



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Advanced Methods and Implementations for the Meta-analyses of Animal Models

**Citation for published version:**

Yang, Y, Macleod, M, Pan, J, Lagisz, M & Nakagawa, S 2022, 'Advanced Methods and Implementations for the Meta-analyses of Animal Models: Current Practices and Future Recommendations', *Neuroscience & Biobehavioral Reviews*, pp. 105016. <https://doi.org/10.1016/j.neubiorev.2022.105016>

**Digital Object Identifier (DOI):**

[10.1016/j.neubiorev.2022.105016](https://doi.org/10.1016/j.neubiorev.2022.105016)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Peer reviewed version

**Published In:**

Neuroscience & Biobehavioral Reviews

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

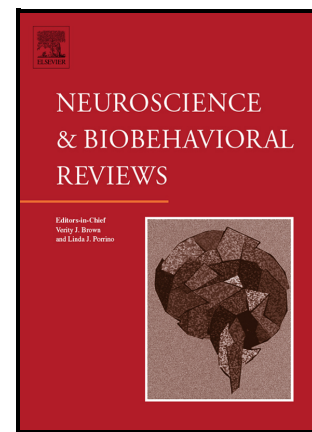
**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



Advanced Methods and Implementations for the  
Meta-analyses of Animal Models: Current Practices  
and Future RecommendationsRunning title:  
Advanced meta-analysis methods

Yefeng Yang, Malcolm Macleod, Jinming Pan,  
Malgorzata Lagisz, Shinichi Nakagawa



PII: S0149-7634(22)00505-X  
DOI: <https://doi.org/10.1016/j.neubiorev.2022.105016>  
Reference: NBR105016

To appear in: *Neuroscience and Biobehavioral Reviews*

Received date: 21 June 2022  
Revised date: 19 December 2022  
Accepted date: 20 December 2022

Please cite this article as: Yefeng Yang, Malcolm Macleod, Jinming Pan, Malgorzata Lagisz and Shinichi Nakagawa, Advanced Methods and Implementations for the Meta-analyses of Animal Models: Current Practices and Future RecommendationsRunning title: Advanced meta-analysis methods, *Neuroscience and Biobehavioral Reviews*, (2022) doi:<https://doi.org/10.1016/j.neubiorev.2022.105016>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

*Running title: Advanced meta-analysis methods*

**Advanced methods and implementations for the meta-analyses of animal models:  
current practices and future recommendations**

Yefeng Yang<sup>a,b,c,\*</sup>, Malcolm Macleod<sup>d</sup>, Jinming Pan<sup>b,\*</sup>, Malgorzata Lagisz<sup>a,1</sup>, Shinichi Nakagawa<sup>a,1,\*</sup>

<sup>a</sup> Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia

<sup>b</sup> Department of Biosystems Engineering, Zhejiang University, Hangzhou 310058, China

<sup>c</sup> Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China

<sup>d</sup> Centre for Clinical Brain Sciences, The University of Edinburgh, UK

\* Corresponding author.

E-mail addresses: s.nakagawa@unsw.edu.au (S. Nakagawa),

yefeng.yang1@unsw.edu.au (Y. Yang), panhouse@zju.edu.cn (J. Pan).

<sup>1</sup>Contributed equally as joint senior authors.

**Open Science**

An online hands-on tutorial on advanced meta-analytic techniques outlined in the manuscript is available at GitHub:

[https://github.com/Yefeng0920/advanced\\_animal\\_MA\\_tutorial](https://github.com/Yefeng0920/advanced_animal_MA_tutorial). Raw data and statistical code to reproduce examples presented in the manuscript is archived at Zenodo: Yefeng, & Malgorzata Lagisz. (2022).

Yefeng0920/advanced\_animal\_MA\_tutorial: A tutorial of advanced methods for the meta-analyses of animal models (v1.5.0). Zenodo.

<https://doi.org/10.5281/zenodo.7314683>.

### **Abstract**

Meta-analytic techniques have been widely used to synthesize data from animal models of human diseases and conditions, but these analyses often face two statistical challenges due to complex nature of animal data (e.g., multiple effect sizes and multiple species): statistical dependency and confounding heterogeneity. These challenges can lead to unreliable and less informative evidence, which hinders the translation of findings from animal to human studies. We present a literature survey of meta-analysis using animal models (animal meta-analysis), showing that these issues are not adequately addressed in current practice. To address these challenges, we propose a meta-analytic framework based on multilevel (linear mixed-effects) models. Through conceptualisation, formulations, and worked examples, we illustrate how this framework can appropriately address these issues while allowing for testing new questions. Additionally, we introduce other advanced techniques such as multivariate models, robust variance estimation, and meta-analysis of emergent effect sizes, which can deliver robust inferences and novel biological insights. We also provide a tutorial with annotated R code to demonstrate the implementation of these techniques.

Keywords: Research synthesis; Quantitative method; Publication bias; New effect size; Multilevel meta-analysis; Meta-regression; Multivariate meta-analysis; Systematic review; PRISMA; Animal experiment; Animal research

## 1. Introduction

Meta-analysis has made a significant contribution to many disciplines including behavioural and neurobiological sciences, playing a crucial role in quantitatively summarizing existing findings and informing evidence-based decision-making (Bannach-Brown et al., 2021; Gurevitch et al., 2018; Nakagawa et al., 2020; Sena et al., 2014; Vesterinen et al., 2014). Traditionally, meta-analyses have been used to synthesize data from human randomised controlled trials (RCT), describing the efficacy of treatments and interventions (Chalmers and Haynes, 1994; Schmid et al., 2020). More recently, there has been a surge in meta-analyses using data from animal studies modelling human diseases, physiology, and behaviour (hereafter, “animal meta-analyses”; see Greek and Menache, 2013; Hooijmans et al., 2014; Hunniford et al., 2021).

The aims of animal meta-analyses are diverse with examples including predicting the effectiveness of therapeutic strategies for a neurological disorder (Baldez et al., 2021), identifying study characteristics mediating the effectiveness of a therapy (Figueiredo et al., 2020), and explaining replication failures of preclinical studies (Usui et al.,

2021). Perhaps most importantly, high-quality animal meta-analyses can generate reliable and useful appraisals of preclinical experimental evidence, which can be used to direct human trials and inform decisions to progress to clinical applications (Bahadoran et al., 2020; de Vries et al., 2014; Soliman et al., 2020). Therefore, animal meta-analyses can serve as a useful complement to traditional medical meta-analyses (e.g., clinical meta-analysis on RCT data). Although animal meta-analyses are important and performed with increasing frequency, we know very little about their methodological rigor (Hunniford et al., 2021; Mueller et al., 2014). Indeed, as far as we know there is no systematic profiling of methodological and reporting practice of animal meta-analyses (but see Hooijmans et al., 2022).

We have predicted that two major statistical issues had not been adequately addressed in the current practice of animal meta-analyses because the meta-analytic methods (i.e., fixed- and random-effects models) inherited from human medical meta-analyses, are technically incapable of handling the two problems. The first issue is the violation of the assumption of statistical independence between effect sizes (e.g., multiple effect sizes per study/paper). Indeed, the true effects and sampling errors are inevitably correlated in animal studies that involve multiple measurements, multiple animal cohorts, multiple species, and multiple treatments (Figure 1; Aarts et al., 2014; Pound and Bracken, 2014; Vesterinen et al., 2014). Therefore, an animal meta-analysis often has a hierarchical/multilevel/nested data structure (sometimes a

multivariate structure) which might lead to the dependency among the true effect sizes (i.e., the correlation/covariance of the true effects within the same study). Importantly, the traditional statistical models used in animal meta-analyses often fail to deal with this statistical non-independence (but see Bonapersona et al., 2018; Lagisz et al., 2020), inflating type I error, and thus distorting statistical inference, leading to spurious conclusions (Cheung, 2019; Nakagawa et al., 2017; Thomas et al., 2003). The second issue relates to the first point. The hierarchical nature of the animal datasets means that animal meta-analyses could be improved, in terms of explanatory power, from decomposing variance in effect sizes across different levels, e.g., within- and between-study level heterogeneity (a more complex level: species-specific heterogeneity; Konstantopoulos, 2011; Senior et al., 2016). However, the traditional meta-analytic models treat between-experiment variability as the only source of heterogeneity, confounding between-study heterogeneity with other sources of heterogeneity (e.g., within-study between-experiment heterogeneity), and subsequently leading to less informative quantification.

Methodological innovations and advances in meta-analysis offer us pragmatic and effective strategies to deal with these two issues. The use of multilevel meta-analytic models has been developed explicitly to account for non-independent and heterogeneous effect sizes and has successfully been deployed, for example, in ecology, evolution, psychology, and education (Cheung, 2014, 2019; Nakagawa et al.,

2017; Nakagawa and Santos, 2012). The multivariate models and robust variance estimation (RVE) are also effective approaches to handling dependent effect sizes in the context of meta-analysis (Cheung, 2013; Fisher and Tipton, 2015; Hedges et al., 2010; Jackson et al., 2011; Pustejovsky and Tipton, 2022; Welz et al., 2023).

Furthermore, some underappreciated effect size statistics can inform methodological choices and provide new neurobiological insights for animal meta-analyses. For example, meta-analysing variability-based effect sizes which compare differences in variance (rather than its mean) between two groups (i.e., meta-analysis of variance; Figure 2) enables us to scale out research questions, such as investigating the individual variability in treatment response to drugs in neurological and behavioural disorders (Maslej et al., 2021; Nakagawa et al., 2015; Senior et al., 2020).

Here, we first conduct a systematic literature survey to characterize the current practice of animal meta-analyses by mapping their issues and statistical approaches. While presenting our survey results, we also summarize the traditional and emerging meta-analytic statistical procedures. Second, we illustrate the concepts and rationale of the multilevel models (only a small number of researchers have already applied these models to animal meta-analyses; see Bonapersona et al., 2018; Lagisz et al., 2020); how they can deal with non-independence among effect sizes and how multilevel meta-regression can account for multiple levels of heterogeneity. Finally, we touch upon advanced techniques, including the multivariate model and RVE robust variance



estimation. Additionally, we provide recommended practices and a hands-on workflow (with annotated R code) that can serve as a template to deal with the statistical issues commonly encountered in animal meta-analyses. Future meta-analysts can adapt our example code to undertake their own animal meta-analyses to draw more robust model inferences, generate new neurobiological insights, and better facilitate animal-to-human translation (Bahadoran et al., 2020).

## **2. Survey of current practice in animal meta-analyses**

### *2.1. Survey procedures*

The main aim of the literature survey was to capture the state of current practice in conducting meta-analyses on non-human animal data. As such, we performed a systematic literature search of neurobiology and behavioural journals to identify meta-analytic papers published in the last 10 years (2011–2021). The details of methodological procedures of the search strategy are detailed in Supplementary Materials (file 1), reported according to PRISMA guidelines (Moher et al., 2009).

