PLOS BIOLOGY

META-RESEARCH ARTICLE

Meta-analysis of variation suggests that embracing variability improves both replicability and generalizability in preclinical research

Takuji Usui[1,2¤*], **Malcolm R. Macleod**[3], **Sarah K. McCann**[4,5], **Alistair M. Senior**[2☯*], **Shinichi Nakagawa**[1,2☯*]

1 Evolution and Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, Australia, **2** The Charles Perkins Centre and School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia, **3** Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom, **4** QUEST Center for Transforming Biomedical Research, Berlin Institute of Health (BIH), Berlin, Germany, **5** Charité—Universitätsmedizin Berlin Corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany

☯ These authors contributed equally to this work.
¤ Current address: Biodiversity Research Centre, University of British Columbia, Vancouver, Canada
* usuitakuji@gmail.com (TU); alistair.senior@sydney.edu.au (AMS); s.nakagawa@unsw.edu.au (SN)

OPEN ACCESS

Citation: Usui T, Macleod MR, McCann SK, Senior AM, Nakagawa S (2021) Meta-analysis of variation suggests that embracing variability improves both replicability and generalizability in preclinical research. PLoS Biol 19(5): e3001009. https://doi.org/10.1371/journal.pbio.3001009

Academic Editor: Isabelle Boutron, University Paris Descartes, FRANCE

Received: October 27, 2020

Accepted: May 4, 2021

Published: May 19, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pbio.3001009

Data Availability Statement: All relevant data and code are available on Figshare: https://doi.org/10.6084/m9.figshare.14527317.v4

Abstract

The replicability of research results has been a cause of increasing concern to the scientific community. The long-held belief that experimental standardization begets replicability has also been recently challenged, with the observation that the reduction of variability within studies can lead to idiosyncratic, lab-specific results that cannot be replicated. An alternative approach is to, instead, deliberately introduce heterogeneity, known as "heterogenization" of experimental design. Here, we explore a novel perspective in the heterogenization program in a meta-analysis of variability in observed phenotypic outcomes in both control and experimental animal models of ischemic stroke. First, by quantifying interindividual variability across control groups, we illustrate that the amount of heterogeneity in disease state (infarct volume) differs according to methodological approach, for example, in disease induction methods and disease models. We argue that such methods may improve replicability by creating diverse and representative distribution of baseline disease state in the reference group, against which treatment efficacy is assessed. Second, we illustrate how meta-analysis can be used to simultaneously assess efficacy and stability (i.e., mean effect and among-individual variability). We identify treatments that have efficacy and are generalizable to the population level (i.e., low interindividual variability), as well as those where there is high interindividual variability in response; for these, latter treatments translation to a clinical setting may require nuance. We argue that by embracing rather than seeking to minimize variability in phenotypic outcomes, we can motivate the shift toward heterogenization and improve both the replicability and generalizability of preclinical research.

Funding: AMS was supported by a Discovery Early Career Researcher Award from the Australian Research Council (ARC DECRA: DE180101520: https://www.arc.gov.au/grants/discovery-program/discovery-early-career-researcher-award-decra), and formally by a Coffey Fellowship from the University of Sydney. SN was supported by the Australian Research Council Discovery Grants (DP180100818 and DP200100367: https://www.arc.gov.au/grants/discovery-program/discovery-projects). 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.

Abbreviations: CAMARADES, Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies; CI, credible interval; CV, coefficient of variation; HBOT, hyperbaric oxygen therapy; lnCV, log coefficient of variation; lnCVR, log coefficient of variation ratio; lnRR, log response ratio; lnSD, log standard variation; MAR, missing at random; MLMA, multilevel meta-analysis; MLMR, multilevel meta-regression; PEESE, precision-effect estimate with standard errors; PET, precision-effect test; STAIR, Stroke Academic Industry Roundtable; tPA, tissue plasminogen activator.

Introduction

Replicability of research findings—"obtaining the same results from the conduct of an independent study whose procedures are as closely matched to the original experiment as possible," otherwise known as "Results reproducibility" [1]—is integral to scientific progress. Compelling evidence, however, suggests that non-replicability pervades basic and preclinical research [1–5]. Moreover, animal studies motivate the development of novel treatments to be tested in clinical studies, but failure to observe effects in humans which have been reported in animal studies is commonplace [6,7]. The conventional approach to preclinical experimental design has been to minimize heterogeneity in experimental conditions within studies to reduce the variability between animals in the observed outcomes [8]. Such rigorous standardization procedures have long been endorsed as the way to improve the replicability of studies by reducing within-study variability and increasing statistical power to detect treatment effects, as well as reducing the number of animals required [8,9]. This well-established notion that standardization begets replicability, however, has recently been challenged.

An inadvertent consequence of standardization is that an increase in internal validity may come at the expense of external validity [10]. By reducing within-study variability, standardization may inflate between-study variability as outcomes become idiosyncratic to the particular conditions of a study, ultimately becoming only representative of local truths [10–12]. For example, in animal studies, the interaction between an organism's genotype and its local environment (i.e., phenotypic plasticity due to gene-by-environment interactions) can result in variable and discordant outcomes across laboratories using otherwise concordant methodology [13–16]. Such inconsistent outcomes may result from distinct plastic responses of animals to seemingly irrelevant and minor, unmeasured differences in environmental conditions and experimental procedures [13–18]. Through amplifying the effects of these unmeasured variables, standardization may thus weaken, rather than strengthen, replicability in preclinical studies.

A potential counter to this "standardization fallacy" [10] then is to improve replicability by embracing, rather than minimizing, heterogeneity [10–12]. Practical solutions to enhance external validity include conducting studies across multiple laboratories to deliberately account for differences in within-lab variability [19–21], and perhaps more radically, to systematically introduce variability into experimental designs within studies [12,22,23]. Both simulation [11,14,20,21] and empirical studies [19,22,24,25] show that deliberate inclusion of more heterogeneous study samples and experimental conditions (i.e., "heterogenization") improve external validity, and hence replicability, by increasing within-study (or within-lab) variability and minimizing among-study (or among-lab) variability.

Despite the promise of heterogenization, standardization remains the conventional approach in preclinical studies [26–28]. This has been partly fuelled by Russel and Birch's [29] injunction to a "reduction in the numbers of animals used to obtain information of a given amount and precision." Consequently, within-study variability is typically treated as a biological inconvenience that is to be minimized, rather than an outcome of interest in its own right. Embracing and quantifying heterogeneity, however, may benefit preclinical science in at least 2 ways. First, through comparative analyses of the variability associated with experimental procedures, we may identify methodologies that introduce variation. As discussed above, by using methods that induce variation, one may design a deliberately heterogeneous study with greater replicability [10–12]. Second, by explicitly investigating interindividual heterogeneity in the response to drug/intervention outcomes, we may quantify the generalizability of a treatment and its translational potential. That is, a treatment with low interindividual variation in efficacy despite heterogenization is more generalizable, while a treatment with high interindividual

variation indicates the effect may be individual specific. This may be relevant in the context of personalized medicine: A treatment associated with interindividual variation in outcomes may require tailoring in its clinical use [30]. Taking these 2 points together, one could argue that an ideal trial would use a technical design that typically generated variation in disease state, which was then attenuated by a treatment of interest that might consistently (in all animals) or selectively (in some animals) improve outcome.

An illustrative case where the issues of replicability and lack of translation have been highlighted repeatedly is that of animal models of ischemic stroke [31–33]. Several systematic reviews [34,35] and meta-analyses [36–38] have questioned the propriety of experimental design and the choice of experimental procedures in stroke animal studies. The consequent recommendation for improving replicability in the field has usually been to adopt methodological procedures that minimize heterogeneity (and/or mitigate sources of bias) in phenotypic outcomes (e.g., in infarct volume or neurobehavioral outcomes) [34–38]. Furthermore, while potentially beneficial treatments have been identified in individual trials at the preclinical stage, intravenous thrombolysis remains the only regulatory-approved treatment for ischemic stroke [33,39,40]. This lack of transferable results from the preclinical to clinical stage highlights a major shortcoming for the generalizability of stroke animal models and is emblematic of translation failures generally across preclinical studies [6,7,33,34].

Using the case of rat animal models of stroke as a guiding example, we highlight how recently developed methods for the meta-analysis of variation can be used to better understand biological heterogeneity. First, through analysis of variability using the log coefficient of variation (lnCV; CV representing variance relative to the mean) in control groups, we identify methodological procedures that increase variability in outcomes. Second, we show how, through the concurrent meta-analysis of mean and variance in treatment effects using the log response ratio (lnRR; i.e., ratio of means) and log coefficient of variation ratio (lnCVR), one gains additional information about the generalizability of an intervention at the individual level. Overall, we argue that the quantification of heterogeneity in phenotypic outcomes can be exploited to improve both the replicability and translation of animal studies.

