (true positive rate / false positive rate) / (false negative rate / true negative rate).

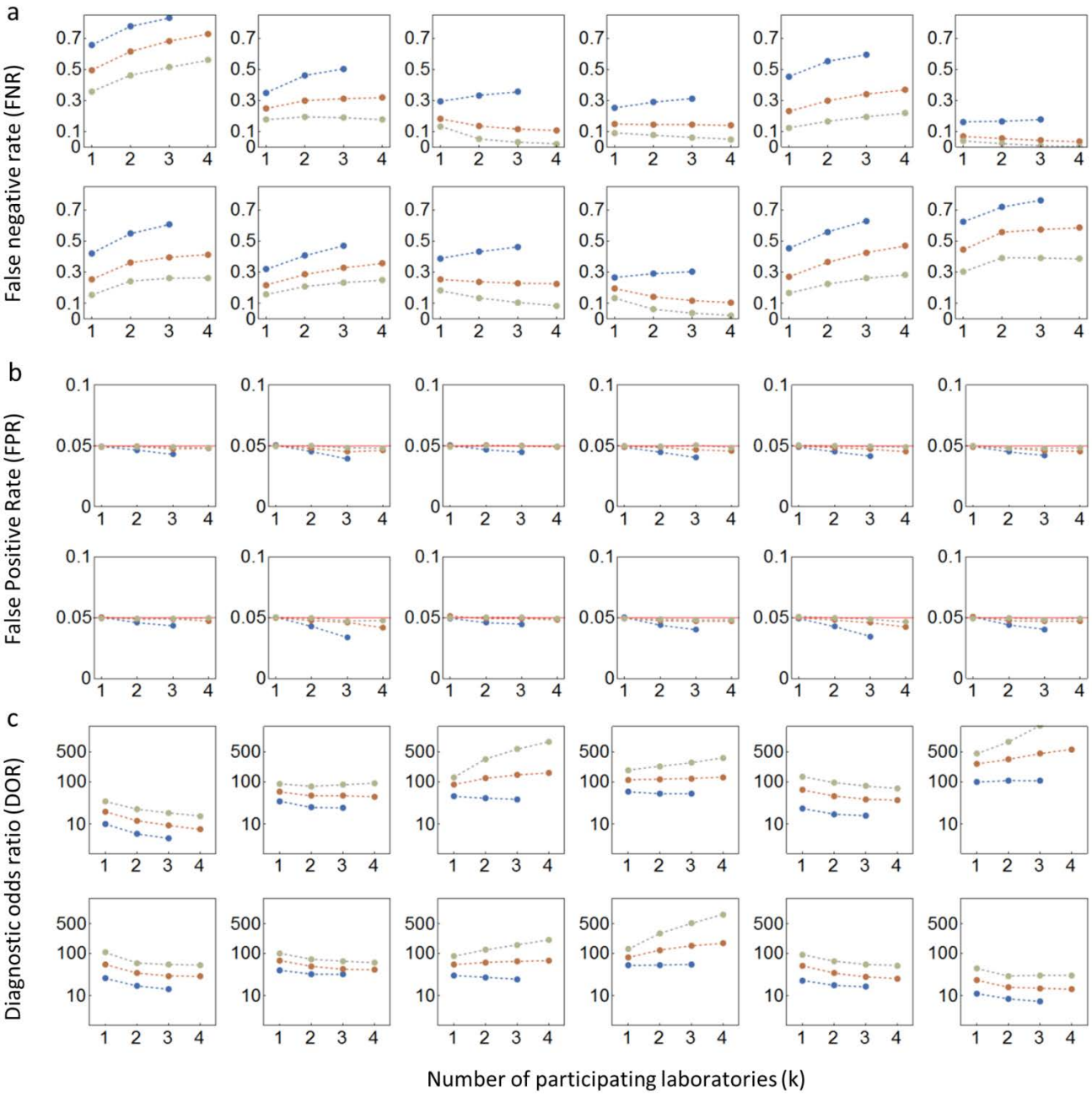

Figure C. False negative rate (a), false positive rate (b) and diagnostic odds ratio (c) for the 12 replicate data sets (from left to right: row 1: D1-D6, row 2: D7-D12), based on 10^5 simulations for overall sample sizes of N=12 (blue), N=24 (orange) and N=48 (grey) animal subjects.