Briefly, we used the ISI InCites Journal Citation Reports to collect the ISSN of journals classified under ‘Neurology’ and/or ‘Behaviour’ (including 115 journals; see details in Supplementary Materials file 1). We then searched within these included journals using the online database PubMed (last updated in October 2021) for meta-analytic papers published within the last 10 years. We restricted our searches to

studies labelled as non-human studies and as 'meta-analysis' in the database. This search yielded a total of 188 bibliographic references. Two authors (YY and ML) then performed two-stage screening to identify animal meta-analyses using the following inclusion criteria: (1) the paper addressed a question in the fields of neurology or behavioural sciences; (2) claimed to conduct a meta-analysis; and (3) used data from the animal models (i.e., in *vivo* animal studies for preclinical research) rather than exclusively human studies. There were 2% conflicting decisions, which were resolved via discussions. Finally, we included 78 papers claiming to be meta-analyses, of which 62 papers were eligible for data extraction of methodological approach.

## 2.2. Data coding

Two authors (YY and ML) independently assessed the full text of the eligible papers to retrieve the following information: (1) whether the authors conducted a formal meta-analysis of animal data (e.g., fixed-effects or random-effects statistical models, or related model, to aggregate effect sizes that were collected from animal studies), (2) the type of effect size used, (3) the type of statistical models employed, (4) whether and how the study accounted for heterogeneity, (5) the type of heterogeneity indices reported, (6) whether the authors reported the number of primary studies ( $N$ ) and effect sizes ( $k$ ), (7) whether the data used was potentially at risk of issues of statistical non-independence (e.g., if the ratio of  $N$  to  $k$  is larger than 1, or if one primary study contributes more than one effect size), (8) whether the authors of meta-

analysis acknowledged the presence of statistical non-independence, (9) whether the authors used any procedure to deal with non-independence among effect sizes (i.e., correlated effect sizes), (10) whether the authors addressed the issue of sampling errors non-independence (i.e., correlated sampling errors), (11) whether the authors used any approaches to test for publication bias, and (12) whether the authors reported the software used to perform the animal meta-analyses. In addition, we extracted the bibliometric information of each included animal meta-analysis (e.g., title, authors, the journal published, publish year). We describe the survey questions and associated options in Supplementary Materials (file 1).

## *2.2. Methodological and reporting practice in animal meta-analyses*

For the included 78 self-proclaimed animal meta-analyses, 21% (16) did not conduct formal meta-analyses of animal data (Supplementary Materials file 1). These ‘non-formal’ meta-analyses used other than formal ‘effect-size-based’ meta-analysis modelling approaches, which included three categories: (1) meta-analyses of genetics, genomics or bioinformatics (e.g., transcriptional, genome-wide association studies), (2) meta-analyses of brain imaging (e.g., fMRI), and (3) other self-proclaiming as containing meta-analyses but not fitting into any traditional (“formal”) meta-analysis framework. For the 62 “formal” meta-analyses, we coded their methodological and reporting practice. We reported the main results in Figure 3 (for the full survey results, see Supplementary Materials file 1). In the later section, we will discuss the

details of each methodological and reporting characteristic of the surveyed animal meta-analyses, in combination with the explanations of the advanced meta-analytic methods.

### **3. Effect sizes in meta-analysis of animal data**

#### *3.1. Common effect sizes in the meta-analysis of animal data*

Effect size is a statistic that can measure the size and direction of an effect (e.g., the effectiveness of an antidepressant or the prevalence of depression occurring; referred to as ‘effect statistic’; see Nakagawa and Cuthill, 2007). It also can refer to the estimate (value) of a given effect statistic (e.g., ‘effect estimate’; see Hentschke and Stüttgen, 2011). In this paper, we use the two definitions interchangeably. Our survey indicates that the effect sizes used in animal meta-analyses can be categorized roughly into four types: (1) standardized mean difference between two (experimentally created or intrinsically occurring) arms, SMD (68%, see Figure 3; e.g., ‘*d*’ family – common estimators of SMD are Cohen's *d* and Hedges' *g*; Hedges, 1982), (2) the incidence of two events (10%; e.g., 2-by-2 contingency table: odds ratio, relative risk, and risk difference), (3) strength of the association/correlation between two variables (1%; correlation coefficient, *r*, or its Fisher's *z* transformation, *Zr*), (4) others (18%; e.g., unstandardized raw mean difference, and determination coefficient,  $R^2$ ). Our survey clearly shows that the most popular effect size is the standardized mean difference, SMD (the ‘*d*’ family – Cohen's *d* or Hedges' *g*). The formulas used to calculate point

estimates of these effect sizes and their sampling variances can be found elsewhere (Hentschke and Stüttgen, 2011; Nakagawa and Cuthill, 2007; Noble et al., 2017).

### *3.2. Emerging effect sizes in the meta-analysis of animal data*

Our literature survey revealed the use of three important but underappreciated effect sizes in animal meta-analyses. The first is the normalised mean difference, NMD (9%; Figure 3), which divides the raw mean difference between two arms (the experimental group vs. control/reference group, e.g., placebo, sham, or wild-type group) by the mean of the control group (Vesterinen et al., 2014). The second is the log-transformed response ratio, lnRR (1%; Figure 3), which uses the natural logarithm of the ratio of means between two arms to measure mean difference (Figure 2; estimating the average treatment effect; Hedges et al., 1999; Lajeunesse, 2011). The third is the log-transformed variability ratio, lnVR (1%; Figure 3), which can quantify the difference in variance (standard deviation) around the mean between two arms (Figure 2; i.e., estimating inter-individual variability between two arms or heterogeneity of treatment effect; Nakagawa et al., 2015; Senior et al., 2020). The log-transformed coefficient of variation ratio (lnCVR) is a mean-adjusted version of lnVR, where the indirect impact of mean on its variability is controlled for (i.e., accounting for the mean-variance relationship; Nakagawa et al., 2015; Volkmann et al., 2020).

The use of InVR and InCVR provides the opportunity to reveal new or neglected neurobiological insights to the field. For example, meta-analysis of traditional effect sizes (e.g., SMD) mainly focuses on how therapy can mitigate neurobiological- and behavioural- disorders (i.e., detecting average treatment effect; Mills et al., 2021). In contrast, meta-analysis of variation (e.g., InVR and InCVR) can examine inter-individual variability in response to treatment in mental symptoms (i.e., detecting heterogeneous treatment effect; Hieronymus et al., 2020). This means meta-analysis of variation can be used to examine whether a treatment shows a consistent (in all animals: low inter-subject variability) or selective (in some animals: high inter-subject variability) efficacy (Usui et al., 2021). Examining such inter-subjective variability brings important implications to precision and personalized medicine and behavioural therapies (Lorenzo-Luaces et al., 2021; Luedtke and Kessler, 2021). For example, if a treatment response manifests high inter-subjective variability, it indicates the treatment effect may be subjective-specific and warrants further research on the sources of variability and personalized prescriptions in the clinic practice (Haggarty et al., 2021; Schork, 2015).

### *3.3. The choice of effect sizes in the meta-analysis of animal data*

In general, the choice of effect sizes is straightforward because the types of effect sizes should be aligned with the scientific question or hypothesis of a meta-analysis (Nakagawa et al., 2017). To put it another way, the goal of a given animal meta-

analysis will generally determine which types of effect sizes should be used. For example, if interested in the effectiveness of a given intervention, researchers might consider effect sizes such as SMD (e.g., mean differences between antidepressants and placebo group expressed in the units of standard deviation) or lnRR (e.g., % increase in mean in antidepressants compared with placebo). If the goal is to examine the external validity of animal behavioural assays (e.g., forced swimming or Morris water maze), lnVR or lnCVR (when a mean-variance relationship exists) is preferable because an optimal animal assay should have less inter-individual variability so that it is more reproducible and generalizable.

Recent empirical and simulation work from ecology and animal sciences indicated lnRR and NMD are in general more statistically powerful than SMD and less vulnerable to overestimation (Wang et al., 2018; Yang et al., 2022a; Yang et al., 2022b). In conjunction with other merits (e.g., better interpretability, heteroscedasticity-robustness, scale-independence), lnRR and NMD provide ‘animal’ meta-analysts with a complement to SMD when mean difference is the focal interest (Sánchez-Tójar et al., 2020; Spake et al., 2021). But note that NMD and lnRR are only applicable for ratio-scale data (i.e., bounded at zero, for example, brain size; Houle et al., 2011; Nakagawa et al., 2015). Also, SMD represents additive effects while NMD and lnRR indicate multiplicative effects. Therefore, the dual use of these two types of effect statistics may be advisable whenever possible (e.g., SMD as the

main effect size; InRR for sensitivity analysis; see Supplementary Materials file 2 for a real example).

#### **4. The prevalence of statistical non-independence in the meta-analysis of animal data**

As mentioned, multiple effect size estimates from the same study can lead to non-independence, which violates the assumption of statistical independence of data in traditional meta-analytic models (also see *section 6.1*). Non-independent effect sizes mean that they contribute similar information to the fitted model, which requires each effect size should contribute unique information. Our survey has shown that the issue of non-independence was near ubiquitous in animal meta-analyses: 89% (55/62) of animal meta-analyses used more than one effect size from the same study (Figure 3). Multiple effect sizes originating from the same study might be statistically dependent (i.e., correlated) because they share the same subjects, methodologies, instruments, measurement procedures or other contexts that might induce correlations. Multiple effect size estimates from the same study introduce two types of non-independencies: correlation in the true underlying effects and the observed effect size estimates (i.e., sampling errors). We summarize the common scenarios of non-independence in animal meta-analyses in Figure 4. The general causes are: (1) shared study identity – multiple effect sizes are derived from a single study (e.g., traits measured repeatedly at follow-up times or separately for males and females; Abbott et al., 2019; Bird et al., 2016), (2) shared animal identity – multiple traits, outcomes, or endpoints measured



on a common set of animals (e.g., measuring depression using forced swim test and tail suspension test for the same cohort; Barha et al., 2017; Burgueno et al., 2020), (3) shared control – multiple treatments compared to a common control group (e.g., multiple trial arms; England et al., 2015; Neville et al., 2020), and (4) shared species (or strains) or evolutionary history – species similarities due to shared evolutionary history (e.g., genetic similarities and phylogenetic relatedness; Khorshidi et al., 2021; Lagisz et al., 2020; Zoerle et al., 2012). Scenarios 1 – 4 can result in correlations of the true underlying effects. At the same time, scenarios 2 and 3 also lead to correlations of the correlated sampling errors because of the ‘overlapping’ animals or repeated use of the same animals when computing effect size estimates (see Figure 4 and *section 9.2*).