Results

Dataset

We obtained data for rat animal models of ischemic stroke from the Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES) database [41], focusing our meta-analysis on animal models that reported outcomes in infarct volume (see Materials and methods for inclusion criteria of studies). We extracted data for infarct volume from 1,318 control group cohorts from 778 studies for our analyses, investigating the effects of methodology and variability. We extracted data for the effect of treatment on infarct volume from 1,803 treatment/control group cohort pairs from 791 studies for our analyses, investigating the effects of drug treatment on interindividual variability (see Data Availability Statement section for full data and code).

Methodology and variability

To identify methodological procedures that generated variability in disease state, we first meta-analyzed variability in infarct volume for control group animals. We quantify variability as the lnCV rather than the log of standard deviation because we found that our data showed a linear log mean–variance relationship (i.e., Taylor's law, where the variance increases with an increase in the mean [42]; S1 Fig). Overall, the coefficient of variation (CV) in infarct volume across control groups was around 23.6% of the mean (lnCV = −1.444, CI = −1.546 to −1.342,

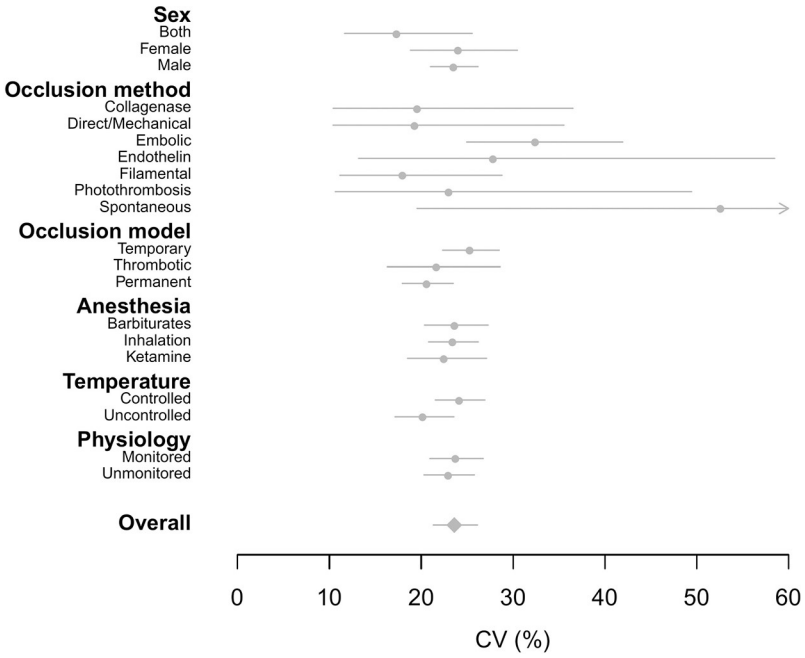

Fig 1. The effects of methodological parameters on variability (CV) in infarct volume across control groups. Mean estimates of unconditional (marginalized), group-specific coefficients of variation (%) are indicated as gray circles, while the overall estimate is indicated as a gray diamond. Moreover, 95% CIs are shown as gray lines and are asymmetric due to back-transformation of log coefficient of variation (lnCV) to the natural scale. Spontaneous occlusion generated the highest estimate of variability as indicated by the arrowhead. The overall and group-specific estimates were obtained from MLMA and MLMR models, respectively. The data underlying this figure can be found at https://doi.org/10.6084/m9.figshare.14527317.v4. CI, credible interval; CV, coefficient of variation; lnCV, log coefficient of variation; MLMA, multilevel meta-analysis; MLMR, multilevel meta-regression.

https://doi.org/10.1371/journal.pbio.3001009.g001

$\tau^2 = 0.565$; Fig 1). We found large differences in variability of infarct volume ($I^2_{total} = 93.7\%$), suggesting that sampling variance alone cannot account for differences in the reported variability across control groups (Table 1). The I^2 attributable to study was 49.6%, suggesting that methodological differences across studies explained some of this heterogeneity, although a moderate amount (42.9%) of I^2 remained unexplained (Table 1).

Table 1. Heterogeneity (I^2) estimates for analyses of methodology on variability (lnCV) and drug treatment on mean (lnRR) and variance (lnCVR) in rat infarct volume.

Model	Total	Study	Strain	Residual (within-study)
lnCV				
MLMA	93.7%	49.6%	1.3%	42.9%
MLMR	93.3%	46.3%	1.7%	45.3%
lnRR				
MLMA	95.7%	54.5%	1.7%	39.5%
MLMR	94.9%	46.3%	2.2%	46.4%
lnCVR				
MLMA	71.2%	38.8%	0.9%	31.6%
MLMR	70.3%	36.1%	1.2%	33.1%

Estimates (%) are shown for MLMAs and MLMR models.

lnCV, log coefficient of variation; lnCVR, log coefficient of variation ratio; lnRR, log response ratio; MLMA, multilevel meta-analysis; MLMR, multilevel meta-regression.

https://doi.org/10.1371/journal.pbio.3001009.t001

We detected statistically significant differences in variability of infarct volume between various methodological approaches (Fig 1; see S1 and S2 Tables for unconditional and conditional model coefficients, respectively). Among occlusion methods, models with spontaneous occlusion produced the greatest variability in infarct volume (CV = 52.5%; lnCV = −0.644, −1.633 to 0.345), while filamental occlusion had lowest variability (CV = 17.9%; lnCV = −1.720, −2.195 to −1.244). Studies using temporary models of ischemia had higher variability in infarct volume (CV = 25.2%; lnCV = −1.377, −1.500 to −1.255) compared with permanent models. Variability was slightly but significantly lower with longer time of damage assessment (lnCV = −1.404, −1.521 to −1.288) and greater median weight of the control group cohort (lnCV = −1.366, −1.486 to −1.245).

Drug treatment effects and interindividual variation

To quantify generalizability in drug treatment outcomes, we meta-analyzed the mean and the CV in infarct volume for the effects observed in control/experimental contrasts. We quantified the mean and interindividual variability as the lnRR and lnCVR, respectively. Overall, mean infarct volume in experimental groups was around 33.1%, smaller than in control groups (lnRR = −0.402, −0.461 to −0.343; Fig 2A), while the CV in experimental groups was around 32.4% higher than in control groups (lnCVR = 0.280, 0.210 to 0.351; Fig 2B). Overall, heterogeneity in lnRR was very high, while that for lnCVR was moderate, and moderate amounts of heterogeneity were partitioned into the study level for both (Table 1).

Both the mean and variability in infarct volume differed significantly across drug treatment groups (Fig 2; see S3 and S4 Tables for unconditional and conditional model coefficients, respectively). Treatment with hypothermia resulted in the largest reduction of mean infarct volume in experimental groups relative to controls (around 49.7% lower in experimental groups than controls; lnRR = −0.687, −0.775 to −0.599). However, hypothermia also had the most variable and inconsistent effect (i.e., intersubject variation) in reducing infarct volume, with the largest ratio of CV between experimental and control groups (interindividual variability around 60.0% higher in experimental groups compared with controls; lnCVR = 0.470, 0.349 to 0.591). In contrast, environmental treatments were the least effective in reducing mean infarct volume (around 7.3% greater in experimental groups than controls; lnRR = 0.071, −0.166 to 0.308). Hyperbaric oxygen therapy (HBOT) has the least variable and most consistent effect on infarct volume (variability around 45.3% less in experimental groups relative to controls; lnCVR = −0.603, −1.483 to 0.277).

Thrombolytics, which include the only regulatory-approved treatment (i.e., tissue plasminogen activator; tPA [33]), reduced mean infarct volume by around 29.6% in experimental relative to control groups (lnRR = −0.351, −0.446 to −0.256). The CV across experimental groups for thrombolytics was around 17.4% higher than control groups (lnCVR = 0.160, 0.031 to 0.289), but it is notable that this increased intersubject variability is much less than that seen with hypothermia. Through quantifying variability in drug treatment outcomes, we propose that treatments be considered generalizable if they reduced mean infarct volume and concurrently show low interindividual variability (i.e., negative lnRR and lnCVR estimates; Fig 3). Drug treatments that on average reduced infarct volume but had variable and inconsistent effects (i.e., had negative lnRR and positive lnCVR estimates; Fig 3) are ungeneralizable but might be appropriate for clinical exploitation in selected patients [30,43]. Conversely, the least successful treatments can be identified as those that consistently do not reduce mean infarct volume (i.e., positive lnRR and lnCVR estimates; Fig 3). We explored whether the sex of groups used in experiments affected lnRR or lnCVR (see Materials and methods for multilevel meta-regression [MLMR] model parameters), but differences in mean or variability of infarct

