1   # Embrace heterogeneity to improve reproducibility: A perspective

2   # from meta-analysis of variation in preclinical research

3

4   Takuji Usui[1,2,#a*], Malcolm R. Macleod[3], Sarah K. McCann[4,5], Alistair M. Senior[2¶*] and

5   Shinichi Nakagawa[1,2¶*]

6   [1] Evolution and Ecology Research Centre and School of Biological, Earth and Environmental
7   Sciences, University of New South Wales, Sydney, Australia
8
9   [2] The Charles Perkins Centre, and School of Life and Environmental Sciences, The
10   University of Sydney, Sydney, Australia
11
12   [3] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom
13
14   [4] QUEST Center for Transforming Biomedical Research, Berlin Institute of Health (BIH),
15   Berlin, Germany
16
17   [5] Charité - Universitätsmedizin Berlin Corporate member of Freie Universität Berlin,
18   Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
19
20   [#a] Current address: Biodiversity Research Centre, University of British Columbia,
21   Vancouver, Canada
22

23   *** Corresponding authors:**

24   Emails: usuitakuji@gmail.com (TU), alistair.senior@sydney.edu.au (AMS),

25   s.nakagawa@unsw.edu.au (SN)

26   ¶ These authors contributed equally to this work.

27

28   **Short title:** Embrace heterogeneity to improve reproducibility

29   **Keywords:** effect size, heterogenization, meta-regression, precision medicine, rat,

30   standardization, stroke, translation, variance

31

Abstract

The reproducibility of research results has been a cause of increasing concern to the scientific community. The long-held belief that experimental standardization begets reproducibility has also been recently challenged, with the observation that the reduction of variability within studies can lead to idiosyncratic, lab-specific results that are irreproducible. 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 ischaemic stroke. First, by quantifying inter-individual variability across control groups we illustrate that the samount 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 reproducibility 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 inter-individual variability), as well as those where there is high inter-individual variability in response; for these latter treatments translation to a clinical setting may require nuance. We argue that by embracing rather than seeking to minimise variability in phenotypic outcomes, we can motivate the shift towards heterogenization and improve both the reproducibility and generalizability of preclinical research.

2

Introduction

Reproducibility 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" [1] – is integral to scientific progress. Compelling evidence, however, suggests that irreproducibility 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 minimise 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 reproducibility 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 reproducibility, 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

3

81   unmeasured variables, standardization may thus weaken, rather than strengthen,

82   reproducibility in preclinical studies.

83

84   A potential counter to this "standardization fallacy" [10] then, is to improve reproducibility

85   by embracing, rather than minimizing, heterogeneity [10-12]. Practical solutions to enhance

86   external validity include conducting studies across multiple laboratories to deliberately

87   account for differences in within-lab variability [19-21], and perhaps more radically, to

88   systematically introduce variability into experimental designs within studies [12, 22, 23].

89   Both simulation [11, 14, 20, 21] and empirical studies [19, 22, 24, 25] show that deliberate

90   inclusion of more heterogeneous study samples and experimental conditions (i.e.,

91   "heterogenization") improve external validity, and hence reproducibility, by increasing

92   within-study (or within-lab) variability and minimizing among-study (or among-lab)

93   variability.

94

95   Despite the promise of heterogenization, standardization remains the conventional approach

96   in preclinical studies [26-28]. This has been partly fuelled by Russel and Birch's [29]

97   injunction to a "reduction in the numbers of animals used to obtain information of a given

98   amount and precision". Consequently, within-study variability is typically treated as a

99   biological inconvenience that is to be minimised, rather than an outcome of interest in its own

100   right. Embracing and quantifying heterogeneity, however, may benefit preclinical science in

101   at least two ways. First, through comparative analyses of the variability associated with

102   experimental procedures we may identify methodologies that introduce variation. As

103   discussed above, by using methods that induce variation one may design a deliberately

104   heterogeneous study with greater reproducibility [10-12]. Second, by explicitly investigating

105   inter-individual heterogeneity in the response to drug/intervention outcomes, we may

106 quantify the generalisability of a treatment and its translational potential. That is, a treatment

107 with low inter-individual variation in efficacy despite heterogenization is more generalizable

108 while a treatment with high inter-individual variation indicates the effect may be individual-

109 specific. This may be relevant in the context of personalized medicine: A treatment

110 associated with inter-individual variation in outcome may require tailoring in its clinical use

111 [30]. Taking these two points together, one could argue an ideal trial would use a technical

112 design that typically generated variation in disease state, which was then attenuated by a

113 treatment of interest that might consistently (in all animals) or selectively (in some animals)

114 improve outcome.

115

116 An illustrative case where the issues of reproducibility and lack of translation has been

117 highlighted repeatedly is that of animal models of ischaemic stroke [31-33]. Several

118 systematic reviews [34, 35] and meta-analyses [36-38] have questioned the propriety of

119 experimental design and the choice of experimental procedures in stroke animal studies. The

120 consequent recommendation for improving reproducibility in the field has usually been to

121 adopt methodological procedures that minimize heterogeneity (and/or mitigate sources of

122 bias) in phenotypic outcomes (e.g. in infarct volume or neurobehavioral outcomes) [34-38].