## **5. Approaches to handle statistical non-independence in the meta-analysis of animal data**

We have found that three broad strategies were used when non-independence was encountered in animal meta-analyses (Figure 5): (1) ignoring non-independence (Lages et al., 2021; Mancini et al., 2020), (2) eliminating non-independence (e.g., averaging or sampling; Currie et al., 2019; Frantzias et al., 2011), and (3) modelling dependence explicitly (Bonapersona et al., 2018; Creutzberg et al., 2021). The first approach was the most used method employed in animal meta-analyses (48%; Figure 3), which is concerning as it ignores the dependency among effect sizes and treats them as if they were statistically independent (i.e., via the use of simple fixed- and

random-effects models for analysing non-independent effect sizes). For example, Lages et al., 2021 included 17 effect sizes from six papers but the review authors used a random-effects model in their analyses without accounting for any non-independence.

The second common strategy in the current practice of animal meta-analyses (29%; Figure 3) also uses a standard meta-analytic model but after aggregating effect sizes or sampling a single effect size from each study to eliminate statistical non-independence of data (Figure 5). For example, Egan et al. (2014) used a fixed-effect meta-analysis for studies having multiple effect sizes to obtain a ‘synthetic’ effect size per study (also see Frantzias et al., 2011). The third method, the least common (11%, seven animal meta-analyses), was to explicitly account for dependence among effect sizes or sampling variances using a multilevel model with a suitable random-effects structure. We elaborate on this technique in *Section 6*.

Using the first and second approaches to account for non-independence in data is not necessarily incorrect, but they have obvious pitfalls (Cheung, 2014; Nakagawa et al., 2021c; Song et al., 2020). In brief, ignoring non-independence (the first approach) does not necessarily overestimate the model coefficients (e.g., pooled effect size or overall/mean effect or grand mean) but could bias the estimate of model coefficients and underestimated standard errors and, distort the subsequent hypothesis tests with

the non-nominal level of Type I error rates and coverage rates of confidence intervals (Cinar et al., 2021a; Table 1). Eliminating non-independence (the second approach) could reduce the statistical power and precision of the model coefficients due to the loss of “sample size” (in this case, the number of effect sizes), and limit the capacity to ask new questions, for example, investigating the drivers of effect size heterogeneity because of the loss of information (explanatory variables, predictors or moderators; Cheung, 2014; Nakagawa et al., 2021c; Song et al., 2020; see Figure 5). Importantly, neither ignoring nor eliminating non-independence can deal with a more complex non-independence issue such as phylogenetic relatedness (or phylogenetic correlation), which arises when incorporating data from multiple species (Nakagawa and Santos, 2012; Noble et al., 2017). Because species with a shared evolutionary history appear to be more similar to each other, leading to the effect sizes derived within the same species are not independent (Figure 4D; Hadfield and Nakagawa, 2010). We suppose that the widespread use of ‘ignoring’ and ‘eliminating’ strategies is caused in part by the low uptake of the suitable easy-to-implement methods that can directly model non-independence and practitioners might prefer simple methods (e.g., random-effects model) or software (e.g., Review Manager) they are familiar with, although the lack of awareness of the importance of accounting for non-independence may make a major contribution. For example, Stukalin et al., 2020 treated one experiment with different antidepressant doses as two or more independent experiments in their meta-analytic models.

## 6. Modelling non-independence using multilevel models

### 6.1. Fixed- and random-effects meta-analytic model

In animal meta-analytic practice, the random-effects model was the most commonly employed (50%; Figure 3). As noted above, however, the random-effects model is not capable of accounting for non-independent data. In this section, we re-formulate the traditional meta-analytic model as a multilevel model to better capture the hierarchical data structures in animal studies (Figure 6), such that we can explicitly model the non-independence. Following the conventions in animal meta-analytic practice, the random-effects model can be written as (Vesterinen et al., 2014):

$$ES_j = \beta_0 + u_j + e_j, (1)$$

$$u_j \sim \mathcal{N}(0, \tau^2), \quad e_j \sim \mathcal{N}(0, v_j)$$

where  $ES_j$  = an estimated/observed effect size for study  $j$  ( $j = 1, \dots, N$ ;  $N$  is the number of animal studies included in a meta-analysis; per animal study contributes one  $ES_j$ );  $ES_j$  can be any effect size type commonly used in animal meta-analyses (e.g., SMD; see **section 3**);  $\beta_0$  = an intercept denoting average true effect/pooled effect size (also known as the global estimate, overall mean or meta-analytic mean);  $u_j$  = a random effect corresponding to the between-study effect for study  $j$ , whose mean  $E[u_j] = 0$  and variance  $\text{Var}[u_j] = \tau^2$  ((i.e., assumed to be normally distributed with mean 0 and variance  $\tau^2$ );  $e_j$  = an error term corresponding to the sampling error effect for study  $j$ , where  $\text{Var}[e_j] = v_j$  is the sampling variance, which is assumed to be known and

estimated, usually, from formulas (e.g.,  $1/(n-3)$  for  $Zr$  with  $n$  = the number of subjects; see Nakagawa and Cuthill, 2007; Noble et al., 2017, for formulas to estimate  $v_j$  for the common effect sizes). The fixed-effects (common-effects or equal-effects) model (10% of the surveyed studies; Figure 3) can be considered to be a special case of the random-effects model; that is, the random-effects model assumes heterogeneous between-study effects ( $\tau^2 > 0$ ) and the fixed-effects model assumes homogeneous between-study effects ( $\tau^2 = 0$ ;  $ES_1 = \dots = ES_N \equiv \beta_0$ ). Notably, although the fixed-effects model is appropriate on some occasions, most meta-analyses assume non-zero heterogeneity in the underlying effects (across animal cohorts, species, and settings; IntHout et al., 2016; Senior et al., 2016).

### 6.2. Multilevel meta-analytic models

Importantly, both fixed- and random-effects models assume  $N = k$  (the number of studies is the same as the number of effect sizes included in a meta-analysis). In animal meta-analytic practice,  $k$  is usually larger than  $N$ . In our survey results, the medium  $N$  was 25 (range = 2 – 414, mean = 50; Supplementary Materials file 1), while the medium  $k$  was 56 (range = 5 – 3288, mean = 203). The medium  $k$  to  $N$  ratio was 2 (range = 1 – 28, mean = 3), indicating that most surveyed meta-analyses extracted multiple effect sizes per study and thus have the issue of statistical non-independence.

A traditional meta-analytic model can be reformulated as a multilevel model to handle the issue of statistical non-independence (Cheung, 2014; Nakagawa and Santos, 2012; Van den Noortgate et al., 2013). In our survey, only seven animal meta-analyses employed the multilevel meta-analytic model. Such a model can be formulated as (i.e., three-level meta-analysis):

$$ES_{[i]} = \beta_0 + u_{b[j]} + u_{w[i]} + e_{[i]}, \quad (2)$$

$$u_{b[j]} \sim \mathcal{N}(0, \sigma_b^2), \quad u_{w[i]} \sim \mathcal{N}(0, \sigma_w^2), \quad e_{[i]} \sim \mathcal{N}(0, v_{[i]})$$

where  $ES_i$  = an effect size estimate; note that for a given animal study  $j$  ( $j = 1, \dots, N$ ), it often contains multiple effect sizes  $i$  ( $i = 1, \dots, k; k > N$ );  $\beta_0 = E[ES_{[i]}]$ , an intercept, representing the overall / pooled mean, (i.e., average true effect – the true effect sizes across  $N$  studies follow a distribution with mean  $\beta_0$ );  $u_{b[j]}$  = a random-effects term corresponding to between-study effect for study  $j$  applied to effect size  $i$  (the subscript ‘b’ denotes ‘between-study’), which captures study-specific heterogeneity  $\text{Var}[u_{b[j]}] = \sigma_b^2$  (whose magnitude is determined by the degree of inconsistency between true effect sizes in study  $1, \dots, N$ );  $u_{w[i]}$  = a random-effects term corresponding to within-study effect for effect size  $i$  in study  $j$  (the subscript ‘w’ denotes ‘within-study’), which captures effect-size-specific heterogeneity  $\text{Var}[u_{w[i]}] = \sigma_w^2$  (whose value is defined by the degree of inconsistency between multiple effect sizes in study  $j$ ; also referred to as residual heterogeneity);  $e_{[i]}$  = sampling error in effect size  $i$  in study  $j$ , which captures sampling variance  $\text{Var}[e_i] = v_i$  (if sampling errors are correlated (where the same cohort animal is repeatedly used for effect size computation;

scenarios 2 and 3 in Figure 4),  $v_i$  will become a variance and covariance matrix; details see *Section 9.3*). Because the only model coefficient in Equation 2 is the intercept,  $\beta_0$  (the estimate of the overall effect/pooled effect size/grand mean), we usually call it an intercept-only multilevel meta-analytic model. As shown in Figure 6, when estimating  $\beta_0$ , Equation 2 can exactly capture the multilevel/hierarchical data structure (e.g., multiple effect sizes are clustered/correlated within a study) that arises the statistical non-independence among effect sizes.

### 6.3. The degrees of dependency

Compared to the random-effects model (Equation 1), Equation 2 explicitly models the non-independence due to multiple effect sizes within one study (i.e., the random-effect term  $u_{w[i]}$  assumes heterogeneous effect sizes within studies). In the random-effects model, sampling variance ( $\text{Var}[e_i] = v_i$ ) is treated as the only source of within-study variance, whereas it is not the case for the multilevel model ( $u_{w[i]}$  lead to effect-size specific variance, which belongs to within-study variance). Therefore, the multilevel model can distinguish between sampling variance and within-study variance (see *section 6.3* for details). Moreover, the degrees of dependency among effect sizes can be quantified via the estimated correlation between the true underlying effects: intra-class correlation  $ICC = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2)$ . In case of no dependency or effect sizes not correlated within clusters/studies ( $ICC = 0$ ), all effect sizes derived from the same study are independent, meaning that they contribute fully

unique information to the fitted model, which is an implicit assumption of fitting dependent effect sizes to traditional meta-analytic models. If  $ICC = 1$ , all effect sizes derived from the same study are non-independent, meaning that they contribute the very same information to the fitted model, which is an implicit assumption of the ‘averaging’ method dealing with dependent effect sizes (see *Section 5*).