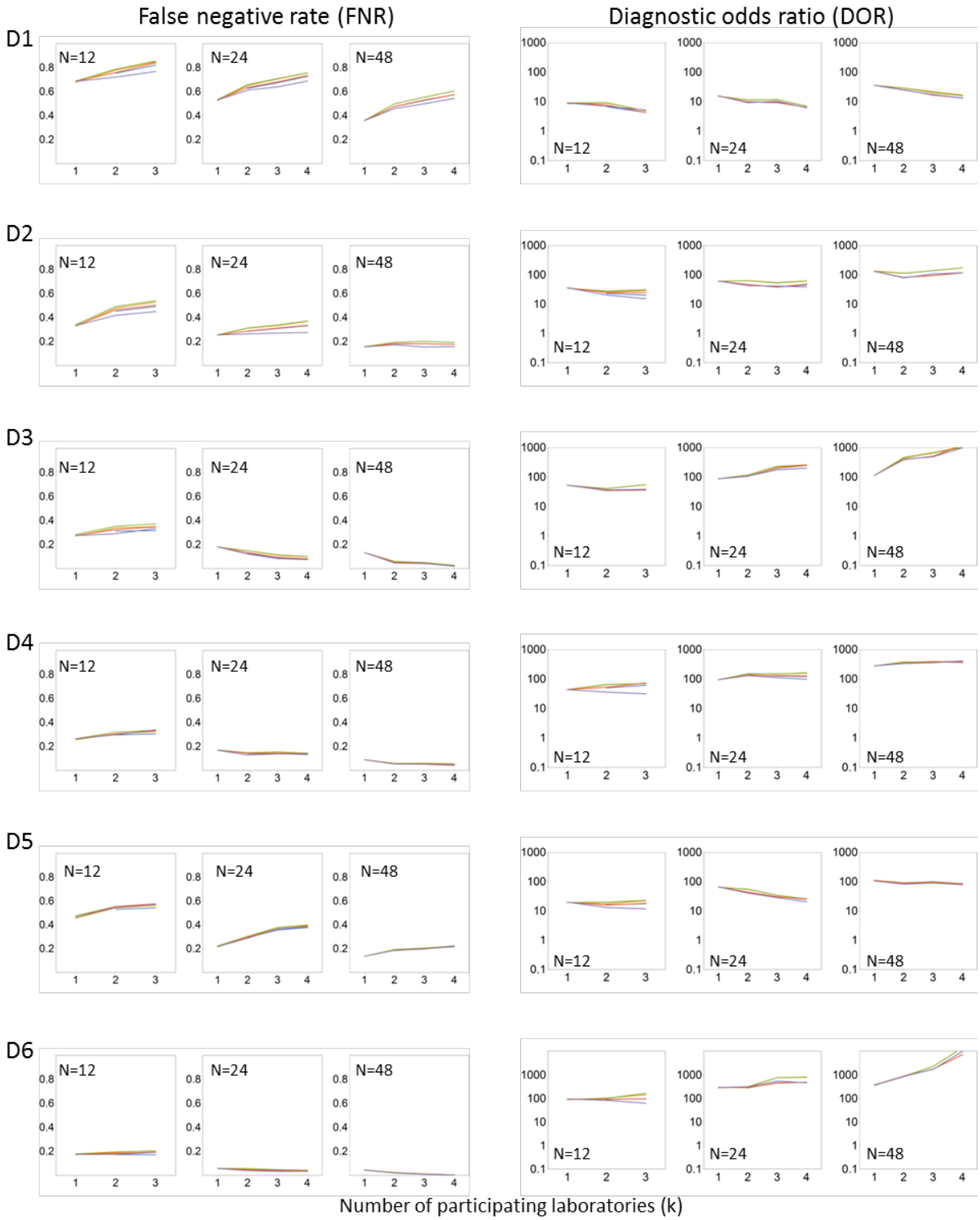

Figure D. False negative rate and diagnostic odds ratio for the 12 replicate data sets (D1-D12) with 10^3 simulations per data set. Inference based on: inclusion of zero by the parametric 95% confidence interval (yellow), t-test on pooled data (green), ANOVA with main effects treatment and laboratory only (red), ANOVA with main effects and interaction term (violet), general linear mixed model Y~treatment+(1|lab) with lab as random effect (blue). In almost all cases diagnostics based on all 5 inference techniques showed very similar behaviour, showing that the findings are not specific to the method of statistical analysis.