123 Furthermore, whilst potentially beneficial treatments have been identified in individual trials

124 at the preclinical stage, intravenous thrombolysis remains the only regulatory approved

125 treatment for ischaemic stroke [33, 39, 40]. This lack of transferable results from the

126 preclinical to clinical stage highlights a major shortcoming for the generalizability of stroke

127 animal models and is emblematic of translation failures generally across preclinical studies

128 [6, 7, 33, 34].

129

130 Using the case of rat animal models of stroke as a guiding example, we highlight how

131 recently developed methods for the meta-analysis of variation can be used to better

132 understand biological heterogeneity. First, through analysis of variability using the log

133 coefficient of variation (lnCV; CV representing variance relative to the mean) in control

134 groups, we identify methodological procedures that increase variability in outcomes. Second,

135 we show how, through the concurrent meta-analysis of mean and variance in treatment

136 effects using the log response ratio (lnRR; i.e. ratio of means) and log coefficient of variation

137 ratio (lnCVR), one gains additional information about the generalisability of an intervention

138 at the individual level. Overall, we argue that the quantification of heterogeneity in

139 phenotypic outcomes can be exploited to improve both the reproducibility and translation of

140 animal studies.

141

Results

Dataset

144 We obtained data for rat animal models of ischaemic stroke from the Collaborative Approach

145 to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES)

146 database [41], focusing our meta-analysis on animal models that reported outcomes in infarct

147 volume (see Materials and Methods for inclusion criteria of studies). We extracted data for

148 infarct volume from 1318 control group cohorts from 778 studies for our analyses

149 investigating the effects of methodology and variability. We extracted data for the effect of

150 treatment on infarct volume from 1803 treatment/control group cohort pairs from 791 studies

151 for our analyses investigating the effects of drug treatment on inter-individual variability (see

152 S1 Data for extracted database used in this study).

153

Methodology and variability

155   To identify methodological procedures that generated variability in disease-state, we first

156   meta-analysed variability in infarct volume for control group animals. We quantify variability

157   as the log coefficient of variation (lnCV) rather than the log of standard deviation because we

158   found that our data showed a mean-variance relationship (i.e. Taylor's Law, where the

159   variance increases with an increase in the mean [42]; S1 Fig). Overall, the coefficient of

160   variation (CV) in infarct volume across control groups was around 23.6% of the mean (lnCV

161   = –1.444, CI = –1.546 to –1.342 τ^2 = 0.565; Fig 1). We found large differences in variability

162   of infarct volume ($I^2{}_{total}$ = 93.7%), suggesting that sampling variance alone cannot account

163   for differences in the reported variability across control groups (Table 1). The I^2 attributable

164   to study was 49.6% suggesting that methodological differences across studies explained some

165   of this heterogeneity, although a moderate amount (42.9%) of I^2 remained unexplained

166   (Table 1).

167

168   **Table 1. Heterogeneity (I^2) estimates for analyses of methodology on variability (lnCV)**

169   **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%

7

lnCVR

MLMA	71.2%	38.8%	0.9%	31.6%
MLMR	70.3%	36.1%	1.2%	33.1%

170 Estimates (%) are shown for multi-level meta-analyses (MLMA) and multilevel meta-

171 regression (MLMR) models.

172

173 We detected statistically significant differences in variability of infarct volume between

174 various methodological approaches (Fig 1; see S1 and S2 Tables in S1 Text for unconditional

175 and conditional model coefficients, respectively). Among occlusion methods, models with

176 spontaneous occlusion produced the greatest variability in infarct volume (CV = 52.5%; lnCV

177 = –0.644, –1.633 to 0.345), whilst filamental occlusion had lowest variability (CV = 17.9%;

178 lnCV = –1.720, –2.195 to –1.244). Studies using temporary models of ischaemia had higher

179 variability in infarct volume (CV = 25.2%; lnCV = –1.377, –1.500 to –1.255) compared to

180 permanent models. Variability was slightly but significantly lower with longer time of

181 damage assessment (lnCV = –1.404, –1.521 to –1.288) and greater median weight of the

182 control group cohort (lnCV = –1.366, –1.486 to –1.245).

183

184 **Drug treatment effects and inter-individual variation**

185 To quantify generalizability in drug treatment outcomes, we meta-analysed the mean and the

186 coefficient of variation in infarct volume for the effects observed in control/experimental

187 contrasts. We quantified the mean and inter-individual variability as the log response ratio

188 (lnRR) and log coefficient of variation ratio (lnCVR), respectively. Overall, mean infarct

189 volume in experimental groups was around 33.1%, smaller than in control groups (lnRR =

190 –0.402, –0.461 to –0.343; Fig 2A); whilst the coefficient of variation in experimental groups

8

191   was around 32.4% higher than in control groups (lnCVR = 0.280, 0.210 to 0.351; Fig 2B).

192   Overall, heterogeneity in lnRR was very high, while that for lnCVR was moderate, and

193   moderate amounts of heterogeneity were partitioned into the study-level for both (Table 1).