#### 6.4. Parameter estimation and statistical inference

The parameters of the two random-effects terms (i.e.,  $\sigma_b^2$  and  $\sigma_w^2$ ) can be estimated from the data along with  $E[ES_{[i]}] = \beta_0$  (note that  $ES_{[i]}$  and  $v_{[i]}$  can be directly computed from the data). There are various estimators to approximate  $\sigma_b^2$  and  $\sigma_w^2$ , such as maximum likelihood (ML), restricted maximum likelihood (REML), DerSimonian-Laird (DL), and Empirical Bayes (EB). Although DL is a common method in many meta-analyses (a default estimator in Review Manager, RevMan), simulation studies indicate REML and Empirical Bayes outperform over DL in different simulated data (McCann et al., 2016; Tanriver-Ayder et al., 2021; see Langan et al., 2019; Viechtbauer et al., 2015). Note that only ML and REML estimators are implementable in multilevel models in the current main R packages for conducting meta-analyses, for example, *metafor* package (Viechtbauer, 2010). Together with the estimated variance components (e.g.,  $\sigma_b^2$  and  $\sigma_w^2$ ) and sampling variance (i.e.,  $v_i$ ), the variance (and covariance if involving correlated  $v_i$ ) matrix can be constructed and model coefficients can be estimated under the inverse-variance



weighting scheme (Marin-Martinez and Sánchez-Meca, 2010). Finally, statistical inferences (e.g., statistical tests of model intercept  $\beta_0$  and CIs construction) can be made based on null-hypothesis tests (e.g., Wald-type tests with standard normal distribution or  $t$ -distribution), likelihood ratio tests, resampling methods (e.g., a permutation test and bootstrapping) or cluster-robust inference (sandwich-type estimator; see section 9.3). The use of methods based on  $t$ -distribution (with adjusted degrees of freedom), permutation test and cluster-robust inference are preferable in the case of a meta-analysis with a small number of studies (Joshi et al., 2022; Nakagawa et al., 2021c; Sánchez-Meca and Marín-Martínez, 2008; Viechtbauer et al., 2015). We illustrate how to implement multilevel models with recommended estimators (REML) and improved inference methods ( $t$ -distribution with adjusted degrees of freedoms) and interpret corresponding model results using `rma.mv()` function in `metafor` package in **section 11** (Supplementary Materials file 2; Viechtbauer, 2010).

### 6.5. Flexible random-effects structures

One of the multilevel model's advantages is the capacity to incorporate a flexible random-effects structure, and therefore, to account for various types of non-independence and heterogeneity due to different levels of clustering variables (see **section 7**). For example, animal meta-analyses often encounter stratified data structure, by multiple strains or species (e.g., more than 10 antidepressants included in Kara et al., 2018; 13 strains of rodents included in Bird et al., 2016; 8 species in

Lagisz et al., 2020). In this regard, we can add a corresponding random-effects term, for example, strain-specific effect, to Equation 2 to account for this:

$$ES_{[i]} = \beta_0 + u_{s[k]} + u_{b[j]} + u_{w[i]} + e_{[i]}, \quad (3)$$

$$u_{s[k]} \sim \mathcal{N}(0, \sigma_s^2)$$

where  $u_{s[k]}$  = a random-effects term corresponding to strain-specific effect for strain  $k$  (it also can be species-, drug-, dose-specific effects – depending on which clustering variable is incorporated into the model), wherein  $\text{Var}[u_{s[k]}] = \sigma_s^2$  denotes the strain-specific variance. As a rule of thumb, a proper random-effects term needs at least five levels to make the estimation of the respective variance components feasible and stable (Bolker et al., 2009; Gomes, 2021). Therefore, the strain-specific effect is preferable to the species-specific effect in practice (if there are fewer than 5 species included – in the case of biomedical studies usually only rats and mice). In a more complicated spectrum, we can further extend Equation 3 to account for additional sources of non-independence by adding random-effects terms – for example, authorship dependence (labs, research groups), non-phylogenetic species similarities, phylogeny relatedness, and temporal and spatial correlations (Hadfield and Nakagawa, 2010; Moulin and Amaral, 2020; Nakagawa et al., 2019; Maire et al., 2019). We only found one paper that accounted for these complex sources of non-independence, for example, Lagisz et al., 2020 employed a phylogenetic multilevel meta-analytic model to address the phylogeny relatedness among 22 non-human species. Given the complexities, we do not elaborate on the methodological details of

these complicated models (for interested researchers, see Chamberlain et al., 2012; Cinar et al., 2021a; Nakagawa and Santos, 2012).

Theoretically, any cluster or grouping variable can serve as a random-effects candidate (e.g., study identities, drug types, or species). However, the levels of one cluster variable might be strongly overlapping with another (e.g., study vs. animal cohort). Therefore, a practical problem to consider is to select the best random-effects structure when conducting a meta-analysis. The rationale of testing the random-effects structure is to investigate whether the examined random-effects terms are (1) of neurobiologically interest, and (2) true sources of heterogeneity. For a random-effects term, we often use information-theoretic approaches alongside likelihood methods to examine whether it is a suitable random-effects term that should be included in the meta-analytic model, such as Akaike Information Criterion (AIC) and likelihood ratio tests. It is worth noting that when comparing models with different candidate random-effects structures, the ML method should be used rather than REML, because the log-likelihoods ratio (the index of information criteria) is not estimable for models incorporated with different fixed-effects (Gurka, 2006). The methodological details of calculating AIC and log-likelihood ratio are not the focal interest in this paper (but see Cinar et al., 2021b for details). We provide an example showing how to use information-theoretic approaches to decide the best random-effects structure in Supplementary Materials file 2.

## 7. Quantifying heterogeneity in the multilevel model

### 7.1. Multilevel version of $I^2$ statistic and variance components

It is common for an animal meta-analysis to combine studies with experimental designs with multiple species/strains, multiple outcomes, and multiple trials, each with multiple arms (Sandercock and Roberts, 2002; Hunniford et al., 2021). All these neurobiological and methodological differences are likely to lead to inconsistency among effect sizes. In the meta-analytic context, this ‘inconsistency’ is typically referred to as “heterogeneity” of the true underlying effects, and animal studies have a high amount of heterogeneity indeed (Kafkafi et al., 2018; Richter et al., 2009; Voelkl et al., 2020). As with the traditional meta-analyses (fixed- and random-effects meta-analyses), the multilevel model also can measure the amount of heterogeneity in the true underlying effects. In animal meta-analytic practice, as revealed by our survey, there are three widely used statistics for determining the amount of heterogeneity:  $I^2$  (43%; Figure 3), Cochran’  $Q$  (20%), and between-study variance  $\tau^2$  (19%). In general, Cochran’  $Q$  could be useful because it facilitates dichotomous decisions regarding whether the effect sizes are homogeneous (Cochran’  $Q$  is a test statistic for testing the null hypothesis of homogeneous effect sizes). It also can be used to assess the uncertainty of between-study variance on some occasions (but see Van Aert et al., 2019a). But it is not as informative as  $I^2$  and  $\tau^2$ . Specifically,  $I^2$  and  $\tau^2$  can measure the amount of heterogeneity among effect sizes (the former is the relative heterogeneity, and the latter is the absolute heterogeneity; Borenstein et al., 2017).

Importantly, the multilevel meta-analytic can partition the two statistics across different levels corresponding to different random-effects terms. The random-effects terms in Equation 2 indicate that total variance components can be decomposed into between- and within-study-specific variances in the multilevel model ( $\sigma_b^2$ , and  $\sigma_w^2$ , respectively; Figure 7). The strain-specific variance  $\sigma_s^2$  also can be separated from the total variance if fitting Equation 3 (not shown in Figure 7).

Applying a random-effects model to non-independent data leads to model misspecification because a random-effects model treats the between-study variance as the total variance in ( $\tau^2 = \tau_{total}^2$ ; Equation 1), while a multilevel model treats between-study variance as one of the components of the total variance ( $\tau_{total}^2 \geq \sigma_b^2$  in Equation 2; Figure 7). Therefore, the true total variance ( $\sigma_{total}^2 = \sigma_b^2 + \sigma_w^2$ ) is incorrectly attributed to the between-study variance  $\tau^2$  in a random-effects model (where  $\tau^2$  should be equal to  $\sigma_b^2$  rather than  $\sigma_{total}^2$ ; see Supplementary Materials file 2 for a real example).  $I^2$  statistic is defined as the relative amount of variance between effect sizes after taking out sampling error effects (Higgins and Thompson, 2002). We formulate  $I^2$  statistics in the context of a multilevel meta-analytic model as follows (Cheung, 2014; Nakagawa and Santos, 2012):

within-study specific  $I^2$

$$I_w^2 = \frac{\sigma_w^2}{\text{Var}[ES_i]} = \frac{\sigma_w^2}{\sigma_{total}^2 + \sigma_{sampling}^2}, (4)$$

between-study specific  $I^2$

$$I_b^2 = \frac{\sigma_b^2}{\text{Var}[ES_i]} = \frac{\sigma_b^2}{\sigma_{total}^2 + \sigma_{sampling}^2}, \quad (5)$$

total level  $I^2$

$$I_{total}^2 = \frac{\sigma_{total}^2}{\text{Var}[ES_i]} = \frac{\sigma_{total}^2}{\sigma_{total}^2 + \sigma_{sampling}^2}, \quad (6)$$

where  $\text{Var}[ES_i] = \sigma_{total}^2 + \sigma_{sampling}^2$  signifies the total variance of the observed/estimated effect size (i.e.,  $ES_i$ );  $\sigma_{total}^2$  denotes the total variance in the true effects (true heterogeneity;  $\sigma_{total}^2 = \sigma_b^2 + \sigma_w^2$  in Equation 2), which is caused by neurobiological-, and methodological relevant variability (and can be explained by corresponding moderator variables; see Figure 7 and **section 8**);  $\sigma_{sampling}^2 = a$  “typical” sampling error variance, which is driven by the finite ‘sampling’ of the population;  $\sigma_{sampling}^2$  can be estimated using (independent) sampling variance  $v_{[i]}$  (Higgins and Thompson, 2002; but see **section 10.2** for non-independent  $v_{[i]}$ ):

$$\sigma_{sampling}^2 = \frac{(k-1) \sum_{i=1}^k 1/v_{[i]}}{(\sum_{i=1}^k 1/v_{[i]})^2 - \sum_{i=1}^k 1/v_{[i]}^2}, \quad (7)$$

where  $\sigma_{sampling}^2$  is also called a “typical” within-study variance  $\bar{v}$ , since sampling variance is the only source of within-study variance in the framework of the random-effects model (Equation 1:  $N = k$ ).  $\sigma_{sampling}^2$  or  $\bar{v}$  can be conceptually treated as a surrogate of the average value of sampling variances  $v_{[i]}$ .

The multilevel versions of  $I^2$  statistic have three merits: (1) intuitive (range from 0 to 100% enabling us to have a clear sense of the amount of heterogeneity in a given meta-analysis), (2) with commonly used guidelines ( $I^2 = 25, 50, 75\%$  can be

interpreted as low, moderate and high levels of heterogeneity; Higgins et al., 2003), and (3) interpretable ( $I_w^2$ , and  $I_b^2$  are the proportions of the effect size variation attributed to within- and between-study inconsistencies, respectively). We show the calculations of the multilevel version of  $I^2$  index using *i2\_ml* function in *orchard* package in **section 11** (Nakagawa et al., 2021b).