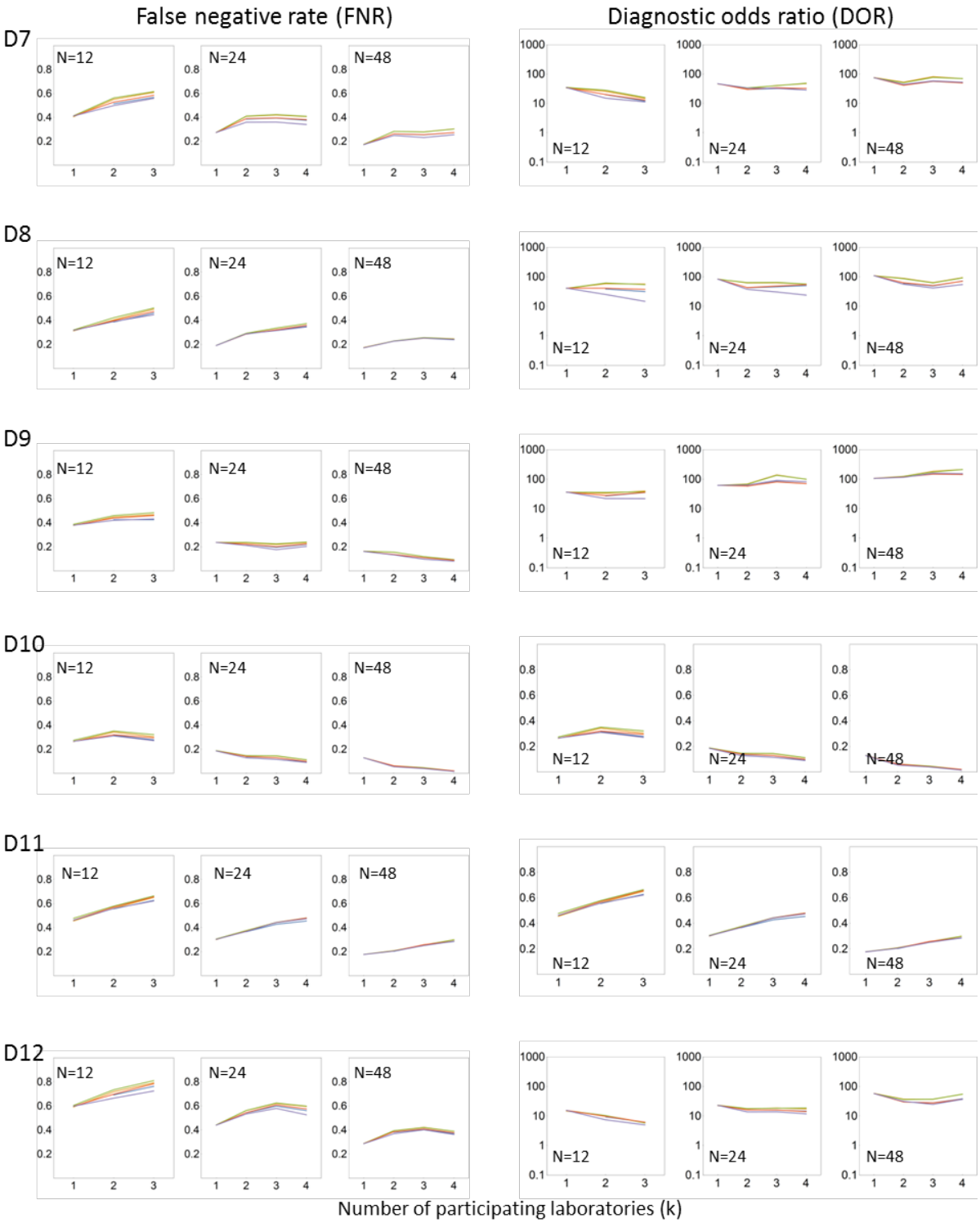

Figure D cont.

Pseudocode for simulating multi-lab studies

k = number of laboratories
n = total number of animals

REPEAT the following 100.000 times:
    create a list of *k* laboratories by sampling without replacement from the study pool
    FOR each laboratory
        control = sample *n*/(2*k*) values from a normal distribution with mean and standard deviation as reported for the control group and divide those values by the reported mean value for the control group
        treatment = sample *n*/(2*k*) values from a normal distribution with mean and standard deviation as reported for the treatment group and divide those values by the reported mean value for the control group
    ENDFOR
    perform a two-way fixed effect ANOVA on the simulated data
    pool the control values of all laboratories
    pool the treatment values of all laboratories
    calculate the means difference and 95% confidence interval for the pooled data
    test whether confidence interval includes the estimate for the means difference of the meta-analysis
ENDREPEAT