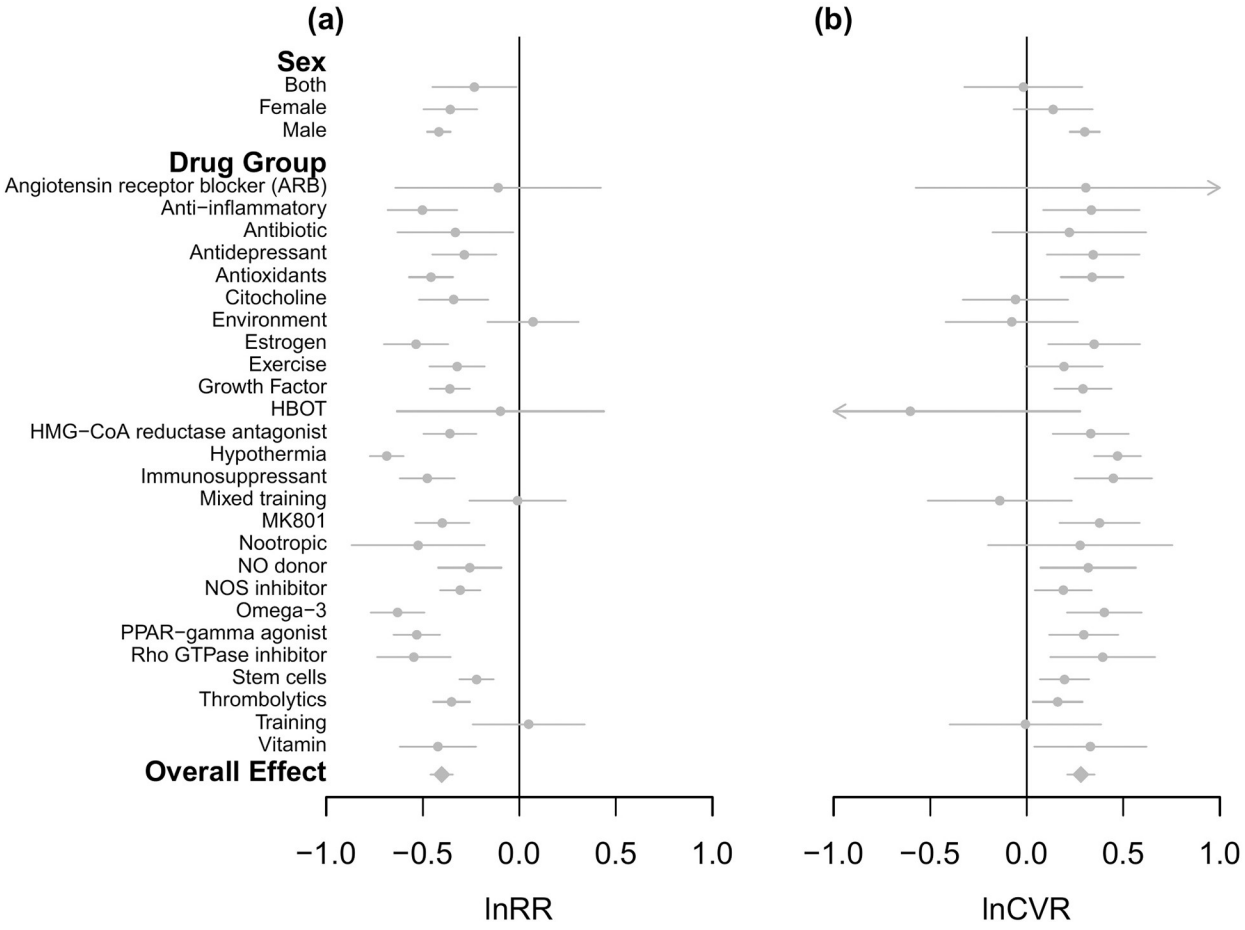

Fig 2. The effects of drug treatments on the difference in (a) mean (lnRR); and (b) variability (lnCVR) in infarct volume across control and experimental rat groups. Mean estimates of unconditional (marginalized), group-specific effects are shown as gray circles, while the overall estimate is indicated by the gray diamonds. Moreover, 95% CIs are shown as gray lines. Negative lnRR estimates indicate that mean infarct volume is smaller in experimental versus control rats. Negative lnCVR estimates show that interindividual variability in infarct volume is smaller in experimental versus control rats (e.g., HBOT indicated by left-pointing arrowhead), while positive lnCVR estimates show that variability in infarct volume is greater in experimental versus control rats (e.g., ARBs indicated by right-pointing arrowhead). The overall and group-specific estimates were obtained from MLMA and MLMR models, respectively. The data underlying this figure can be found at https://doi.org/10.6084/m9.figshare.14527317.v4. ARB, angiotensin receptor blocker; CI, credible interval; HBOT, hyperbaric oxygen therapy; HMG-CoA, β-Hydroxy β-methylglutaryl-CoA; lnCVR, log coefficient of variation ratio; lnRR, log response ratio; MLMA, multilevel meta-analysis; MLMR, multilevel meta-regression; NO, nitric oxide; NOS, nitric oxide synthase; PPAR, peroxisome proliferator–activated receptor.

https://doi.org/10.1371/journal.pbio.3001009.g002

volume did not vary significantly between female and male cohorts (see S5 and S6 Tables for contrast model estimates for sex effects).

Discussion

We propose that the current failures in replicability and translation of preclinical trials may be due, at least in part, to the way studies are designed and assessed, which is to minimize within-study variation and ignore heterogeneity in outcomes [8,9,26–28]. Here, we have illustrated the potential utility of embracing such heterogeneity, through meta-analyzing variability (relative variance or CV) in outcomes for rat animal models of stroke. First, by estimating the variability generated by different methodological designs applied to control animal groups, we have identified procedures that generate variability in disease states (Fig 1). Second, we have, for the first time, quantified both the efficacy and stability (i.e., changes in the mean and

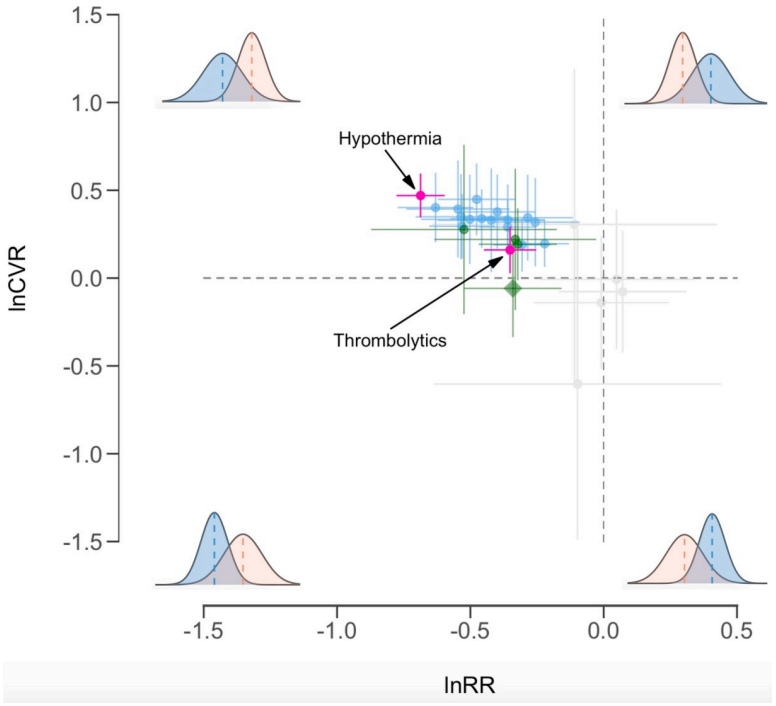

Fig 3. Categorization of treatment effects based on mean efficacy (lnRR) and interindividual variability in efficacy (lnCVR). Estimates (circles) represent unconditional (marginalized), treatment-specific means (lnRR), variability (lnCVR), and their 95% CIs (solid lines) obtained from MLMR models. Treatments that significantly reduce infarct volume (negative lnRR) without significantly affecting the variation are highlighted green, with citicoline indicated by a diamond as the only treatment to significantly reduce infarct volume and also have a negative point estimate of lnCVR. Treatments that significantly reduce infarct volume and increase interindividual variability (positive lnCVR) are highlighted blue. The effects of hypothermia (most negative and positive mean and variability estimates, respectively) and thrombolytics (which include the only regulatory-approved treatment) are highlighted in pink. Histograms show the relationship of the mean and variance in infarct volume between control (orange) and treatment (blue) groups in each quadrant of the graph. The data underlying this figure can be found at https://doi.org/10.6084/m9.figshare.14527317.v4. CI, credible interval; lnCVR, log coefficient of variation ratio; lnRR, log response ratio; MLMR, multilevel meta-regression.

https://doi.org/10.1371/journal.pbio.3001009.g003

variance, respectively) of stroke treatments applied to the experimental animal models (Figs 2 and 3), identifying potential treatments that may be generalizable versus those that require tailoring. We further discuss these results below in the context of their implications for improving the replicability (also defined as "Results reproducibility" [1]) and generalizability of preclinical studies.

Generate variability through methodology

Among stroke animal models, studies may differ in the design of a number of parameters, including the genetic composition of animals (e.g., the sex and strain of rats used [32,44]) as well as laboratory and operational environments (e.g., methods for stroke induction, the duration of ischemia, and the type of anesthesia used [37,38,45]). However, an impediment to heterogenization is that we have not previously had reliable estimates for which methodological parameters may be most successful in generating variability in phenotypic outcomes [15]. Our results therefore quantify heterogeneity and rank the experimental factors that can generate variability in disease state into animal models so that we can most efficiently capture heterogeneity in experimental design.