194

195   Both the mean and variability in infarct volume differed significantly across drug treatment

196   groups (Fig 2; S3 and S4 Tables in S1 Text for unconditional and conditional model

197   coefficients, respectively). Treatment with hypothermia resulted in the largest reduction of

198   mean infarct volume in experimental groups relative to controls (around 49.7% lower in

199   experimental groups than controls; lnRR = –0.687, –0.775 to –0.599). However, hypothermia

200   also had the most variable and inconsistent effect (i.e. inter-subject variation) in reducing

201   infarct volume, with the largest ratio of CV between experimental and control groups (inter-

202   individual variability around 60.0% higher in experimental groups compared to controls;

203   lnCVR = 0.470, 0.349 to 0.591). In contrast, environmental treatments were the least

204   effective in reducing mean infarct volume (around 7.3% greater in experimental groups than

205   controls; lnRR = 0.071, –0.166 to 0.308). Hyperbaric oxygen therapy (HBOT) has the least

206   variable and most consistent effect on infarct volume (variability around 45.3% less in

207   experimental groups relative to controls; lnCVR = –0.603, –1.483 to 0.277).

208

209   Thrombolytics, which include the only regulatory approved treatment (i.e., tissue

210   plasminogen activator; tPA [42]), reduced mean infarct volume by around 29.6% in

211   experimental relative to control groups (lnRR = –0.351, –0.446 to –0.256). The CV across

212   experimental groups for thrombolytics was around 17.4% higher than control groups (lnCVR

213   = 0.160, 0.031 to 0.289), but it is notable that this increased inter-subject variability is much

214   less than that seen with hypothermia. Through quantifying variability in drug treatment

215   outcomes, we propose that treatments be considered generalizable if they reduced mean

216   infarct volume and concurrently show low inter-individual variability (i.e. negative lnRR and

217   lnCVR estimates; Fig 3). Drug treatments that on average reduced infarct volume but had

218   variable and inconsistent effects (i.e. had negative lnRR and positive lnCVR estimates; Fig 3)

219   are ungeneralizable but might be appropriate for clinical exploitation in selected patients [30;

220   43]. Conversely, the least successful treatments can be identified as those that consistently do

221   not reduce mean infarct volume (i.e. positive lnRR and lnCVR estimates; Fig 3). We

222   explored whether the sex of groups used in experiments affected lnRR or lnCVR (see

223   Methods for multilevel meta-regression model parameters) but differences in mean or

224   variability of infarct volume did not vary significantly between female and male cohorts (see

225   S5 and S6 Tables in S1 Text for contrast model estimates for sex effects).

226

Discussion

228   We propose that the current failures in reproducibility and translation of preclinical trials may

229   be due, at least in part, to the way studies are designed and assessed, which is to minimise

230   within-study variation and ignore heterogeneity in outcomes [8, 9, 26-28]. Here, we have

231   illustrated the potential utility of embracing such heterogeneity, through meta-analysing

232   variability (relative variance or CV) in outcomes for rat animal models of stroke. First, by

233   estimating the variability generated by different methodological designs applied to control

234   animal groups, we have identified procedures that generate variability in disease-states (Fig

235   1). Second, we have, for the first time, quantified both the efficacy and stability (i.e., changes

236   in the mean and variance, respectively) of stroke treatments applied to the experimental

237   animal models (Fig 2; Fig 3), identifying potential treatments that may be generalizable

238   versus those that require tailoring. We further discuss these results below in the context of

239   their implications for improving the reproducibility and generalizability of preclinical studies.

240

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 anaesthesia 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.

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]).

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 et al. (2013) [37] investigated similar

266   components of experimental design on variation in infarct volume in rats. There are a number

267   of methodological differences between their analyses and ours (e.g. differences in size of

268   dataset and use of formal meta-analytic models). Despite these differences our quantitative

269   results are largely concordant. Where we differ substantially is in interpretation of what is a

270   desirable outcome. Ström et al. (2013) [37] concluded that intraluminal filament procedures

271   are optimal as they generate minimal variation in disease outcome and maximise statistical

272   power. Our analyses also identify that filament methods have low variation (CV = 17.9%),

273   however, we argue that these gains in statistical power come at the cost of reduced

274   reproducibility.

275

276   Considering genetic factors, proposals to include more heterogeneous study samples

277   recommend the inclusion of both sexes over just male or female animals [49-51], as well as

278   the use of multiple strains of inbred-mice and rats (or even, multiple species) [27, 52, 53].

279   Recent meta-analyses of variability in male and female rodents show that males may be as or

280   more variable than females in their phenotypic response [54, 55]. We also find that male (CV

281   = 23.5%) and female (CV = 23.9%) rats generate quantitatively equal amounts of variability,

282   but counterintuitively find that studies that used both sexes produce the most consistent

283   outcomes (CV = 17.3%; see S1 Table for full model coefficients). We caution that a

284   moderate amount of the total heterogeneity remained unexplained (i.e. residual variation;

285   Table 1), and thus these outcomes of sex on estimates of variability may be due to

286   confounding effects of unaccounted for differences in experimental design. We therefore

287   emphasize the importance of considering both genetic and environmental parameters for

288   effective heterogenization of studies [56, 57].