## 7.2. Prediction intervals

Our survey also found one useful heterogeneity index used in animal meta-analyses – prediction interval (PI; Figure 3; Mancini et al., 2020). 95% PI is defined as the estimate of an interval (a plausible value range) where 95% of the future measurements (i.e., true effect sizes of new studies) would fall when no sampling errors exist (Riley et al., 2011; van Aert et al., 2021). For example, assume an antidepressant with a mean SMD = -0.4 and 95% PI = -0.1 to -0.7 – this means 95% of new trials using this antidepressant will decrease the manifestation of depressive behaviours by between 0.1 to 0.7 standard deviations over different experimental contexts. We note that PI is distinct from confidence interval (CI). A CI quantifies the precision of the mean effect size (i.e.,  $\beta_0$ ), which is dependent on the standard error (i.e.,  $SE[\beta_0]$ ):  $95\%CI = \beta_0 \pm t_{df;0.975}\sqrt{SE[\beta_0]^2}$ , where  $t_{df;0.975}$  = 97.5th percentile of a Student's *t*-distribution with *df* degrees of freedom. In contrast, PI captures the dispersion of the mean effect size, which accommodates heterogeneity in the true

underlying effects, namely, neurobiologically and methodologically relevant uncertainties (i.e., variance in the true effects –  $\sigma_{total}^2$ ; **section 7.1**):

$$95\%PI = \beta_0 \pm t_{0.975} \sqrt{\sigma_{total}^2 + SE[\beta_0]^2}, \quad (8)$$

where the exact value of  $df$  of the Student's  $t$ -distribution is controversial; Some common approximations include  $df = k - 1$ ,  $k - 2$ , or  $k - 4$ , where  $k$  is the number of effect size estimates (a detailed discussion see Knapp and Hartung, 2003; Riley et al., 2011; van Aert et al., 2021; Viechtbauer, 2010). The *metafor* package uses  $k - 1$  as the default.

With the same scale as its mean effect size, the neurobiological interpretation of PI is straightforward. This merit makes it a good complementary statistic to the  $I^2$  index since  $I^2$  index provides no information on the absolute amount of heterogeneity among effect sizes (although it directly measures the percentage of heterogeneity due to ‘true’ neurobiological-relevant variation as opposed to chance (i.e., sampling variance) – Equation 7; Borenstein et al., 2017). Specifically, for a given animal meta-analysis with small  $\sigma_{sampling}^2$ , even a tiny true effect heterogeneity  $\sigma_{total}^2$  can lead to a high value of  $I_{total}^2$  (see Equation 6). However, this large  $I^2$  does not necessarily mean a high amount of true heterogeneity of neurobiological-relevant variation in the effect sizes. In contrast,  $\sigma_{total}^2$  can directly reflect the genuine differences underlying the true effects because the square root of  $\sigma_{total}^2$  can be interpreted as the standard deviation of the true effect sizes. Moreover, in the framework of the multilevel model,



$\sigma_{total}^2$  also can be partitioned at different levels (e.g., within- and between-study levels:  $\sigma_w^2$  and  $\sigma_b^2$ ; see Figure 7). Therefore, both PI and  $\sigma_{total}^2$  should be reported as a complement to the commonly used  $I^2$  index. The calculation of PI and partition of  $\sigma_{total}^2$  are readily available in existing packages (see examples in Supplementary Materials file 2).

## 8. Multilevel meta-regression

### 8.1. Multilevel meta-regression to explain heterogeneity in effect sizes

When heterogeneity is detected (which is indicated by a large  $I_{total}^2$  and significantly meaningful  $\sigma_{total}^2$ ; see Supplementary Materials file 2), it is necessary to find the drivers of such variability and try to explain (at least part of) this heterogeneity using variables extracted from primary studies as explanatory variables or predictors (known as moderator variables in the meta-analytic context because they moderate the strength of the effect on effect size  $ES_{[i]}$ ). There are three common drivers of heterogeneity: neurobiological, methodological and sociological (or meta-scientific) moderators (Nakagawa and Santos, 2012). The neurobiological and methodological moderators can be used to account for heterogeneity due to neurobiological processes (e.g., different doses of drug, or sex tested) and methodological differences (e.g., different drugs or dosages), respectively. The sociological moderators are the drivers of publication bias (see *section 8.2* for details). In other words, including moderator variable (s) in a multilevel model (Equation 5) allows us to test moderator effects –

examining whether the effect size estimates systematically change in response to different levels of moderator variables (e.g., whether the overall efficacy of an antidepressant drug depends on sex). We found 89% (55/62) of animal meta-analyses accounted for heterogeneity using either meta-regression or subgroup analysis (Supplementary Materials file 1), among which 33 employed meta-regression.

The above multilevel meta-analytic model involving a moderator, in essence, is a meta-analytic regression model via multilevel (linear) mixed-effects models with both fixed effects and random effects, also well-known as (mixed-effects) meta-regression models. It can be expressed as the following mathematical notations:

$$ES_{[i]} = \beta'_0 + \beta_1 x_{b[j]} + u_{b[j]} + u_{w[i]} + e_{[i]}, (9)$$

where  $\beta'_0$  = an intercept (which is different from the overall mean,  $\beta_0$ , in Equation 2, and this has an important implication in certain circumstances; see **section 9.1.2**);  $\beta_1$  = a slope representing effect size changes for (one-unit increase in)  $x_{b[j]}$ , a moderator corresponding to between-study characteristics (e.g.,  $x_{b[j]}$  = antidepressant types:

fluoxetine vs. sertraline) or strain-specific characteristics (e.g.,  $x_{s[k]}$  = animal

taxonomy: rats vs. mice). Equation 9 is known as a three-level mixed-effects model.

Importantly, Equation 9 also enables us to examine the relationship between effect size changes and a moderator variable that varies within studies (e.g., within-study

level moderator variable  $x_{w[i]}$  = a series of doses; Figures 5 and 7). Categorical

moderator variables also can be incorporated into Equation 9 using a dummy-coding

strategy (Schielzeth, 2010). For example, animal sex can be dummy coded as  $D_{sex}$ : 0 (male) and 1 (female). This is equivalent to subgroup analysis. In the framework of meta-regression, subgroup analysis is achieved by incorporating  $D_{sex}$  into Equation 9:

$$ES_{[i]} = \beta_{male} + \beta_{\Delta}D_{sex} + u_{b[j]} + u_{w[i]} + e_{[i]}, \quad (11)$$

where  $\beta_{male}$  = an intercept indicates the estimate of the mean effect size of the subgroup male;  $\beta_{\Delta}$  = a slope equals to  $\beta_{female} - \beta_{male}$ . Equation 11 is intuitive: when  $D_{sex} = 0$  (subgroup = male), we can obtain the estimate of  $\beta_{male}$ ; when  $D_{sex} = 1$  (subgroup = female), we obtain the estimate of  $\beta_{female} = \beta_{male} + \beta_{\Delta}$ . Importantly,  $\beta_{\Delta}$  enables us to examine the difference of a categorical moderator variable at different levels (e.g., the differences between females and males in responses to a given antidepressant; see Supplementary Materials file 2).

Equation 9 explicitly indicates that the strategies of aggregating or selecting effect sizes will prevent the meta-regression from providing information about effect-size level moderator variables (see Figure 5 and **section 5**). Importantly, Equation 9 can further partition the total variance (i.e.,  $\sigma_{total}^2$ ) into two parts: (1) the variance of fixed-effects,  $\sigma_f^2 = \text{Var}[\beta_1 x_{b[j]}]$ , which denotes the heterogeneity explained by the included moderator variable  $x_{between[j]}$  (Figure 7); and (2) the variance of random-effects terms ( $\sigma_b^2$ , and  $\sigma_w^2$ ), which now becomes “residual” heterogeneity that is not explained by the associated moderator variable. The goodness-of-fit index  $R^2$  is also applicable to quantify the percentage of variance explained by the included moderator

variable (Aloe et al., 2010). A general and widespread measure of  $R^2$  is the marginal  $R^2$  (Nakagawa and Schielzeth, 2013), which can be calculated by:

$$R_{marginal}^2 = \frac{\sigma_f^2}{\sigma_{total}^2} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_b^2 + \sigma_w^2} \quad (10)$$

In contrast to  $I_{total}^2$  (Equation 6),  $R_{marginal}^2$  does not contain the sampling error variance ( $\sigma_{sampling}^2$ ; Equation 7) in the denominator because this variance component is assumed to be known before including moderator variables to explain the heterogeneity. The calculation of  $R_{marginal}^2$  is readily available, for example, using *orchaRd* packages (see **Section 11**).

Two points are worth noting here. First, various methods can be used to deal with covariates with missingness under the assumption of a random missing mechanism (e.g., data are missing completely at random and unrelated to any other variable), including simple deletion (filtering the incomplete cases prior to model fitting; embedded in *metafor*; Viechtbauer, 2010) and advanced imputation methods (i.e., full information maximum likelihood embedded in *metaSEM*; Cheung, 2015; Jak and Cheung, 2020). Second, a random meta-regression requires each of its moderators to have at least five studies (Hedges and Pigott, 2004). The minimal number of studies or effect sizes required by a multilevel meta-regression remains unknown, albeit some simulation studies suggest that the estimates of model coefficients of a multilevel meta-regression are generally stable under various simulated situations (Jamshidi et al., 2020; López-López et al., 2017).

## 8.2. Extended Egger's regression to test publication bias

Identifying publication bias is a crucial and mandatory procedure of a meta-analysis because the validity of meta-analytic evidence would be undermined if publication bias occurs (Augusteijn et al., 2019; Nakagawa et al., 2017; Van Aert et al., 2019b).

The most common testable form of publication bias is the small-study effect where small studies (i.e., small sample size) often tend to report large effect sizes (Sterne et al., 2000). In our survey, we found that 86% of animal meta-analyses dealt with publication bias in their analyses in some way (Figure 3). The most common method used to examine publication bias was: funnel plots (35%), (simple) regression-based methods (e.g., Egger's regression; 30%), and trim-and-fill tests (14%). However, all these procedures are not appropriate if effect sizes are statistically dependent or heterogeneous or both (Rodgers and Pustejovsky, 2021; Sterne et al., 2001a).