Our analyses of operational factors reveal that heterogeneity in outcomes may be induced by incorporating spontaneous (CV = 52.5%), embolic (CV = 32.3%), and endothelin (CV = 27.8%) methods of occlusion. Temporary models of occlusion also generate significantly more variability in disease state than permanent models (CV = 25.2% and 20.5%, respectively). Where choices permit, we suggest that these operational design considerations are a valuable approach for introducing variability into animal models, in conjunction with more familiar proposals to diversify the laboratory environment (e.g., through differences in animal housing conditions and feeding regimens [16,19]). Depending on the type and purpose of study, such operational and laboratory design considerations that increase heterogeneity in outcomes through environmental effects may be especially valuable when variability cannot be introduced through the animal's genetic composition (e.g., for studies that are interested in sex-specific [46,47] or strain-specific outcomes [44,48]).

Considering genetic factors, proposals to include more heterogeneous study samples recommend the inclusion of both sexes over just male or female animals [49–51], as well as the use of multiple strains of inbred mice and rats (or even, multiple species) [52,53]. Recent meta-analyses of variability in male and female rodents show that males may be as or more variable than females in their phenotypic response [54,55]. We also find that male (CV = 23.5%) and female (CV = 23.9%) rats generate quantitatively equal amounts of variability. Counterintuitively, however, we find that studies that used both sexes produce the most consistent outcomes (CV = 17.3%; see S1 Table for full, unconditional model coefficients). We caution that a moderate amount of the total heterogeneity remained unexplained (i.e., residual variation; Table 1). Thus, these outcomes of sex on estimates of variability may be due to confounding effects of unaccounted for differences in experimental design. We therefore emphasize the importance of considering both genetic and environmental parameters for effective heterogenization of studies [56,57].

An alternative approach to heterogenization of experimental designs within studies is to introduce variability by conducting experiments across multiple research laboratories (i.e., multi-laboratory approach) [20,24,58]. Importantly, such an approach inherently captures "unaccounted" sources of variability in experimental conditions that are difficult to systematically manipulate within a single-center study [16,19]. We argue that, especially where logistical constraints may hinder multi-laboratory approaches (e.g., for earlier, basic, and exploratory studies), introducing heterogeneity within studies may provide the most practical alternative [23]. Indeed, by meta-analyzing the variability introduced by differences in experimental methodology across studies, we can begin to find ways in which to heterogenize single studies in order to best capture the variation that exist across laboratories and studies [16,20].

Systematically introducing variability into a system comes at the cost of reduced statistical sensitivity [8,9] and necessitates larger studies [8,26,29]. While in the long-term increased replicability may reduce waste and outweigh the initial costs, these economic and ethical costs must, of course, be minimized. This can be done by identifying from a spectrum of all available methodological choices the most efficient means of introducing heterogeneity within experiments (Fig 1, S1 Table). For some methodological aspects such as operational factors, this will mean replacing current methodologies with choices that induce greater variability in baseline and control group outcomes (e.g., by changing methods of occlusion). For other design parameters that may traditionally be standardized such as genetic or lab environmental factors, this will mean deliberately incorporating these types of heterogeneity in a systematic manner, by including levels of these categorical (e.g., strain) or continuous (e.g., time of assessment) variables using a randomized block or fully factorial design (i.e., "controlled heterogenization" [59]). Regardless of the manner in which heterogeneity is incorporated, however, it is necessary to quantify the amount of variability that different experimental designs introduce, with

the aim that researchers can then make informed decisions about how to most efficiently incorporate heterogeneity into study design [14–16,20]. Identifying sources of variability through meta-analysis of variance in existing animal data as we have done here is the most practical and economic way of establishing this much needed knowledge base.

Our analysis is not the first to assess the effects of experimental methodology on variation in disease state in rodent models of stroke [37 38]. Ström and colleagues [37] investigated similar components of experimental design on variation in infarct volume in rats. There are a number of methodological differences between their analyses and ours (e.g., differences in size of dataset and use of formal meta-analytic models). Despite these differences, our quantitative results are largely concordant. Where we differ substantially is in the interpretation of what is a desirable outcome. For example, Ström and colleagues [37] concluded that intraluminal filament procedures provide optimal occlusion methods as they generate minimal variation in disease outcome and maximize statistical power. Our analyses also identify that filament methods have low variation (CV = 17.9%); however, we argue that these gains in statistical power come at the cost of reduced replicability.

We attempted to provide formal statistical support for the hypothesis that heterogeneous methods result in more repeatable treatment effects. We used a second-order meta-regression to assess whether the amount of variation (lnCV) in disease states induced by occlusion methodology predicts heterogeneity in effect sizes for drug treatments using those methods (quantified as ln$H$ [60]). As predicted, there is a negative relationship (slope = −0.876, −2.047 to 0.295; $P = 0.142$; S2 Fig), suggesting that methodologies that induce greater variability in baseline disease states are associated with more consistent treatment outcomes. We note, however, that our slope estimate is statistically nonsignificant and that our analysis was based on a small number of methodological groups ($N = 7$) with an unbalanced distribution of drugs/rat strains across those groups (see S7 Table for analysis details and full model results). Nonetheless, our results are encouraging, and we are excited to see further studies formally quantify the relationship between variability induced by methodological procedures and replicability in reported outcomes. Meta-analyses that quantify both variability in control and treatment outcomes, as done here, provide a useful approach for quantifying the relationship between methodological heterogenization and replicable outcomes.

Quantify variability to improve drug translation

Our second approach of simultaneously assessing both the mean and variation in treatment outcomes allows us to place potentially useful treatments into 2 distinct categories for further exploration: (1) beneficial and generalizable interventions, which are those that consistently reduce infarct volume across individuals; and (2) beneficial but non-generalizable interventions, which on average reduce infarct volume but result in large interindividual heterogeneity in outcomes. This latter group could even include treatments that do not necessarily reduce mean state, but have a large enough variance response to be beneficial to some [30,43,61].

Overall, we find that the stroke treatments in our dataset are usually effective, reducing infarct volume on average by 33.1% compared with controls. Out of these effective treatments, we identify 4 treatments that significantly reduced infarct volume but did not induce significant differences in the CV across experimental and control groups (green highlights in Fig 3). Nootropic treatments reduced infarct volume on average by 40.8%, while citicoline, antibiotic, and exercise treatments reduced infarct volume by around 27.5% to 28.8% compared with control groups. None of these treatments were estimated to significantly affect the CV, although estimated effects ranged from 5.7% smaller in experimental relative to controls for citicoline (highlighted with a triangle symbol in Fig 3) to 21.3% to 31.9% greater for the other

treatments. We emphasize that these treatments may potentially be more generalizable in that the outcomes of these treatments are on average favorable and are relatively consistent at the individual level [33,34].

Second, we identify a handful of effective treatments that on average reduce infarct volume, but also generate significant amounts of variability in experimental groups (blue highlights in Fig 3; see S3 Table for rank order of unconditional estimates in mean and CV across treatments). Of particular interest to note is that while thrombolytics significantly increase variability in experimental groups relative to controls, they are still relatively consistent in reducing mean infarct volume (on average reducing infarct volume by 29.6%, while the CV in experimental groups is only 17.4% greater than controls). Out of treatments that significantly reduce mean infarct volume, thrombolytics rank second in terms of its consistency in effect, with overlapping confidence intervals in their effects on the CV with those of citicoline (Fig 3).

On the other hand, hypothermia is much more effective in reducing infarct volume (on average reducing infarct volume by 49.7%) but is the least consistent in doing so, estimating the greatest CV (60.0% greater in hypothermia treated groups than concurrent controls). Interestingly, efforts to exploit hypothermia for stroke in clinical trials have so far failed to identify a patient group who might reliably benefit [62]. Other treatments that greatly reduce average infarct volume while increasing the variation include, for example, omega-3, rho GTPase inhibitors, and estrogen treatments. As such, while these treatments confer a mean beneficial effect, this effect may not be generalizable across animals. Any future translation into clinical trials would require tailoring with effort put in to predicting response at the individual level [30]. To our knowledge, such tailoring has not been attempted because a treatment with high variability (inconsistency) is less likely to be statistically significant and pass the preclinical stage (even if it does improve a disease state) [30,43,61,63]. Our study represents the first meta-analyses to quantify both the efficacy and consistency of treatment effects in animal models. We believe that this approach will forge new opportunities for improving the generalizability and translation of preclinical trials by embracing both the mean and variability in outcomes.