289

290 An alternative approach to heterogenization of experimental designs within studies is to

291 introduce variability by conducting experiments across multiple research laboratories (i.e.,

292 multi-laboratory approach) [20, 24, 58]. Importantly, such an approach inherently captures

293 'unaccounted' sources of variability in experimental conditions that are difficult to

294 systematically manipulate within a single centre study [16, 19]. We argue that, especially

295 where logistical constraints may hinder multi-laboratory approaches (e.g., for earlier, basic

296 and exploratory studies), introducing heterogeneity within studies may provide the most

297 practical alternative [23]. Indeed, by meta-analysing the variability introduced by differences

298 in experimental methodology across studies, we can begin to find ways in which to

299 heterogenize single studies in order to best capture the variation that exist across laboratories

300 and studies [16; 20].

301

302 Systematically introducing variability into a system comes at the cost of reduced statistical

303 sensitivity [8, 9] and necessitates larger studies [8, 26, 29]. These economic and ethical costs

304 must, of course, be minimised, which can be done by identifying the most efficient means of

305 introducing heterogeneity within experiments. It is therefore necessary to quantify the amount

306 of variability that different experimental designs introduce, with the aim that researchers can

307 then make informed decisions about how to most efficiently incorporate heterogeneity into

308 study design [14-16, 20]. Identifying sources of variability through meta-analysis of variance

309 in existing animal data as we have done here is the most practical and economic way of

310 establishing this much needed knowledge base.

311

312 **Quantify variability to improve drug translation**

313 Our second approach of simultaneously assessing both the mean and variation in treatment

314 outcomes allows us to place potentially useful treatments into two, distinct categories for

315   further exploration: 1) beneficial and generalizable interventions, which are those that

316   consistently reduce infarct volume across individuals and; 2) beneficial but non-generalizable

317   interventions, which on average reduce infarct volume but result in large inter-individual

318   heterogeneity in outcomes. This latter group could even include treatments that do not

319   necessarily reduce mean state, but have a large enough variance response to be beneficial to

320   some [30, 43, 59].

321

322   Overall, we find that the stroke treatments in our dataset are usually effective, reducing

323   infarct volume on average by 33.1% compared to controls. Out of these effective treatments,

324   we identify four treatments that significantly reduced infarct volume but did not induce

325   significant differences in the coefficient of variation across experimental and control groups

326   (green highlights in Fig 4). Nootropic treatments reduced infarct volume on average by

327   40.8%, whilst citicoline, antibiotic and exercise treatments reduced infarct volume by around

328   27.5% to 28.8% compared to control groups. None of these treatments were estimated to

329   significantly affect the CV, although estimated effects ranged from 5.7% smaller in

330   experimental relative to controls for citicoline (highlighted with a triangle symbol in Fig 4),

331   to 21.3% to 31.9% greater for the other treatments. We emphasise that these treatments may

332   potentially be more generalizable in that the outcomes of these treatments are on average

333   favourable, and are relatively consistent at the individual level [33, 34].

334

335   Second, we identify a handful of effective treatments that on average reduce infarct volume,

336   but also generate significant amounts of variability in experimental groups (blue highlights in

337   Fig 3; see S3 Table in S1 Text for rank order of unconditional estimates in mean and

338   coefficient of variation across treatments). Of particular interest to note is that whilst

339   thrombolytics significantly increase variability in experimental groups relative to controls,

340   they are still relatively consistent in reducing mean infarct volume (on average reducing

341   infarct volume by 29.6% whilst the coefficient of variation in experimental groups is only

342   17.4% greater than controls). Out of treatments that significantly reduce mean infarct volume,

343   thrombolytics rank second in terms of its consistency in effect, with overlapping confidence

344   intervals in their effects on the coefficient of variation with those of citicoline (Fig 3).

345

346   On the other hand, hypothermia is much more effective in reducing infarct volume (on

347   average reducing infarct volume by 49.7%) but is the least consistent in doing so, estimating

348   the greatest coefficient of variation (CV is 60.0% greater in hypothermia treated groups than

349   concurrent controls). Interestingly, efforts to exploit hypothermia for stroke in clinical trials

350   have so far failed to identify a patient group who might reliably benefit [60]. Other treatments

351   that greatly reduce average infarct volume whilst increasing the variation include, for

352   example, omega-3, rho GTPase inhibitors, and oestrogen treatments. As such, whilst these

353   treatments confer a mean beneficial effect, this effect may not be generalizable across

354   animals. Any future translation into clinical trials would require tailoring with effort put in to

355   predicting response at the individual level [30]. To our knowledge, such tailoring has not

356   been attempted because a treatment with high variability (inconsistency) is less likely to be

357   statistically significant and pass the preclinical stage (even if it does improve a disease state)

358   [30, 43, 59, 61]. Our study represents the first meta-analyses to quantify both the efficacy and

359   consistency of treatment effects in animal models. We believe that this approach will forge

360   new opportunities for improving the generalizability and translation of preclinical trials by

361   embracing both the mean and variability in outcomes.