Mathematica Code for simulating multi-lab studies
(Effect size estimate, CI_{95}, and 2-way ANOVA)

```
getOneES[labslist_] := Module[
        {nplc, controls, treatments, se, meansdiff, lower, upper, ci, anovap},
         nplc = totalsamplesize/(k*2);        (* k is the number of labs and                          *)
                                              (* nplc is the number of animals per lab per condition  *)
         controls = Flatten[Table[RandomReal[NormalDistribution[data[[labslist[[i]], 4]],
                  data[[labslist[[i]],11]]], nplc]/data[[labslist[[i]], 4]], {i, k}]];
        treatments = Flatten[Table[RandomReal[NormalDistribution[data[[labslist[[i]], 7]],
                  data[[labslist[[i]], 12]]], nplc]/data[[labslist[[i]], 4]], {i,k}]];
                                              (* this generates values sampled from normal distributions  *)
                                              (* with parameters as reported in the original studies      *)
        se = Sqrt[(StandardDeviation[treatments]^2 + StandardDeviation[controls]^2)/(totalsamplesize/2)];
                                              (* se is the standard error for the mean difference        *)
        meansdiff = Mean[controls - treatments];   (* this is the mean difference                        *)
        anovap =                              (* the p-value of a fixed effect ANOVA                     *)
        If[k == 1,
                ANOVA[Transpose[{Join[Table[0, { totalsamplesize/2}], Table[1, { totalsamplesize/2}]],
                        Join[controls, treatments]}]][[1, 2, 1, 1, 5]],
                                              (* for the one-lab condition a one-way ANOVA is made       *)
                                              (* and the p-value is assigned to the variable anovap      *)
                ANOVA[Transpose[{Join[Table[0, {{ totalsamplesize/2}], Table[1, {{ totalsamplesize/2}]],
                        Flatten[Join[Table[Table[i, {nplc}], {i, k}], Table[Table[i, {nplc}], {i, k}]]],
                        Join[controls, treatments]}], {x, y}, {x, y}][[1, 2, 1, 1, 5]]
                                              (* for more than one lab a two-way ANOVA is made          *)
                                              (* and the p-value is assigned to the variable anovap      *)
        ];
        lower = meansdiff - zvalue*se;        (* this gives the lower 95% CI                            *)
        upper = meansdiff + zvalue*se;        (* this gives the upper 95% CI                            *)
        {truees >= lower && truees <= upper, lower > 0 || upper < 0, anovap, anovap < 0.05}
        (* this gives a list with: the first entry giving True if the true effect size lies within the 95%  *)
        (* confidence interval, and False otherwise, the second entry gives True if the 95% confidence      *)
        (* interval is not including zero and False otherwise, the third entry is the p-value estimate*)    *)
        (* from the ANOVA, and the fourth entry is True if the p-value of the ANOVA is less than 0.05       *)
        (* and False otherwise                                                                             *)
 ]

results = Table[getOneES[RandomSample[Range[numberofstudies], k]], {100 000}];
(*this repeats the simulation 100.000 times. The number of labs (k) and the true effect size (truees)     *)
(* and the number of studies in the data matrix (numberofstudies) must be specified before executing      *)
(* the function. The data of the original studies must be provided as matrix with the observed mean       *)
(* of the control group in column 4, the mean of the treatment group in column 7, the standard            *)
(* deviation of the control group in column 11 and the standard deviation for the treatment group in      *)
(* column 12.                                                                                             *)
```

R-code for meta analyses

```
library("metafor")

data<-read.table("dataset_meta.csv", header=TRUE, sep=';')
ncontrol<-data$Number.in.Control.Group
mcontrol<-data$Reported.Mean.in.Control.Group/data$Reported.Mean.in.Control.Group
sdcontrol<-data$Calculated.SD.in.Control.Group/data$Reported.Mean.in.Control.Group
ntreatment<-data$Number.in.Treatment.Group
mtreatment<-data$Reported.Mean.in.Treatment.Group/data$Reported.Mean.in.Control.Group
sdtreatment<-data$Calculated.SD.in.Treatment.Group/data$Reported.Mean.in.Control.Group
# This block reads in observed values from data for reported sample size of the control
# group (ncontrol), reported mean of the control group (mcontrol), reported standard deviation
# for the control group (sdcontrol), reported sample size of the treatment group (ntreatment),
# reported mean of the treatment group (mtreatment), and reported standard deviation for
# the treatment group (sdtreatment).

result.meta <- rma(m1=mcontrol, m2=mtreatment,
            sd1=sdcontrol, sd2=sdtreatment, n1= ncontrol, n2=ntreatment,
            method="REML", measure="MD")            # ”MD” indicates means difference

summary(result.meta)
```