Conclusions

We have demonstrated how researchers can quantitatively embrace heterogeneity in phenotypic outcomes with the aim of improving both the replicability and generalizability of animal models. Prior to experimentation, researchers may design their experiments by deliberately selecting methodologies that generate variability in disease state, creating a heterogenous, but broadly representative backdrop of disease states against which treatment efficacy can be assessed [10–12]. Since the magnitude and direction of phenotypic expression and outcomes are determined by the interaction of genetic and environmental contexts within studies [14–16], both of these methodological factors require heterogenization in order to avoid context-specific and non-replicable outcomes across studies [16]. Post-experimentation, studies may further incorporate analyses that estimate the magnitude and direction of variability generated by treatments to identify potentially generalizable versus non-generalizable approaches. Recent meta-analyses of variability in phenotypic outcomes of animal models are beginning to illuminate the potential use of embracing different types of heterogeneities for improving replicability, generalizability, and translation [60–62]. We offer that comparative analyses of variability in both control and treatment groups has the potential to inform experimental design and lead to changes in both the approach and direction of follow-up studies, ultimately leading to a more successful program of replicability, drug discovery, and translation.

Materials and methods

Data collection and imputation

We identified studies of rat animal models for stroke from the CAMARADES electronic database (see S3 Fig for database query and selection). For our analysis, we only included experimental studies that reported mean infarct volume (and their associated sample size and standard deviation or standard error) in both control and experimental groups. Where necessary, we calculated the standard deviation from the standard error multiplied by the square root of (n– 1), where n is the sample size of the control or experimental group. Furthermore, when a study used multiple treatment groups for a control group (28% of identified studies), we divided the sample size of the control group equally among the treatment groups, which dealt with correlated errors and prevented sampling (error) variances being overly small [64]. Before calculating the effect sizes, we excluded data where (i) the standard error was reported as 0; or (ii) the sample size of the control group when divided was equal to or less than 1. We also excluded categorical predictors that were represented by fewer than 5 data points. Overall, 2.9% and 1.2% of all identified studies were excluded for methodological and drug treatment analyses, respectively (S3 Fig).

For meta-analysis of variance across methodological parameters, we focused on control groups with sufficient group-level information on the methodology of the experiment. Specifically, we collected and coded methodological predictors as closely as possible to the predictors used by Ström and colleagues [37] to produce a comparable meta-analysis (see full model parameters in S1 Table). For meta-analysis of variance across drug treatment, we included data from studies with sufficient group-level information on the drug group, rat strain, and sex of experimental/control groups (see full model parameters in S3 Table). For all analyses, we dealt with missing data via multiple imputation [65,66] using the package *mice* [67] as follows: We first generated multiple, simulated datasets ($m = 20$) by replacing missing values with possible values under the assumption that data are missing at random (MAR) [68,69]. After imputation, meta-analyses were performed on each imputed dataset (as described in the Statistical analysis section), and model estimates were then pooled across analyses into a single set of estimates and errors.

Calculating effect sizes

For meta-analyzing variance across methodological predictors, we calculated the lnCV and its associated sampling variance (s^2_{lnCV}) for each control group. Since many biological systems appear to exhibit a relationship between the mean and the variance on the natural scale (i.e., Taylor's law; [42,70]), an increase in the mean may correspond to an increase in variance. Our data indeed appears to exhibit a positive and linear relationship between log standard variation (lnSD) and log mean infarct volume (S1 Fig). When such a relationship holds in data, it may be most preferable to use an effect size such as lnCV, which estimates variance accounting for the mean, and this is the approach we have taken.

For meta-analyzing variance across drug treatments, we calculated the log coefficient of variance (lnCVR) and its associated sampling variance (s^2_{lnCVR}) as given in equations (11) and (12) in Nakagawa and colleagues [69] (S8 Table). When meta-analyzing variance in the presence of Taylor's law as it appears in our dataset, it may be most preferable to use lnCVR (over the log variance ratio, lnVR), which gives the variance of a contrast group accounting for differences in the mean. We therefore report all results using lnCVR in the manuscript. We note that both lnCV and lnCVR assume a linear relationship between log mean and log variance with the slope coefficient of 1 on the log scale. When slope estimates are closer to 0 or

nonlinearities are present in the mean–variance relationship, other metrics of variability such as log variability ratio (lnVR) or an approach that directly estimates the strength of association between log mean and log variance (i.e., an arm-based meta-analysis [69]) based on log SD may be more appropriate (for an example of an arm-based approach, see S9 Table and S4 Fig for galaxy plot of lnRR on lnSD). We advise that future analyses of heterogeneity pick the most appropriate statistic and model of variability based on the mean–variance relationships present in their dataset [71]. In addition to assessing the effects of treatments on variance, we further quantified differences in mean infarct volume by calculating the lnRR of the mean for each control/experimental group within a study (lnRR) and its associated sampling variance (s^2_{lnRR}). For both lnRR and lnCVR, we calculated effect sizes so that positive values corresponded to a larger mean or variance in the experimental group.

Statistical analysis

We implemented multilevel meta-analytic models in a likelihood-based package using the function "rma.mv" in the *metafor* package [72] as described in Eq 1:

$$y_{ij} = \mu + \beta x_{ij} + s_j + t_j + e_{ij} + m_{ij}, \quad (1)$$

where y_{ij} (the ith effect size of variability or mean infarct volume from a set of n effect sizes ($i = 1,2,\ldots,n$) in the jth study from a set of k studies $j = 1,2,\ldots,k$) is given by the grand mean (μ), the effects of fixed predictors (βx_{ij}), and random effects due to study (s_j), strain (t_j), residual (e_{ij}), and measurement error (m_{ij}) for the ith effect size in the jth study. Since variability in observed effects may be explained by measurement error (m_{ij} in Eq 1), we present total I^2 (the percentage of variance that cannot be explained by measurement error) and study I^2 (the percentage of variance explained by study-effects) to estimate the true variance in observed effects (i.e., meta-analytic heterogeneity) [60]. We interpreted I^2 of 25%, 50%, and 75% as small, medium, and large variance, respectively [60].

To estimate variance (lnCV) in outcome as a function of methodology in control groups, we constructed 2 meta-analytic models. First, we fitted a multilevel meta-analysis (MLMA) with the objective of estimating the overall average variability in infarct volume across studies. MLMA included a fixed intercept and random effects described in Eq 1. Second, we fitted a MLMR with the objective of estimating effects of methodological predictors on variability in infarct volume, by fitting the following fixed predictors: (i) method of occlusion; (ii) sex of animal cohort; (iii) type of ischemic model; (iv) type of anesthetic; (v) whether experiments were temperature controlled; (vi) whether rats were physiologically monitored; (vii) mean cohort weight; and (viii) time for evaluation of damage after focal ischemia (S1 Table). Mean cohort weight and time for evaluation were z-transformed prior to model fitting. We similarly constructed MLMA and MLMR models for lnRR and lnCVR (fitting each effect size as the response in separate models) to estimate the mean and variance in outcome as a function of drug treatment in our control/experimental groups, respectively. For these MLMR models, we included (i) drug treatment group, and (ii) sex of animal cohort as fixed predictors (S3 Table). Fixed effects were deemed statistically significant where their 95% credible intervals (CIs) did not span zero. For interpretation of results, we back-transformed model estimates from the log to the natural scale.

Since reported outcomes may be prone to within-study biases particularly with regard to mean estimates, we conducted a sensitivity analysis including publication quality as a random effect in our MLMR model. Publication quality was determined according to guidelines set out by the Stroke Academic Industry Roundtable (STAIR), which scored studies based on whether they implemented strategies to mitigate against both selection and detection bias [73].

Our sensitivity analysis did not lead to any qualitative changes in our main reported outcomes, and publication quality accounted for little in terms of differences in mean infarct volume (I^2_{lnRR} = 0.7%; for full sensitivity model estimates, see S5 and S6 Figs and S10–S12 Tables). Finally, we tested for signs of publication bias (systematic bias in the published data due to the preferential publication of more significant results) in our data by visual inspection of funnel plots (S7 Fig) and conducting a type of Egger regression (precision-effect test and precision-effect estimate with standard errors, PET-PEESE) on lnRR [74]. Egger regression on lnRR suggested a small effect of publication bias in our mean estimate (1.6% difference between bias-corrected and uncorrected estimates of our meta-analytic mean; see S13 Table for publication bias test results). Egger regression cannot be used for lnCVR, and further, it is unlikely that publication bias occurs for lnCVR because such biases are not driven by the difference in standard deviations between the experimental and control groups [75]. All meta-analyses were conducted on the statistical programming environment R (v 3.2.2 [76]).