362

363   **Conclusion**

15

364 We have demonstrated how researchers can quantitatively embrace heterogeneity in

365 phenotypic outcomes with the aim of improving both the reproducibility and generalizability

366 of animal models. Prior to experimentation, researchers may design their experiments by

367 deliberately selecting methodologies that generate variability in disease-state creating a

368 heterogenous, but broadly representative back drop of disease states against which treatment

369 efficacy can be assessed [10-12]. Since the magnitude and direction of phenotypic expression

370 and outcomes are determined by the interaction of genetic and environmental contexts within

371 studies [14-16], both of these methodological factors require heterogenization in order to

372 avoid context-specific and irreproducible outcomes across studies [16]. Post-experimentation,

373 studies may further incorporate analyses that estimate the magnitude and direction of

374 variability generated by treatments to identify potentially generalizable versus non-

375 generalizable approaches. Recent meta-analyses of variability in phenotypic outcomes of

376 animal models are beginning to illuminate the potential use of embracing different types of

377 heterogeneities for improving reproducibility, generalizability, and translation [61-63]. We

378 offer that comparative analyses of variability in both control and treatment groups has the

379 potential to inform experimental design and lead to changes in both the approach and

380 direction of follow-up studies, ultimately leading to a more successful program of

381 reproducibility, drug discovery and translation.

382

Materials and methods

Data collection and imputation

385 We identified studies of rat animal models for stroke from the CAMARADES electronic

386 database. For our analysis, we only included experimental studies that reported mean infarct

387 volume (and their associated standard deviation and sample size) in both control and

388 experimental groups. Where necessary we calculated the standard deviation from the standard

389   error multiplied by the square root of $(n-1)$, where n is the sample size of the control or

390   experimental group. Furthermore, when a study used multiple treatment groups for a control

391   group, we divided the sample size of the control group equally amongst the treatment groups,

392   which dealt with correlated errors and prevented sampling (error) variances being overly

393   small [64]. Before calculating the effect sizes, we excluded data where: (i) the standard error

394   was reported as zero; or (ii) the sample size of the control group when divided was equal to or

395   less than one. We also excluded categorical predictors that were represented by fewer than

396   five data points.

397

398   For meta-analysis of variance across methodological parameters, we focused on control

399   groups and only included data from studies that provided sufficient group-level information

400   on the methodology of the experiment. Specifically, we collected and coded methodological

401   predictors as closely as possible to the predictors used by Ström et al. (2013) [37] to produce

402   a comparable meta-analysis (see full model parameters in S1 Table in S1 Text). For meta-

403   analysis of variance across drug treatment, we included data from studies that provided

404   sufficient group-level information on the drug group, rat strain, and sex of

405   experimental/control groups (see full model parameters in S3 Table in S1 Text). For all

406   analyses, we dealt with missing data via multiple imputation [65, 66] using the package *mice*

407   [67] as follows: We first generated multiple, simulated datasets ($m = 20$) by replacing missing

408   values with possible values under the assumption that data are missing at random (MAR) [66,

409   78]. After imputation, meta-analyses were performed on each imputed dataset (as described

410   in *Statistical Analysis*) and model estimates were then pooled across analyses into a single set

411   of estimates and errors.

412

413   **Calculating effect sizes**

17

414   For meta-analysing variance across methodological predictors we calculated the log

415   coefficient of variation (lnCV) and its associated sampling variance (s^2_{lnCV}) for each control

416   group. Since many biological systems appear to exhibit a relationship between the mean and

417   the variance on the natural scale (i.e., Taylor's Law; [42, 69]), an increase in the mean may

418   correspond to an increase in variance. Our data indeed appears to exhibit a positive

419   relationship between log standard variation (lnSD) and log mean infarct volume (S1 Fig).

420   When such a relationship holds in data it may be most preferable to use an effect size such as

421   lnCV, which estimates variance accounting for the mean, and this is the approach we have

422   taken.

423

424   For meta-analysing variance across drug treatments, we calculated the log coefficient of

425   variance (lnCVR) and its associated sampling variance (s^2_{lnCVR}) as given in equations (11)

426   and (12) in Nakagawa et al. (2015) [70] (S7 Table in S1 Text). When meta-analysing

427   variance in the presence of Taylor's Law as it appears in our dataset, it may be most

428   preferable to use lnCVR (over the log variance ratio, lnVR), which gives the variance of a

429   contrast group accounting for differences in the mean. We therefore report all results using

430   lnCVR in the manuscript. We note, however, that both lnCV and lnCVR assumes a linear

431   relationship between the mean and variance on the natural scale, whilst Taylor's law states a

432   power relationship. In addition to assessing the effects of treatments on variance, we further

433   quantified differences in mean infarct volume by calculating the log response ratio of the

434   mean for each control/experimental group within a study (lnRR) and its associated sampling

435   variance (s^2_{lnRR}). For both lnRR and lnCVR we calculated effect sizes so that positive values

436   corresponded to a larger mean or variance in the experimental group.