Recently, a multilevel version of Egger's regression has been proposed to tackle these limitations (Fernández-Castilla et al., 2021; Nakagawa et al., 2021a). Briefly, we need to add sampling error ( $se_{[i]} = \sqrt{v_{[i]}}$ ) as a moderator variable into a multilevel model (equivalent to set  $x_{within[i]}$  as  $se_{[i]}$  in Equation 9; a potentially better approach is to use 'effective sample size' to replace  $se_{[i]}$ ; see Nakagawa et al., 2021a). Then, a statistically significant  $\beta_1$  (i.e., the slope of  $se_{[i]}$ ) means that studies with large  $se_{[i]}$  (i.e., small sample size) have large effect size. This indicates that a small study effect

exists in the meta-analytic dataset. Likewise, if including publication year as a moderator variable, Equation 9 can detect another form of publication bias, the decline effect (i.e., time-lag bias, which is defined as the temporal instability of the magnitude of effect sizes), the implication of which is underappreciated (Grainger et al., 2020; Koricheva and Kulinskaya, 2019; Nakagawa et al., 2021a; see Supplementary Materials file 2)

## **9. Extensions to the multilevel models**

### *9.1. Multi-moderator multilevel meta-regression*

In practice, it is common to test one moderator variable at a time to explain the heterogeneity among effect sizes (i.e., Equation 9; known as the single-moderator meta-regression model or univariable meta-regression). Theoretically, multiple moderators ( $x_{w[i]}$  and  $x_{b[j]}$ ) can be examined simultaneously. This leads to a multi-moderator multilevel meta-regression (i.e., multivariable meta-regression). In contrast to the univariable meta-regression (i.e., Equation 9), a multi-moderator multilevel meta-regression can provide more neurobiological and meta-scientific insights. For example, it enables us to ask (1) whether there exists an interactive effect between two moderator variables, and (2) what is the adjusted effect size of an animal meta-analysis after correcting for publication bias (Kvarven et al., 2020). Yet, the complexity of parameterization of such a meta-regression requires a large dataset to make optimization algorithms free of convergence issues (Bates et al., 2015; Cinar et

al., 2021a). Given that (at least some) datasets in our surveyed animal meta-analyses are not small ( $k$ : medium = 56, range = 5 – 3288, mean = 203; Supplementary Materials file 1), it is feasible to introduce this more complex model to the field.

### 9.1.1. Investigating the effects of multiple moderators and interactions

For the sake of illustration, we use the simplest form of the multi-moderator multilevel meta-regression model as an illustration:

$$ES_{[i]} = \beta'_0 + \beta_1 x_{w[i]} + \beta_2 x_{b[j]} + u_{b[j]} + u_{w[i]} + e_{[i]}, \quad (12)$$

Equation 12 builds upon Equation 9 and contains two moderator variables, whose slopes  $\beta_1$  and  $\beta_2$  can be used to quantify the (average) effects of  $x_{w[i]}$  (e.g., dose) and  $x_{b[j]}$  (e.g., sex) on effect size changes, separately. In practice, the effect of  $x_{w[i]}$  (e.g., dose) might be confounded by  $x_{b[j]}$  (e.g., sex). For example, high doses of antidepressants are more likely to mitigate the depression symptoms of females, while antidepressants are less effective on males (Figure 8; Mauvais-Jarvis et al., 2020; Tannenbaum et al., 2019). This example requires us to control for the impact of  $x_{b[j]}$  (e.g., sex) when quantifying the effect of  $x_{w[i]}$  (e.g., dose) on effect sizes. To do so, an interaction term needs to be added to Equation 12:

$$ES_{[i]} = \beta'_0 + \beta_1 x_{w[i]} + \beta_2 x_{b[j]} + \beta_3 x_{w[i]} x_{b[j]} + u_{b[j]} + u_{w[i]} + e_{[i]}, \quad (13)$$

where  $x_{w[i]} x_{b[j]}$  = interaction between  $x_{w[i]}$  and  $x_{b[j]}$ ;  $\beta_3$  = slope of the interaction term, which captures the magnitude of the interactive effect. If the significance test of

the model coefficients shows a statistically significant  $\beta_3$ , we conclude that the two moderator variables can interact with each other.

Equations 12 and 13 can be extended to a more general form:

$$ES_{[i]} = \beta'_0 + \sum \beta_{mod}x_m + u_{b[j]} + u_{w[i]} + e_{[i]}, (14)$$

where the variable  $x_m$  can be any moderator variable denoting within- and between-study level characteristics;  $\beta_{mod}$  = the moderator variable  $x_m$ 's slope, which is interpreted as the magnitude of the moderator effect for  $x_m$  (e.g., the effect of antidepressants on depression symptoms);  $\sum \beta_{mod}x_m$  = the sum of all moderator effects. Though Equation 14 is not commonly used in the discipline (but see Vendl et al., 2021), it has versatile functionality. It can be used to predict the combined effects of two moderator variables, for example, examining how the effects of an antidepressant drug on females at a series of doses even if these doses have not been tested by empirical studies (e.g., conditional [marginal] estimates). Given a large enough number of effect sizes, Equation 14 also can be used to construct a linear and non-linear relationship (e.g., quadratic, cubic polynomial, and spline) between a continuous moderator variable and effect size estimates  $ES_{[i]}$  (Gasparrini et al., 2012; Orsini et al., 2012). It is worth noting that multi-moderator multilevel meta-regression models share limitations with other types of linear models: for example, they are susceptible to overfitting and multi-collinearity. For a detailed illustration of these complex applications, see Supplementary Materials file 2.



### 9.1.2. Correcting meta-analysis for publication bias

When replacing  $x_{w[i]}$  and  $x_{b[j]}$  by sampling error ( $se_{[i]}$ ) and publication year of a paper ( $year_{[j]}$ ), Equation 13 enables us to test for small-study effect and time-lag bias simultaneously. More importantly, such a model can correct for publication bias in animal studies:

$$ES_{[i]} = \beta'_0 + \beta_1 se_{[i]} + \beta_2 year_{[i]} + u_{b[j]} + u_{w[i]} + e_{[i]}, \quad (15)$$

where  $\beta'_0$  is an intercept, which could serve as the bias-corrected overall effect.

Imagine that if the meta-analytic data does not have publication bias (e.g., no small-study effect and time-lag bias exist), we are more likely to obtain the bias-corrected effect. Theoretically,  $se_{[i]} = 0$  and  $year_{[j]} = 0$  indicates that no small-study effect and time-lag bias occur.  $\beta'_0$  is the estimate which is explicitly conditional on  $se_{[i]} = 0$  and  $year_{[j]} = 0$ . There are two notable issues if our interest is to estimate a bias-corrected overall effect. First, we need to centre  $year_j$  (i.e.,  $c(year_{[j]})$ ) at its mean value (set mean  $c(year_{[j]})$  as 0), such that  $\beta'_0$  is meaningful to be interpreted as a bias-corrected overall effect:

$$ES_{[i]} = \beta'_0 + \beta_1 se_{[i]} + \beta_2 c(year_{[i]}) + u_{b[j]} + u_{w[i]} + e_{[i]}, \quad (16)$$

Second, if  $\beta'_0$  in Equation 16 is significantly different from zero (i.e., a true effect), some researchers recommend replacing sampling error  $se_{[i]}$  by its sampling variance  $se_{[i]}^2$  or  $v_{[i]}$  (Nakagawa et al., 2021a; Stanley et al., 2017):

$$ES_{[i]} = \beta'_0 + \beta_1 se_{[i]}^2 + \beta_2 c(year_{[i]}) + u_{b[j]} + u_{w[i]} + e_{[i]}. \quad (17)$$

Equation 17 assumes a quadratic association between sampling error ( $se_{[i]}$ ) and effect sizes ( $ES_{[i]}$ ) to avoid a downwardly biased estimate of the bias-corrected overall effect ( $\beta'_0$ ). The combination of Equations 16 and 17 is a so-called two-step procedure (Stanley et al., 2017), which provides us with an unprecedented opportunity for testing and correcting for publication bias (see Supplementary Materials file 2 for implementation).

Given that high heterogeneity may invalidate publication bias test (Macaskill et al., 2001; Moreno et al., 2009; Sterne et al., 2001b), it is best to account for the potential heterogeneity when testing publication bias:

$$ES_{[i]} = \beta'_0 + \beta_1 se_{[i]} + \beta_2 c(year_{[j]}) + \sum \beta_{mod} x_m + u_{b[j]} + u_{w[i]} + e_{[i]}, \quad (18)$$

The significance tests of  $\beta_1$  and  $\beta_2$  can be used to indicate the presence of small-study effect and time-lag bias, separately (see *section 8.2*);  $\sum \beta_{mod} x_m$  in Equation 18 is used to accommodate the potential heterogeneity in the animal dataset. Moreover, we can use Equation 18 to correct for publication bias for the moderator effects, for example, to estimate the efficiency of different antidepressant drugs on depression symptoms after adjusting for publication bias. Following the similar rationale of estimating the bias-corrected overall effect (see above), we can estimate the bias-corrected effect for any moderator variable  $x_m$  by estimating  $\beta_{mod}$  conditional upon  $se_{[i]} = 0$  (no small-study effect) and  $year_{[j]} = 0$  (no time-lag bias). Several existing

packages are readily available to implement Equation 18, such as *emmeans*, *lsmeans*, and *orchaRd* (Lenth, 2016; Nakagawa et al., 2021b; Russell, 2021).

## 9.2. Multivariate models for modelling multiple outcomes simultaneously

Some behavioural and neurobiological studies measure more than a single outcome (e.g., two different endpoints: anxiety and depression within the same studies), therefore, resulting in more than one type of effect size or multivariate effect sizes.

When computing multivariate effect sizes, at least two types of non-independence arise due to repeated use of the sample cohort of animals: correlations/covariances between different outcomes and sampling errors (scenarios 2 and 3 in Figure 4).

Multivariate meta-analytic models have been proposed to account for these dependent effect sizes by modelling the multiple types of outcomes simultaneously (also referred to as multi-response and multi-outcome meta-analytic models). Extending univariate meta-analytic models (Equation 1) to multivariate versions is similar to extending ANOVA to MANOVA. In contrast to a (univariate) random-effects meta-analytic model (Equation 1), a multivariate meta-analytic model allows both between-study random effects ( $\mathbf{u}_i$ ) and sampling errors ( $\mathbf{e}_j$ ; see next section) to follow a multivariate normal distribution (i.e.,  $\mathbf{u}_i \sim \mathcal{N}(0, \mathbf{T}^2)$  and  $\mathbf{e}_i \sim \mathcal{N}(0, \mathbf{V})$ ).