Supporting information

S1 Fig. Scatter plot of log mean–variance (log SD) relationship in rat animal data. Point estimates for control (blue) and treatment (yellow) groups are provided. Slopes and 95% CIs from linear regressions for control (0.822, 0.791 to 0.854) and treatment (0.758, 0.728 to 0.788) rat groups, respectively, are shown. The data underlying this figure can be found at https://doi.org/10.6084/m9.figshare.14527317.v4. CI, credible interval.
(TIF)

S2 Fig. Relationship between variability (lnCV) induced by occlusion methodologies and consistency (ln*H*) in drug treatment outcomes. The mean slope of the relationship (slope = −0.876, −2.047 to 0.295) from the MLMR model is shown by the gray line. Circles represent estimates for each occlusion method and solid lines their 95% CIs obtained from multilevel regression (MLMR) models. Each color represents a different occlusion method, and circle sizes represent the number of effect sizes available to estimate ln*H* for each occlusion method. From the highest to lowest ln*H* estimates: orange = Filament [*N* = 973]; yellow = Mechanical/direct [*N* = 438]; blue = Endothelin injection [*N* = 76]; turquoise = Emboli/clot [*N* = 201]; green = Photothrombotic [*N* = 64]; purple = Collagenase injection [*N* = 8]; pink = spontaneous [*N* = 4]. The data underlying this figure can be found at https://doi.org/10.6084/m9.figshare.14527317.v4. CI, credible interval; lnCV, log coefficient of variation; MLMR, multilevel meta-regression.
(TIF)

S3 Fig. PRISMA flowchart of database query and study selection process. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
(TIF)

S4 Fig. Galaxy plot of treatment effects based on mean efficacy (lnRR) and interindividual variability in efficacy as obtained from an arm-based meta-analysis of lnSD. Estimates (circles) represent unconditional (marginalized), treatment-specific means (lnRR), variability (lnSD), and their 95% CIs (solid lines). We reveal differences in variability across treatment groups, with treatments that significantly reduce infarct volume and increase interindividual variability (positive lnSD) highlighted blue. The effects of hypothermia and thrombolytics (the latter of which include the only regulatory-approved treatment) are highlighted in pink. The data underlying this figure can be found at https://doi.org/10.6084/m9.figshare.14527317.v4. CI, credible interval; lnRR, log response ratio; lnSD, log standard variation.
(TIF)

S5 Fig. Sensitivity model CV estimates from multilevel regression (MLMR) of infarct volume in control groups. Mean estimates of unconditional (marginalized), group-specific coefficients of variation (%) are indicated as gray circles. Moreover, 95% CIs are shown as gray lines and are asymmetric due to back-transformation of log coefficient of variation (lnCV) to the natural scale. The data underlying this figure can be found at https://doi.org/10.6084/m9.figshare.14527317.v4. CI, credible interval; lnCV, log coefficient of variation; MLMR, multilevel meta-regression.
(TIF)

S6 Fig. Sensitivity model (a) lnRR and (b) lnCVR estimates from multilevel regression (MLMR) of infarct volume in treatment/control groups. Mean estimates of unconditional (marginalized), group-specific effects are shown as gray circles, and 95% CIs are shown as gray lines. Negative lnRR estimates indicate that mean infarct volume is smaller in experimental versus control rats. Negative lnCVR estimates show that interindividual variability in infarct volume is smaller in experimental versus control rats. The data underlying this figure can be found at https://doi.org/10.6084/m9.figshare.14527317.v4. CI, credible interval; lnCVR, log coefficient of variation ratio; lnRR, log response ratio; MLMR, multilevel meta-regression.
(TIF)

S7 Fig. Funnel plot for lnRR characterizing differences in mean infarct volume for control/treatment groups. Raw effect sizes are plotted against their precision (inverse of the square root of standard error). MLMA model predicted mean effect size (solid vertical line), and its 95% CI (dashed lines) are shown. The data underlying this figure can be found at https://doi.org/10.6084/m9.figshare.14527317.v4. CI, credible interval; lnRR, log response ratio; MLMA, multilevel meta-analysis.
(TIF)

S1 Table. Unconditional (marginalized) estimates and 95% credible intervals for lnCV, obtained from multilevel regression (MLMR) models of control group infarct volume. Continuous predictors were Z-transformed prior to model fitting. lnCV, log coefficient of variation; MLMR, multilevel meta-regression.
(DOCX)

S2 Table. Conditional estimates and 95% credible intervals for lnCV, obtained from multilevel regression (MLMR) models of control group infarct volume. Continuous predictors were Z-transformed prior to model fitting. Bold italicized estimates indicate that the 95% credible intervals do not span zero. lnCV, log coefficient of variation; MLMR, multilevel meta-regression.
(DOCX)

S3 Table. Unconditional (marginalized) estimates and 95% credible intervals for lnRR and lnCVR, obtained from multilevel regression (MLMR) models of infarct volume in treatment/control groups. Treatment effects (DrugGroup) are ordered from groups that produce, on average, the greatest reduction in infarct volume (i.e., the most effective, as indicated by most negative estimates of lnRR) to groups that are, on average, the least effective. lnCVR, log coefficient of variation ratio; lnRR, log response ratio; MLMR, multilevel meta-regression.
(DOCX)

S4 Table. Conditional estimates and 95% credible intervals for lnRR and lnCVR, obtained from multilevel regression (MLMR) models of infarct volume in treatment/control groups. Bold italicized estimates indicate that the 95% credible intervals do not span zero. lnCVR, log

coefficient of variation ratio; lnRR, log response ratio; MLMR, multilevel meta-regression.
(DOCX)

S5 Table. Conditional estimates and 95% credible intervals for lnRR and lnCVR, obtained from contrast multilevel regression (MLMR) models to assess the effect of sex on infarct volume. The intercept here represents studies in which "Both" sexes were used. Bold italicized estimates indicate that the 95% credible intervals do not span zero. lnCVR, log coefficient of variation ratio; lnRR, log response ratio; MLMR, multilevel meta-regression.
(DOCX)

S6 Table. Conditional estimates and 95% credible intervals for lnRR and lnCVR, obtained from contrast multilevel regression (MLMR) models to assess the effect of sex on infarct volume. The intercept here represents studies in which only "Female" sex was used. Bold italicized estimates indicate that the 95% credible intervals do not span zero. lnCVR, log coefficient of variation ratio; lnRR, log response ratio; MLMR, multilevel meta-regression.
(DOCX)

S7 Table. Consistency in drug treatment outcomes across variability induced by occlusion methodologies. For our second-order meta-regression, we first separated our rat infarct volume data by occlusion methods. For each occlusion method data, we conducted a MLMR to estimate heterogeneity (I^2) in lnRR including our original random (study ID, effect size ID, and strain) and fixed effects (sex + drug treatment group). From our MLMR models, we extracted total I^2 of lnRR and from this calculated the heterogeneity statistic ln$H$. ln$H$ is a preferable effect size for downstream analyses as it is unbounded and has a relatively well-defined standard error to act as a measure of its precision [61 in main text]. Using the square of the standard error of ln$H$ as the sampling variance and ln$H$ as our response variable, we then fit a second-order meta-regression using the lnCV estimates of each occlusion method as a fixed predictor and effect size ID as a random effect ($\sigma^2_{Residual} = 0.200$). Unconditional estimates of lnCV were obtained from our MLMR models of methodological variability (S1 Table) described in our main text. Estimates and 95% credible intervals from this second-order MLMR model is reported below. Estimates with credible intervals that do not span zero are considered statistically significant. See S3 Fig for a line plot depicting the relationship between ln$H$ and lnCV with the model fitted line. lnCV, log coefficient of variation; lnRR, log response ratio; MLMR, multilevel meta-regression.
(DOCX)

S8 Table. Effect sizes and sampling variances used in meta-analysis of variance (a) across methodological predictors and (b) across drug treatment groups. Equations and the model type in which the effect size was used are also given. $\underline{x}$ and s are the mean and SD of the group infarct volume, n is the sample size, CV is the coefficient of variation, and ρ is the correlation between the mean and standard deviation on the log scale (ρ is assumed to be 0*). Subscripts C and E refer to control and treatment groups, respectively.
(DOCX)

S9 Table. Model estimates (unconditional) and 95% credible intervals for lnRR and lnSD. Estimates of lnSD were obtained from an arm-based, multilevel regression model (MLMR) of lnSD in infarct volume for both treatment and control groups. Original fixed (drug treatment group and sex) and random effects (study ID, effect size ID, and strain) were fit, in addition to a nested random effect of "Drug treatment group | pairwise ID." Estimates of lnRR are from the main analysis of mean drug treatment effects and are the same as in S3 Table. Treatment effects (DrugGroup) are ordered from groups that produce, on average, the greatest reduction

in infarct volume (i.e., the most effective, as indicated by most negative estimates of lnRR) to groups that are, on average, the least effective. lnRR, log response ratio; lnSD, log standard variation; MLMR, multilevel meta-regression.
(DOCX)

S10 Table. Sensitivity model estimates (unconditional) and 95% credible intervals for lnCV, obtained from multilevel regression (MLMR) models of control group infarct volume. Continuous predictors were Z-transformed prior to model fitting. lnCV, log coefficient of variation; MLMR, multilevel meta-regression.
(DOCX)

S11 Table. Sensitivity model estimates (unconditional) and 95% credible intervals for lnRR and lnCVR, obtained from multilevel regression (MLMR) models of infarct volume in treatment/control groups. Treatment effects (DrugGroup) are ordered from groups that produce, on average, the greatest reduction in infarct volume (i.e., the most effective, as indicated by most negative estimates of lnRR) to groups that are, on average, the least effective. lnCVR, log coefficient of variation ratio; lnRR, log response ratio; MLMR, multilevel meta-regression.
(DOCX)