437

438   **Statistical analysis**

439   We implemented multilevel meta-analytic models in a likelihood-based package using the

440   function 'rma.mv' in the *metafor* package [71] as described in equation 1:

441    $$y_{ij} = \mu + \beta x_{ij} + s_j + t_j + e_{ij} + m_{ij} \qquad \text{eqn 1}$$

442   where, y_{ij} (the ith effect size of variability or mean infarct volume from a set of n effect sizes

443   ($i = 1, 2, \dots, n$) in the jth study from a set of k studies $j = 1, 2, \dots, k$) is given by the grand

444   mean (μ), the effects of fixed predictors (βx_{ij}), and random effects due to study (s_j), strain

445   (t_j), residual (e_{ij}) and measurement error (m_{ij}) for the ith effect size in the jth study. Since

446   variability in observed effects may be explained by measurement error (m_{ij} in equation 1),

447   we present total I^2 (the percentage of variance that cannot be explained by measurement

448   error) and study I^2 (the percentage of variance explained by study-effects) to estimate the

449   true variance in observed effects (i.e. meta-analytic heterogeneity) [72]. We interpreted I^2 of

450   25%, 50% and 75% as small, medium, and large variance, respectively [72].

451

452   To estimate variance (lnCV) in outcome as a function of methodology in control groups we

453   constructed two meta-analytic models. First, we fitted a multilevel meta-analysis (MLMA)

454   with the objective of estimating the overall average variability in infarct volume across

455   studies. MLMA included a fixed intercept and random effects described in equation 1.

456   Second, we fitted a multilevel meta-regression (MLMR) with the objective of estimating

457   effects of methodological predictors on variability in infarct volume, by fitting the following

458   fixed predictors: (i) method of occlusion, (ii) sex of animal cohort, (iii) type of ischaemic

459   model, (iv) type of anaesthetic, (v) whether experiments were temperature controlled, (vi)

460   whether rats were physiologically monitored, (vii) mean cohort weight, and (viii) time for

461   evaluation of damage after focal ischaemia (S1 Table in S1 Text). Mean cohort weight and

462   time for evaluation were z-transformed prior to model fitting. We similarly constructed

463   MLMA and MLMR models for lnRR and lnCVR (fitting each effect size as the response in

464   separate models), to estimate the mean and variance in outcome as a function of drug

465   treatment in our control/experimental groups, respectively. For these MLMR models, we

466   included (i) drug treatment group, and (ii) sex of animal cohort as fixed predictors (S3 Table

467   in S1 Text).

468

469   Fixed effects were deemed statistically significant where their 95% credible intervals (CIs)

470   did not span zero. For interpretation of results, we back-transformed model estimates from

471   the log to the natural scale. Finally, we tested for signs of publication bias (systematic bias in

472   the published data due to the preferential publication of more significant results) in our data

473   by visual inspection of funnel plots (S2 Fig) and conducting a type of Egger regression

474   (precision-effect test and precision-effect estimate with standard errors, PET-PEESE) on

475   lnRR [73] (see S8 Table in S1 Text for publication bias test results). Egger regression cannot

476   be used for lnCVR, and further, it is unlikely that publication bias occurs for lnCVR because

477   such biases are not driven by the difference in standard deviations between the experimental

478   and control groups [74]. All meta-analyses were conducted using the 'rma.mv' function in

479   the likelihood-based package *metafor* [71], on the statistical programming environment R (v

480   3.2.2 [75]).

481

Acknowledgements

485

References

487   1. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean?
488   Sci Transl Med. 2016; 341: 96–102.
489   2. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2:

490   696–701.

491   3. Begley CG, Ioannidis JPA. Reproducibility in science: improving the standard for
492   basic and preclinical research. Circ Res. 2015; 116: 116–126.

493   4. Frye SV, Arkin MR, Arrowsmith CH, Conn PJ, Glicksman MA, Hull-Ryde EA, et
494   al. Tackling reproducibility in academic preclinical drug discovery. Nat Rev Drug
495   Discov. 2015; 14: 733–734.

496   5. Baker M. Is there a reproducibility crisis? A Nature survey lifts the lid on how
497   researchers view the crisis rocking science and what they think will help. Nature.
498   2016; 533: 452–455.

499   6. Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. Nat
500   Rev Neurol. 2014; 10: 37–43.

501   7. Seyhan AA. Lost in translation: the valley of death across preclinical and clinical
502   divide – identification of problems and overcoming obstacles. Transl Med
503   Commun. 2019; 4(18). doi.org/10.1186/s41231-019-0050-7

504   8. Festing MF. Reduction of animal use: experimental design and quality of
505   experiments. Lab Anim. 1994; 28: 212–221.

506   9. Beynen AC, Baumans V, Van Zutphen LFM. Principles of Laboratory Animal
507   Science. Amsterdam: Elsevier; 2001.

508   10. Würbel H. Behaviour and the standardization fallacy. Nat Genet. 2000; 26: 263.

509   11. Richter SH, Garner J P, Würbel H. Environmental standardization: cure or cause of
510   poor reproducibility in animal experiments? Nat Methods. 2009; 6: 257–261.