The principle is that two types of dependence can be directly captured by (co)variance structures of the multivariate models. Specifically, correlations ( $\rho_{[p]}$ ) between

different outcomes are defined in the variance-covariance matrix of the between-study variances  $\text{Cov}[\mathbf{u}_i] = \mathbf{T}^2$  (which are distinct from the one-dimension between-study variances  $\tau^2$  in the univariate random-effects model, Equation 1). Likewise, the correlations ( $\rho_{[s]}$ ) between different sampling errors are defined in the variance-covariance matrix of the sampling errors  $\text{Cov}[\mathbf{e}_i] = \mathbf{V}$  (which are distinct from the one-dimension sampling variances in the random-effects model; details see next section). Assume a simple example of  $\mathbf{T}^2$  involving three outcomes:

$$\mathbf{T}^2 = \begin{bmatrix} \tau_1^2 & \rho_{[p]12}\tau_1\tau_2 & \rho_{[p]13}\tau_1\tau_3 \\ \rho_{[p]21}\tau_2\tau_1 & \tau_2^2 & \rho_{[p]23}\tau_2\tau_3 \\ \rho_{[p]31}\tau_3\tau_1 & \rho_{[p]32}\tau_3\tau_2 & \tau_3^2 \end{bmatrix}, \quad (19)$$

where  $\tau_i^2$  = between-study variance for outcomes  $i$ ;  $\rho_{[p]21} = \rho_{[p]12}$  = population level correlation (also known as between-study correlation; Gasparrini et al., 2012; Jackson et al., 2011), which is the correlation between the first and second population outcomes; covariance of the first and second outcomes  $\text{Cov}[\tau_1^2, \tau_2^2] = \rho_{[p]21}\tau_2\tau_1 = \rho_{[p]12}\tau_2\tau_1$ . The univariate  $I^2$  statistic based on the proportion of between-study variance ( $\tau^2$ ) in the total variance ( $\tau^2 + \bar{v}$ ) can be easily extended to multivariate  $I^2$  statistics (Higgins and Thompson, 2002). An alternative definition of  $I^2$  is based on the variance-covariance matrix of the model coefficients under the multivariate random- and fixed-effects models (Jackson et al., 2012).

The matrix  $\mathbf{T}^2$ , which carries information about the extent to which two pairwise outcomes are correlated, can be used to account for dependent effect sizes while also

improving the accuracy of model estimates and enabling the investigation of new questions, such as whether there is a strong correlation between two outcomes at the population level. Multivariate models can also leverage the “borrowing of strength” among correlated outcomes, allowing the estimation of missing effect sizes feasible even when some outcomes are only partially reported in some studies. Nonetheless, three points should be noted here. First, there might be only a few studies reporting complete paired outcomes. As such, no information (from the paired effect sizes from the same study) will be borrowed to increase the precision of model estimates (Jackson et al., 2011). Second, when the included studies have a large number of outcomes, the multivariate models are highly parameterized and likely to be overparameterized. In this regard, the advantages of multivariate models might be compromised by the increasing the number of parameters that need to be estimated (Boca et al., 2017). For example, 15 neural-behavioural traits in Yang et al. (2021) need to estimate 105 correlations  $\rho_{[p]}$ . To mitigate the estimation and computational challenges posed by high dimensional parameters, some researchers proposed to enforce a simplified structure on the between-study covariance matrix  $\mathbf{T}^2$ , such as a compound-symmetry or a diagonal structure (Gasparrini and Armstrong, 2011; Gasparrini et al., 2012; Ritz et al., 2008). Third, it is challenging to construct a sampling variance-covariance matrix  $\mathbf{V}$  due to the lack of sampling covariances (but see next section for solutions).

### 9.3. Within-study variance-covariance matrix for accounting for correlated sampling errors

Typically, sampling errors are assumed to be independent in the framework of the multilevel model outlined early. However, as mentioned above, repeated use of the same cohort animal or “overlapping” animals (when calculating effect sizes) leads to dependency among sampling errors (see correlated sampling errors due to shared animal and control in Figure 4). Indeed, the nested random-effects structure in a multilevel model fails to capture this type of non-independence (i.e., mis-specified variance structure). As with the multivariate models, constructing a variance-covariance matrix for sampling errors (within-study VCV matrix) is the most straightforward way to account for correlated sampling errors. Theoretically, a multilevel model with a within-study VCV matrix can capture all types of dependence structure of effect sizes, wherein an appropriate specification of random-effects structure ( $u_{b[j]}$  and  $u_{w[i]}$ ) can account for correlations/covariances in the true effects and a within-study VCV matrix can account for the correlations/covariances in the sampling errors. The sampling error effect ( $e_i$ ) follows a multivariate normal distribution with mean zero and variances of  $V$  ( $e_i \sim \mathcal{N}(0, V)$ ). A simple example  $V$  having two studies with three effect sizes (where the first study contributes two effect sizes) can be expressed as:

$$V = \begin{bmatrix} se_1^2 & \rho_{[s]12}se_1se_2 & 0 \\ \rho_{[s]21}se_2se_1 & se_2^2 & 0 \\ 0 & 0 & se_3^2 \end{bmatrix}, (22)$$

where  $se_i$  = the sampling error (i.e., standard error) of effect size  $i$  (square root of sampling variance  $v_i$ );  $\rho_{[s]12} = \rho_{[s]21}$  = sampling level correlation (also known as within-study correlation or sampling correlation), which is the correlation between the first two observed effect size estimates, or more precisely  $se_1$  and  $se_2$ ; the sampling covariance  $\text{Cov}[v_1, v_2] = \rho_{[s]12}se_1se_2 = \rho_{[s]21}se_2se_1$ .

Yet, 63% of surveyed animal meta-analyses ignored correlated sampling errors in their analyses (but see Lagisz et al., 2020; Supplementary Materials file 1), where a within-study VCV matrix is incorrectly treated as a diagonal matrix with the sampling variance  $se_i^2$  (or  $v_i$ ) along the diagonal and zero along the off-diagonal (i.e.,  $\text{Cov}[v_i, v_k] = 0, i \neq k$ ; model misspecification). The challenge of constructing a within-study VCV matrix is that  $\rho_{[s]}$  or individual data used to compute  $\rho_{[s]}$  are rarely reported in the primary studies. We propose three approaches to construct a within-study VCV matrix to account for correlated sampling errors: (1) “empirical sampling correlation”, (2) “general sampling correlation”, and (3) “partially empirical sampling correlation” solutions. First, for studies reporting  $\rho_{[s]}$  or sampling covariances, we can directly extract them (contacting authors if missing). For studies involving multiple-treatment and multiple-outcome (Figure 4), we can use specific formula based on summary statistics to compute (asymptotic) sampling covariances for common effect size statistics (“empirical sampling correlation” solution; see Gleser and Olkin, 2009; Lajeunesse, 2011 for formula). Second, for studies involving other types of correlated

sampling errors (e.g., repeated measurements of the same outcome), we can use a “general” formula to compute sampling covariances, that is, “guesstimate” a constant sampling correlation  $\rho_{[s]}$  to calculate sampling covariances (e.g.,  $\rho_{[s]} = 0.5$  or  $0.8$ ; “general sampling correlation” solution; Fisher and Tipton, 2015; Noble et al., 2017; Pustejovsky and Tipton, 2022). The constant  $\rho_{[s]}$  can be rough, arbitrary, or educated guess. The validity of the guessed  $\rho_{[s]}$  can be ensured by robust variance estimation (see next section) or a sensitivity analysis, through which the robustness of the model parameter estimates (e.g.,  $\beta_0$ ) to the choice of  $\rho_{[s]}$  values would be tested. Third, use the “empirical sampling correlation” solution where possible, and supplement the “general sampling correlation” solution to take care of the remaining covariances (“partially empirical sampling correlation” solutions; Pustejovsky and Tipton, 2022).

#### *9.4. Robust variance estimation for counteracting model misspecification*

Our survey found that two papers used the robust variance estimation, RVE, methods to deal with statistical non-independence (Supplementary Materials file 1; Shields et al., 2015; Zajitschek et al., 2020). The superiority of RVE methods is that they can handle statistical non-independence even without knowing the exact dependence structure of effect sizes (Hedges et al., 2010; Tipton, 2013). RVE methods estimate sampling covariances from the data if the meta-analysis includes a larger number of studies (Hedges et al., 2010). Subsequently, standard errors of model coefficients are adjusted (robust errors) to avoid inflated Type I error rates and incorrect hypothesis



tests (e.g.,  $p$ -value). Existing R packages can easily implement RVE methods (Fisher and Tipton, 2015; Pustejovsky and Tipton, 2018). If our primary interest is in parameter estimations and hypothesis tests of model coefficients, then RVE method is an appealing way to deal with non-independence among effect sizes.

While the RVE method has its benefits, it may not be as effective when employed as a standalone rather than a complementary approach to multilevel or multivariate models. For example, RVE cannot decompose variance components into between- and within-study variances as the multilevel models do. Nor can RVE allow variance components to vary for different outcomes within studies as the multivariate models do. Therefore, it is promising to combine the multilevel or multivariate models with RVE (Nakagawa et al., 2021c; Pustejovsky and Tipton, 2022) where the multilevel or multivariate models can provide new insight regarding heterogeneity- or variance-related parameters while RVE can properly deal with all types of non-independence (especially mis-specified VCV matrix with a guesstimate within-study correlation). Fortunately, this hybrid strategy can be implemented by the combination of *metafor* and *clubSandwich* packages (see Supplementary Materials file 2). It should be noted that when the number of studies is small and the moderators are imbalanced, RVE cannot perform nominally (Tanner-Smith et al., 2016; Tipton, 2015). As such, applying RVE to a small-sample size meta-analyses might provide a biased estimate of the sampling variance and covariances matrix, leading to inaccurate robust errors

and inflated Type I error rates. Several small sample-size adjustment methods can be used to address these issues (Pustejovsky and Tipton, 2018; Tipton and Pustejovsky, 2015; Welz et al., 2023). Recently, the robust-wild-bootstrapping method has been introduced to control Type I error rates while improving the statistical power of hypothesis tests in RVE (Joshi et al., 2022).

## **10. Recommended practices for animal multilevel meta-analysis**

Given the data-generating processes and mechanisms in the field (e.g., a single construct of interest and nested effect sizes), we outline our practice recommendations for conducting meta-analysis using animal data. We strongly recommend using the multilevel model as the framework for conducting animal meta-analyses. Fitting multilevel models is feasible and conceptually simple (see Supplementary file 2). As such, it provides a good starting point for researchers in this field to properly account for non-independence, while producing reliable parameter estimates and hypothesis tests (as demonstrated in an example in *Section 11*).

More specifically, we recommend:

(I) Using the (three-level) multilevel models (e.g., Equation 2) by default, rather than fixed- and random-effects models (conventional practices; Equation 1), with the option to add additional levels for random-effects when necessary (e.g., when analysing data from multiple species, as described in *Section 6*).

(II) Performing all necessary meta-analytic procedures within the context of the multilevel model, such as using meta-regression to explain heterogeneity or conducting subgroup analyses and testing publication bias (*Section 7*).