S12 Table. Sensitivity model estimates of heterogeneity (I^2) for analyses of methodology on variability (lnCV) and drug treatment on mean (lnRR) and variance (lnCVR) in rat infarct volume. Estimates (%) are shown for MLMAs and MLMR models. lnCV, log coefficient of variation; lnCVR, log coefficient of variation ratio; lnRR, log response ratio; MLMA, multilevel meta-analysis; MLMR, multilevel meta-regression.
(DOCX)

S13 Table. Results from Egger regression (PET-PEESE) on lnRR to test for publication bias. This procedure fits the square root of sampling variance as a moderator (slope estimate and 95% credible intervals shown in the first half of the table). If this estimate is significant, we then fit the sampling variance (second half of the table). The intercept from this latter model indicates a "potentially" bias-corrected, modified meta-analytic mean. In our case, the biased-corrected estimate is a 28.0% decline, compared with the original estimate without correction, which is a 29.6% decline. These values indicate that although this analysis detected a sign of publication bias, the effect of this bias is very small (1.6% difference). Bold italicized estimates indicate that the 95% credible intervals do not span zero. lnRR, log response ratio; PEESE, precision-effect estimate with standard errors; PET, precision-effect test.
(DOCX)

Acknowledgments

We would like to thank the CAMARADES team for help in data access and extraction and the I-DEEL lab for providing the opportunity for TU to conduct this meta-analysis. We thank Megan Szojka and Mia Waters for providing feedback on an earlier draft.

Author Contributions

Conceptualization: Takuji Usui, Alistair M. Senior, Shinichi Nakagawa.

Data curation: Takuji Usui, Malcolm R. Macleod, Sarah K. McCann, Shinichi Nakagawa.

Formal analysis: Takuji Usui, Alistair M. Senior, Shinichi Nakagawa.

Funding acquisition: Alistair M. Senior, Shinichi Nakagawa.

Supervision: Alistair M. Senior, Shinichi Nakagawa.

Writing – original draft: Takuji Usui.

Writing – review & editing: Takuji Usui, Malcolm R. Macleod, Sarah K. McCann, Alistair M. Senior, Shinichi Nakagawa.

References

1. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med. 2016; 341:96–102. https://doi.org/10.1126/scitranslmed.aaf5027 PMID: 27252173
2. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2:696–701. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722
3. Begley CG, Ioannidis JPA. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res. 2015; 116:116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819 PMID: 25552691
4. Frye SV, Arkin MR, Arrowsmith CH, Conn PJ, Glicksman MA, Hull-Ryde EA, et al. Tackling reproducibility in academic preclinical drug discovery. Nat Rev Drug Discov. 2015; 14:733–734. https://doi.org/10.1038/nrd4737 PMID: 26388229
5. Baker M. Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the crisis rocking science and what they think will help. Nature. 2016; 533:452–455. https://doi.org/10.1038/533452a PMID: 27225100
6. Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. Nat Rev Neurol. 2014; 10:37–43. https://doi.org/10.1038/nrneurol.2013.232 PMID: 24247324
7. Seyhan AA. Lost in translation: the valley of death across preclinical and clinical divide–identification of problems and overcoming obstacles. Transl Med Commun. 2019; 4(18). https://doi.org/10.1186/s41231-019-0050-7
8. Festing MF. Reduction of animal use: experimental design and quality of experiments. Lab Anim. 1994; 28:212–221. https://doi.org/10.1258/002367794780681697 PMID: 7967459
9. Beynen AC, Baumans V, Van Zutphen LFM. Principles of Laboratory Animal Science. Amsterdam: Elsevier; 2001.
10. Würbel H. Behaviour and the standardization fallacy. Nat Genet. 2000; 26:263. https://doi.org/10.1038/81541 PMID: 11062457
11. Richter SH, Garner J P, 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
12. Richter SH. Systematic heterogenization for better reproducibility in animal experimentation. Lab Anim. 2017; 46:343–349. https://doi.org/10.1038/laban.1330 PMID: 29296016
13. Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. Science. 1999; 284:1670–1672. https://doi.org/10.1126/science.284.5420.1670 PMID: 10356397
14. 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
15. Karp NA. Reproducible preclinical research–is embracing variability the answer? PLoS Biol. 2018; 16: e2005413. https://doi.org/10.1371/journal.pbio.2005413 PMID: 29505576
16. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et al. Reproducibility of animal research in light of biological variation. Nat Rev Neurosci. 2020; 21:384–393. https://doi.org/10.1038/s41583-020-0313-3 PMID: 32488205
17. Chesler EJ, Wilson SG, Lariviere WR, Rodriguez-Zas SL, Mogil JS. Influences of laboratory environment on behavior. Nat Neurosci. 2002; 5:1101–1102. https://doi.org/10.1038/nn1102-1101 PMID: 12403996
18. Mueller FS, Polesel M, Richetto J, Meyer U, Weber-Stadlbauer U. Mouse models of maternal immune activation: mind your caging system! Brain Behav Immun. 2018; 73:643–660. https://doi.org/10.1016/j.bbi.2018.07.014 PMID: 30026057
19. Richter SH, Garner JP, Zipser B, Lewejohann L, Sachser N, Touma C, et al. Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. PLoS ONE. 2011; 6: e16461. https://doi.org/10.1371/journal.pone.0016461 PMID: 21305027

20. Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biol. 2018; 16:e2003693. https://doi.org/10.1371/journal.pbio.2003693 PMID: 29470495

21. Kafkafi N, Golani I, Jaljuli I, Morgan H, Sarig T, Würbel H, et al. Addressing reproducibility in single-laboratory phenotyping experiments. Nat Methods. 2017; 14:462–464. https://doi.org/10.1038/nmeth.4259 PMID: 28448068

22. Bodden C, von Kortzfleisch VT, Karwinkel F, Kaiser S, Sachser N, Richter H. Heterogenising study samples across testing time improves reproducibility of behavioural data. Sci Rep. 2019; 9:8247. https://doi.org/10.1038/s41598-019-44705-2 PMID: 31160667

23. Karp NA, Speak AO, White JK, Adams DJ, de Angelis MH, Hérault Y, et al. 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

24. Milcu A, Puga-Freitas R, Ellison AM, Blouin M, Scheu S, Freschet GT, et al. Genotypic variability enhances the reproducibility of an ecological study. Nat Ecol Evol. 2018; 2:279–287. https://doi.org/10.1038/s41559-017-0434-x PMID: 29335575

25. Llovera G, Hofmann K, Roth S, Salas-Pérdomo A, Ferrer-Ferrer M, Perego C, et al. Results of a preclinical randomized controlled multicenter trial (pRCT): Anti-CD49d treatment for acute brain ischemia. Sci Transl Med. 2015; 7(299). https://doi.org/10.1126/scitranslmed.aaa9853 PMID: 26246166

26. Festing MF. Refinement and reduction through the control of variation. Altern Lab Anim. 2004; 32:259–263. https://doi.org/10.1177/026119290403201s43 PMID: 23577470

27. Festing MF. Evidence should trump intuition by preferring inbred strains to outbred stocks in preclinical research. ILAR J. 2014; 55:399–404. https://doi.org/10.1093/ilar/ilu036 PMID: 25541542

28. Willmann R, De Luca A, Benatar M, Grounds M, Dubach J, Raymackers J-M, et al. Enhancing translation: guidelines for standard pre-clinical experiments in mdx mice. Neuromuscul Disord. 2012; 22:43–49. https://doi.org/10.1016/j.nmd.2011.04.012 PMID: 21737275

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

30. Schork NJ. Personalized medicine: time for one-person trials. Nature. 2015; 520(7549):609–11. https://doi.org/10.1038/520609a PMID: 25925459

31. Dirnagl U. Bench to bedside: The quest for quality in experimental stroke research. J Cerebr Blood F Met. 2006; 26(12):1465–1478. https://doi.org/10.1038/sj.jcbfm.9600298 PMID: 16525413

32. Howells DW, Porritt MJ, Rewell SSJ, O'Collins V, Sena ES, Van Der Worp HB, et al. Different strokes for different folks: The rich diversity of animal models of focal cerebral ischemia. J Cerebr Blood F Met. 2010; 30(8):1412–1431. https://doi.org/10.1038/jcbfm.2010.66 PMID: 20485296

33. O'Collins VE, Macleod MR, Donnan GA, Horky LL, Van Der Worp BH, Howells DW. 1,026 Experimental treatments in acute stroke. Ann Neurol. 2006; 59(3):467–477. https://doi.org/10.1002/ana.20741 PMID: 16453316

34. Howells DW, Sena ES, O'Collins VE, Macleod MR. Improving the efficiency of the development of drugs for stroke. Int J Stroke. 2012; 7(5):371–377. https://doi.org/10.1111/j.1747-4949.2012.00805.x PMID: 22712738

35. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, et al. Comparison of treatment effects between animal experiments and clinical trials: Systematic review. Brit Med J. 2007; 334(7586):197–200. https://doi.org/10.1136/bmj.39048.407928.BE PMID: 17175568