511   12. Richter SH. Systematic heterogenization for better reproducibility in animal
512   experimentation. Lab Anim. 2017; 46: 343–349.

513   13. Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with
514   laboratory environment. Science. 1999; 284: 1670–1672.

515   14. Voelkl B, Würbel H. Reproducibility crisis: are we ignoring reaction norms? Trends
516   Pharmacol Sci. 2016; 37: 509–510.

517   15. Karp NA. Reproducible preclinical research – is embracing variability the answer?
518   PLoS Biol. 2018; 16: e2005413. doi.org/10.1371/journal.pbio.2005413

519   16. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et
520   al. Reproducibility of animal research in light of biological variation. Nat Rev
521   Neurosci. 2020; 21: 384–393.

522   17. Chesler EJ, Wilson SG, Lariviere WR, Rodriguez-Zas SL, Mogil JS. Influences of
523   laboratory environment on behavior. Nat Neurosci. 2002; 5: 1101–1102.

524   18. Mueller FS, Polesel M, Richetto J, Meyer U, Weber-Stadlbauer U. Mouse models of
525   maternal immune activation: mind your caging system! Brain Behav Immun. 2018;
526   73: 643–660.

527   19. Richter SH, Garner JP, Zipser B, Lewejohann L, Sachser N, Touma C, et al. Effect of
528   population heterogenization on the reproducibility of mouse behavior: a multi-
529   laboratory study. PLoS One. 2011; 6: e16461. doi:10.1371/journal.pone.0016461

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

533   21. Kafkafi N, Golani I, Jaljuli I, Morgan H, Sarig T, Würbel H, et al. Addressing
534   reproducibility in single-laboratory phenotyping experiments. Nat Methods 2017; 14:
535   462–464.

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

23. Karp NA, Speak AO, White JK, Adams DJ, de Angelis MH, Hérault Y, Mott RF. Impact of temporal variation on design and analysis of mouse knockout phenotyping studies. PLoS ONE. 2014; 9: e111239. doi.org/10.1371/journal.pone.0111239

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.

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). doi.org/10.1126/scitranslmed.aaa9853

26. Festing MF. Refinement and reduction through the control of variation. Altern Lab Anim. 2004; 32: 259–263.

27. Festing MF. Evidence should trump intuition by preferring inbred strains to outbred stocks in preclinical research. ILAR J. 2014; 55: 399–404.

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.

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.

31. Dirnagl U. Bench to bedside: The quest for quality in experimental stroke research. J Cerebr Blood F Met. 2006; 26(12): 1465–1478.

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.

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.

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.

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.

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.

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. doi.org/10.1186/1471-2202-14-41

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. doi.org/10.1038/srep21086

39. Van der Worp HB, Van Gijn J. Clinical practice. Acute ischemic stroke. N Engl J Med. 2007; 357: 572–579.

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.

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 Meth. 2014; 221: 92–102.

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.

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.

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

46. Haast RAM, Gustafson DR, Kiliaan AJ. Sex differences in stroke. J Cerebr Blood F Met. 2012; 32(12): 2100–2107.

47. Turtzo LC, McCullough LD. Sex-specific responses to stroke. Future Neurol. 2010; 5(1): 47–59.

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.

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.

50. Clayton JA, Collins FS. NIH to balance sex in cell and animal studies. Nature. 2014; 509(7500): 282–283.

51. Clayton JA. Applying the new SABV (sex as a biological variable) policy to research and clinical care. Physiol Behav. 2018; 187: 2–5.

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.

54. Prendergast BJ, Onishi KG, Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. Neurosci Biobehav Rev. 2014; 40: 1–5.

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. doi.org/10.1186/s13293-016-0087-5

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.

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.

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

59. Naylor S, Chen JY. Unraveling human complexity and disease with systems biology and personalized medicine. Pers Med. 2010; 7(3): 275–289.

60. 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.

61. 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.

62. Brugger SP, Angelescu I, Abi-Dargham A, Mizrahi R, Shahrezaei V, Howes OD. Heterogeneity of striatal dopamine function in schizophrenia: meta-analysis of variance. Biol Psychiat. 2020; 87(3): 215–24.

63. Kuo SS, Pogue-Geile MF. Variation in fourteen brain structure volumes in schizophrenia: A comprehensive meta-analysis of 246 studies. Neurosci Biobehav Rev. 2019; 98: 85–94.

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.

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

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

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

69. Cohen JE, Xu M. Random sampling of skewed distributions implies Taylor's power law of fluctuation scaling. Proc Natl Acad Sci. 2015; 112(25): 7749–7754.

70. 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.

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

72. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21(11): 1539–1558.

73. Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. Res Synth Methods. 2014; 5(1): 60–78.

74. 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.

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

Supporting information

S1 Text. Supporting information including tables of full model coefficients, effect

size/sampling variance equations, and publication bias results. (PDF)

S1 Fig. Scatter plot of mean-variance (SD) relationship in rat animal data. Point estimates for

control (blue) and treatment (yellow) groups are provided, as well as their slope of linear

regressions for control and experimental rat groups, respectively. Note that data points are not

represented in the same units. (PDF)

679   **S2 Fig.** Funnel plot for log response ratio (lnRR) characterizing differences in mean infarct

680   volume for control/treatment groups. Raw effect sizes are plotted against their precision

681   (inverse of the square root of standard error). MLMA-model predicted mean effect size (solid

682   vertical line) and its 95% CI (dashed lines) are shown. (PDF)

683   **S1 Data.** Data files for analysis of lnCV, lnRR and lnCVR in infarct volume, extracted from

684   CAMARADES database. (RDS)

685   **S1 Code.** R code for conducting meta-analyses. (R-CODE)

686

Author contributions

688   **Conceptualization:** Shinichi Nakagawa, Alistair Senior, Takuji Usui

689   **Data curation:** Malcolm Macleod, Sarah McCann, Takuji Usui

690   **Formal analysis:** Alistair Senior, Takuji Usui

691   **Funding acquisition:** Shinichi Nakagawa, Alistair Senior

692   **Supervision:** Shinichi Nakagawa, Alistair Senior

693   **Writing – original draft:** Takuji Usui

694   **Writing – review & editing:** Malcolm Macleod, Sarah McCann, Shinichi Nakagawa, Alistair

695   Senior, Takuji Usui

696

697   **Fig. 1. The effects of methodological parameters on variability (CV) in infarct volume**

698   **across control groups.** Mean estimates of unconditional (marginalized), group-specific

699   coefficients of variation (%) are indicated as grey circles whilst the overall estimate is

700   indicated as a grey diamond. 95% CIs are shown as grey lines and are asymmetric due to

701   back-transformation of log coefficient of variation (lnCV) to the natural scale. Spontaneous

702   occlusion generated the highest estimate of variability as indicated by the arrowhead. The

703   overall and group-specific estimates were obtained from multilevel meta-analysis (MLMA)

704   and multilevel meta-regression (MLMR) models, respectively.

705

706   **Fig. 2. The effects of drug treatments on the difference in: (a) mean (lnRR); and (b)**

707   **variability (lnCVR) in infarct volume across control and experimental rat groups.** Mean

708   estimates of unconditional (marginalized), group-specific effects are shown as grey circles

709   whilst the overall estimate is indicated by the grey diamonds. 95% CIs are shown as grey

710   lines. Negative lnRR estimates indicate that mean infarct volume is smaller in experimental

711   versus control rats. Negative lnCVR estimates show that inter-individual variability in infarct

712   volume is smaller in experimental versus control rats (e.g. HBOT indicated by left-pointing

713   arrowhead) whilst positive lnCVR estimates show that variability in infarct volume is greater

714   in experimental versus control rats (e.g. angiotensin receptor blockers (ARB) indicated by

715   right-pointing arrowhead). The overall and group-specific estimates were obtained from

716   multilevel meta-analysis (MLMA) and multilevel meta-regression (MLMR) models,

717   respectively.

718

719   **Fig. 3. Categorization of treatment effects based on mean efficacy (lnRR) and inter-**

720   **individual variability in efficacy (lnCVR).** Estimates (circles) represent unconditional

721   (marginalized), treatment-specific means (lnRR), variability (lnCVR), and their 95% CIs

722   (solid lines) obtained from multilevel meta-regression (MLMR) models. Treatments that

723   significantly reduce infarct volume (negative lnRR) without significantly affecting the

724   variation are highlighted green, with citicoline indicated by a diamond as the only treatment

725   to significantly reduce infarct volume and also have a negative point estimate of lnCVR.

726   Treatments that significantly reduce infarct volume and increase inter-individual variability

727   (positive lnCVR) are highlighted blue. The effects of hypothermia (most negative and

26

728   positive mean and variability estimates, respectively) and thrombolytics (which include the

729   only regulatory approved treatment) are highlighted in pink. Histograms show the

730   relationship of the mean and variance in infarct volume between control (orange) and

731   treatment (blue) groups in each quadrant of the graph.

Sex
Both
Female
Male

Occlusion method
Collagenase
Direct/Mechanical
Embolic
Endothelin
Filamental
Photothrombosis
Spontaneous

Occlusion model
Temporary
Thrombotic

Anesthesia
Barbiturates
Inhalation
Ketamine

Temperature
Controlled
Uncontrolled

Physiology
Monitored
Unmonitored

Overall

CV (%)

(a)

Sex
Both
Female
Male

Drug Group
Angiotensin receptor blocker (ARB)
Anti-inflammatory
Antibiotic
Antidepressant
Antioxidants
Citocholine
Environment
Estrogen
Exercise
Growth Factor
HBOT
HMG-CoA reductase antagonist
Hypothermia
Immunosuppressant
Mixed training
MK801
Nootropic
NO donor
NOS inhibitor
Omega-3
PPAR-gamma agonist
Rho GTPase inhibitor
Stem cells
Thrombolytics
Training
Vitamin
Overall Effect

-1.0 -0.5 0.0 0.5 1.0

lnRR

(b)

-1.0 -0.5 0.0 0.5 1.0

lnCVR