36. Thomas A, Detilleux J, Flecknell P, Sandersen C. Impact of stroke therapy academic industry roundtable (STAIR) guidelines on peri-anesthesia care for rat models of stroke: A meta-analysis comparing the years 2005 and 2015. PLoS ONE. 2017; 12(1):1–18. https://doi.org/10.1371/journal.pone.0170243 PMID: 28122007

37. Ström JO, Ingberg E, Theodorsson A, Theodorson E. Method parameters' impact on mortality and variability in rat stroke experiments: A meta-analysis. BMC Neurosci. 2013; 14:41. https://doi.org/10.1186/1471-2202-14-41 PMID: 23548160

38. Ingberg E, Dock H, Theodorsson E, Theodorsson A, Ström JO. Method parameters' impact on mortality and variability in mouse stroke experiments: A meta-analysis. Sci Rep. 2016; 6. https://doi.org/10.1038/srep21086 PMID: 26876353

39. Van der Worp HB, Van Gijn J. Clinical practice. Acute ischemic stroke. N Engl J Med. 2007; 357:572–579. https://doi.org/10.1056/NEJMcp072057 PMID: 17687132

40. Adams HP, Adams RJ, Brott T, Del Zoppo GJ, Furlan A, Goldstein LB. Guidelines for the early management of patients with ischemic stroke: A scientific statement from the Stroke Council of the American Stroke Association. Stroke. 2003; 34(4):1056–1083. https://doi.org/10.1161/01.STR.0000064841.47697.22 PMID: 12677087

41. Vesterinen HM, Sena ES, Egan KJ, Hirst TC, Churolov L, Currie GL, et al. Meta-analysis of data from animal studies: A practical guide. J Neurosci Methods. 2014; 221:92–102. https://doi.org/10.1016/j.jneumeth.2013.09.010 PMID: 24099992

42. Taylor BLR. Aggregation, variance and the mean. Nature. 1961; 189:732–735.

43. Plöderl M, Hengartner MP. What are the chances for personalised treatment with antidepressants? Detection of patient-by-treatment interaction with a variance ratio meta-analysis. BMJ Open. 2019; 9 (12):1–6. https://doi.org/10.1136/bmjopen-2019-034816 PMID: 31874900

44. Zhang H, Lin S, Chen X, Gu L, Zhu X, Zhang Y, et al. The effect of age, sex and strains on the performance and outcome in animal models of stroke. Neurochem Int. 2019; 127:2–11. https://doi.org/10.1016/j.neuint.2018.10.005 PMID: 30291954

45. McCullough LD, Liu F. Middle cerebral artery occlusion model in rodents: Methods and potential pitfalls. J Biomed Biotechnol. 2011. https://doi.org/10.1155/2011/464701 PMID: 21331357

46. Haast RAM, Gustafson DR, Kiliaan AJ. Sex differences in stroke. J Cerebr Blood F Met. 2012; 32 (12):2100–2107. https://doi.org/10.1038/jcbfm.2012.141 PMID: 23032484

47. Turtzo LC, McCullough LD. Sex-specific responses to stroke. Future Neurol. 2010; 5(1):47–59. https://doi.org/10.2217/fnl.09.66 PMID: 20190872

48. Walberer M, Müller ESC. Experimental stroke: ischaemic lesion volume and oedema formation differ among rat strains (a comparison between Wistar and Sprague–Dawley rats using MRI). Lab Anim. 2006; 40(1):1–8. https://doi.org/10.1258/002367706775404426 PMID: 16460584

49. Miller LR, Marks C, Becker JB, Hurn PD, Chen W-J, Woodruff T, et al. Considering sex as a biological variable in preclinical research. FASEB J. 2017; 31:29–34. https://doi.org/10.1096/fj.201600781R PMID: 27682203

50. Clayton JA, Collins FS. NIH to balance sex in cell and animal studies. Nature. 2014; 509(7500):282–283. https://doi.org/10.1038/509282a PMID: 24834516

51. Clayton JA. Applying the new SABV (sex as a biological variable) policy to research and clinical care. Physiol Behav. 2018; 187:2–5. https://doi.org/10.1016/j.physbeh.2017.08.012 PMID: 28823546

52. European Medicines Agency. ICH guideline M3(R2) on non-clinical safety studies for the conduct of human clinical trials and marketing authorisation for pharmaceuticals. 2013. EMA/CPMP/ICH/286/1995.

53. Bogue MA, Churchill GA, Chesler EJ. Collaborative cross and diversity outbred data resources in the mouse phenome database. Mamm Genome. 2015; 26:511–520. https://doi.org/10.1007/s00335-015-9595-6 PMID: 26286858

54. Prendergast BJ, Onishi KG, Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. Neurosci Biobehav Rev. 2014; 40:1–5. https://doi.org/10.1016/j.neubiorev.2014.01.001 PMID: 24456941

55. Becker JB, Prendergast BJ, Liang JW. Female rats are not more variable than male rats: a meta- analysis of neuroscience studies. Biol Sex Differ. 2016; 7:34. https://doi.org/10.1186/s13293-016-0087-5 PMID: 27468347

56. Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L. Sex and gender analysis improves science and engineering. Nature. 2019; 575(7781):137–46. https://doi.org/10.1038/s41586-019-1657-6 PMID: 31695204

57. Buch T, Moos K, Ferreira FM, Fröhlich H, Gebhard C, Tresch A. Benefits of a factorial design focusing on inclusion of female and male animals in one experiment. J Mol Med. 2019; 97:871–877. https://doi.org/10.1007/s00109-019-01774-0 PMID: 30980104

58. Ebersole CR, Klein RA, Atherton OE. The Many Lab. 2019 Mar 27 [cited 2020 Oct 15]. Available from: osf.io/89vqh.

59. Voekl B, Würbel H, Krzywinski M, Altman N. The standardization fallacy. Nat Methods. 2021;5–7. https://doi.org/10.1038/s41592-020-01036-9 PMID: 33408399

60. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21 (11):1539–1558. https://doi.org/10.1002/sim.1186 PMID: 12111919

61. Naylor S, Chen JY. Unraveling human complexity and disease with systems biology and personalized medicine. Pers Med. 2010; 7(3):275–289. https://doi.org/10.2217/pme.10.16 PMID: 20577569

62. van der Worp HB, Macleod MR, Bath PMW, Bathula R, Christensen H, Colam B, et al. Therapeutic hypothermia for acute ischaemic stroke. Results of a European multicentre, randomised, phase III clinical trial. Eur Stroke J. 2019; 4(3):254–262. https://doi.org/10.1177/2396987319844690 PMID: 31984233

63. Winkelbeiner S, Leucht S, Kane JM, Homan P. Evaluation of differences in individual treatment response in schizophrenia spectrum disorders: a meta-analysis. JAMA Psychiat. 2019; 76(10):1063–1073. https://doi.org/10.1001/jamapsychiatry.2019.1530 PMID: 31158853

64. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane handbook for systematic reviews of interventions. Chichester: John Wiley & Sons; 2019.

65. Schafer JL. Multiple imputation: A primer. Stat Methods Med Res. 1999; 8(1):3–15. https://doi.org/10.1177/096228029900800102 PMID: 10347857

66. Nakagawa S, Freckleton RP. Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol. 2012; 23(11):592–596.

67. van Buuren S., Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2011; 45(3). https://doi.org/10.18637/jss.v045.i03

68. Little RJ, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons; 2019.

69. Nakagawa S, Poulin R, Mengersen K, Reinhold K, Engqvist L, Lagisz M, et al. Meta-analysis of variation: Ecological and evolutionary applications and beyond. Methods Ecol Evol. 2015; 6(2):143–152.

70. Cohen JE, Xu M. Random sampling of skewed distributions implies Taylor's power law of fluctuation scaling. Proc Natl Acad Sci U S A. 2015; 112(25):7749–7754. https://doi.org/10.1073/pnas.1503824112 PMID: 25852144

71. Volkmann C, Volkmann A, Müller CA. On the treatment effect of heterogeneity of antidepressants in major depression: A Bayesian meta-analysis and simulation study. PLoS ONE. 2020; 15(11): e0241497. https://doi.org/10.1371/journal.pone.0241497 PMID: 33175895

72. Viechtbauer W. Conducting meta-analyses in R with the metafor. J Stat Softw. 2010; 36(3):1–48.

73. Malcolm MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke. 2004; 35(5):1203–1208. https://doi.org/10.1161/01.STR.0000125719.25853.20 PMID: 15060322

74. Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. Res Synth Methods. 2014; 5(1):60–78. https://doi.org/10.1002/jrsm.1095 PMID: 26054026

75. Senior AM, Gosby AK, Lu J, Simpson SJ, Raubenheimer D. Meta-analysis of variance: an illustration comparing the effects of two dietary interventions on variability in weight. Evol Med Public Health. 2016; 1:244–255.

76. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2014. Available from: http://www.R-project.org/.