(III) Transparently reporting all the relevant parameter estimates and hypothesis testing concerning fixed effects (e.g.,  $\beta_0$ , 95CI%, and 95%PI) and random effects (e.g.,  $I^2$ ) (*Section 8*).

While more technically challenging, we still recommend researchers to account for additional non-independence arising from correlated sampling errors if the dataset includes “shared control” and “shared animals” (i.e., the same animal cohort is used repeatedly to compute effect sizes): This can be achieved through the following steps:

(I) Using a multilevel model with a within-study VCV matrix that reflects the extent to which the sampling errors derived from the same animal are correlated (*Section 9.3*). Such a multilevel model with a within-study VCV matrix is technically a special case of a multivariate meta-analysis model without moderators for different outcomes, but the inclusion of random effects at multiple levels can provide additional insights into heterogeneity among effect sizes (*Section 11*).

(II) Using RVE methods to defend against the (potential) model misspecification if the sampling correlation  $\rho_{[s]}$  is not based on empirical data but assumed to construct the within-study VCV matrix (*Section 9.4*).

In the case where the dataset involves a multivariate structure (e.g., more than one type of outcome reported in many studies in a meta-analytic dataset) and multiple outcomes are of primary interest, we suggest the following procedures:

(I) Employing the multivariate meta-analysis models, which allow for the explicit consideration of non-independence due to both effect sizes and sampling errors through the use of between- and within-study variance-covariance matrices (*Section 9.2*).

(II) Using RVE methods as necessary to guard against the (potential) model misspecification.

The dataset with more complexity, including nested and correlated structures, may necessitate the use of more sophisticated approaches. A potential solution is to combine the multilevel and multivariate models, referred to as 'multivariate multilevel meta-analytic models', along with RVE methods. A simpler version of this combined approach is the multilevel model with a within-study VCV matrix, as outlined above. Further investigation, through simulation studies, is necessary to determine the empirical performance of this proposed methodology.

## **11. Worked example**

In this section, we briefly compare the results of our re-analyses of a real dataset, following our recommended practices with the results based on conventional

practices. Our re-analyses explicitly illustrate how failure to account for non-independence using traditional meta-analytic techniques might lead to the incorrect inference, underestimated standard errors and distorted hypothesis testing, and ultimately result in spurious conclusions. This worked example comes from one of our surveyed papers – Ramsteijn et al., 2020, on the relationship between the use of selective serotonin re-uptake inhibitor (SSRI) in animals during pregnancy and their offspring's neurobehavioural phenotype. Effect sizes in Ramsteijn et al., 2020's, dataset were statistically dependent. For example, within a given study included in Ramsteijn et al., 2020, male and female animals were compared separately. Likewise, animals were exposed to different types of SSRI antidepressants, the authors analysed these comparisons (within the same study) as if they were independent studies. Using the random-effects model to fit these non-independent effect sizes runs the risk of getting spurious results. We randomly selected a subset from Ramsteijn et al., 2020 and we dealt with the non-independence (as shown above) using the advanced methods with the practices outlined in *Section 10* and the traditional methods outlined in *Section 5*. It should be noted that these re-analyses are only for illustrative purposes. Readers interested in the biological questions should refer to the original paper.

For the selected subset, Ramsteijn et al., 2020 used a random-effects model (implicitly ignoring non-independence) and found that the use of SSRI in animals

during pregnancy significantly decreased offspring's sensory processing function (pooled SMD = -0.37, 95CI% = [-0.69 to -0.06],  $p$ -value < 0.05). Respective to our re-analysis procedures, first we conducted a random-effects analysis via *rma()* function in *metafor* package to reproduce their original analysis ("RE" in Table 1). Second, we (inappropriately) assumed effects within studies are homogeneous and computed an "averaged" effect size and sampling variance for each study (assuming sampling correlation  $\rho_s = 0.5$ ) and used them for subsequent random-effects models ("Average"). Third, we randomly selected one effect size from each study and used it for the subsequent random-effects model ("Sample"). Fourth, we used a multilevel meta-analytic model specifying between-study and within-study random effects to directly model the dependency among effect sizes ("ML"). Sixth, building upon "ML" method, we accounted for the (potential) correlated sampling errors by approximating a VCV matrix with a constant  $\rho_s = 0.5$  ("ML-VCV"). Seventh, we used the RVE to defend the model misspecification (e.g., the arbitrarily assumed  $\rho_s = 0.5$ ) and make robust model inferences ("ML-VCV-RVE"). It should be noted that because we did not find any dataset suitable for multivariate models in our survey (e.g., meta-analyses with multiple outcomes within studies), Table 1 does not report results corresponding to multivariate models (but see Supplementary Material 3 for an illustration with a fictitious dataset).

Re-analyses based on the multilevel model (“ML”) found that the overall effect of SSRIs exposure has a similar magnitude compared to that obtained from the random-effects model (“RE”), but the overall effect was not statistically significant (SMD = -0.39, 95CI% = [-0.80 to 0.03],  $p$ -value = 0.06). This clearly demonstrates that the standard error has been underestimated and the subsequent hypothesis tests ( $p$ -value) were incorrect. Likewise, the between-study variances ( $\sigma_b^2$ ) derived from RE model (conventional practice) also have been overestimated. For example, the value of  $\sigma_b^2$  in the multilevel model was almost half of that in the random-effects model. Using the random-effects model to fit this dataset might lead to a wrong conclusion that the study level has a high amount of heterogeneity ( $I_b^2 = 69\%$ ). However, the study level only explains 27% of the total heterogeneity. The remaining 49% of the total heterogeneity is due to the within-study level variability. The results of the multilevel model with a VCV matrix are very similar to those of the multilevel model without a VCV matrix (“ML” vs. “ML-VCV”). The results of the multilevel model are also robust after defending against model misspecification (“ML-VCV” vs. “ML-VCV-RVE”). The multilevel model can quantify the degrees of dependency, that is,  $ICC$  (correlation of effect sizes within the same study) suggests that the true underlying effects are weakly correlated with each (0.136).

## 12. A hands-on tutorial of the advanced meta-analytic techniques

We provide an easy-to-implement tutorial to help researchers apply these sophisticated techniques outlined above. Broadly, our tutorial contains two parts. In Part I, we use the dataset of the above worked example to illustrate meta-analysis in the framework of the multilevel model (standard practices; *sections 6 to 8*; see also Assink and Wibbelink, 2016). We recommend every researcher employ these procedures in an animal meta-analysis, such that potentially misleading conclusions can be avoided. In Part II, we use a more complex dataset to show the implementation of the extended methods (recommended practices) outlined in *section 9*. This dataset comes from our published neuroscience meta-analysis (Lagisz et al., 2020), which examined cognition bias across 22 animal species using 71 studies with 459 effect sizes. Given that R language and environment are the most widely used software in animal meta-analyses (Supplementary Materials file 1), we use R code to demonstrate the implementations of these advanced methods. The dataset and R code are freely accessible at an archived repository (See Open Science section). We use R markdown to annotate R code, which allows researchers to easily understand and reproduce our examples, and also easily modify the sample R code to suit their own analyses. The sample R code is based on existing R packages, for example, *metafor*, *orchaRd* and *clubSandwich*. We provide guidance to show the R syntax of these packages implementing the advanced methods outlined in *sections 6 – 9*. The complete R coding-based tutorial can be found at Supplementary Materials file 2.

### **13. Conclusions and future perspectives**



We have profiled the meta-analytic practice in the field of neurobiology and behavioural research on animal models. Researchers in this field mainly rely on traditional meta-analytic techniques (i.e., fixed-effect and random-effects models) for quantitative evidence synthesis. However, the traditional meta-analytic techniques are very limited in handling complex animal datasets (e.g., hierarchical/correlated data structure), which are more prone to statistical issues (e.g., non-independent effect sizes and errors). Researchers should go beyond traditional meta-analytic techniques, embracing the multilevel model and other advanced methods (e.g., multivariate models and robust variance estimation). Currently, these advanced methods are rarely used in animal meta-analyses. We have illustrated the concepts, rationale, examples, and implementations of these advanced methods. We expect their applications to continue to increase in future quantitative evidence synthesis in animal studies, delivering more robust/reliable model estimates and new neurobiological insights.

Furthermore, these advanced methods can be further extended to more sophisticated meta-analytic techniques. For example, A network meta-analysis can rank the effectiveness of multiple treatments by incorporating indirect evidence across separate animal studies (for example, using evidence of  $T_A$  vs.  $T_C$  and  $T_B$  vs.  $T_C$  to infer  $T_A$  vs.  $T_B$ ; Riley et al., 2017). We can employ a meta-analytic mediation (causal) model, path analysis or structural equation modelling to identify how a focal variable mediates the response variable of interest (Cheung, 2022; Shadish and Sweeney, 1991). We can

also take advantage of an individual participant data (IPD) meta-analysis to circumvent Simpson's paradox (i.e., "aggregation" bias or ecological fallacy) and test hypotheses at the individual animal level rather than the study level (Kaufmann et al., 2016; Riley et al., 2010; van Aert, 2022). Besides meta-analysing data across studies, we can also conduct a meta-analysis within a single study to enhance statistical power (i.e., internal meta-analysis; Goh et al., 2016; Nakagawa and Santos, 2012) and across different meta-analyses to ask high-order questions (i.e., second-order-meta-analysis; Fanelli et al., 2017; Nakagawa et al., 2019). Importantly, researchers in animal meta-analyses should review methodological developments and applications of meta-analytic techniques in other fields. In this regard, researchers can harness more appropriate and powerful meta-analytic techniques to gain not only new neurobiological insights, but methodological and meta-scientific insights. Ultimately, the use of these advancing meta-analyses can lead to better animal-to-human translation of this new knowledge.

### **Funding**

We thank the Faculty of Science and the office of Deputy Vice-chancellor of Research, UNSW, Sydney for the support to YY and SN. YY and JP were funded by the National Natural Science Foundation of China (NO. 31972609 and 32102597) and China Agriculture Research System (CARS-40). SN and ML were funded by the Australian Research Council Discovery Grant (DP210100812).

**CRedit authorship contribution statement**

Yefeng Yang: conceptualization, methodology, literature survey, formal analysis, original draft, review & editing, tutorial preparation. Malcom Macleod: review & editing. Jinming Pan: review & editing, funding acquisition. Malgorzata Lagisz: literature survey, formal analysis, review & editing, tutorial preparation, supervision. Shinichi Nakagawa: conceptualization, methodology, review & editing, supervision, project administration, funding acquisition.

**Conflicts of interest**

The authors declare no conflict of interest.

**Data availability**

Survey method and results, online tutorial (R scripts) and data to reproduce the examples presented in the current article and online tutorial are archived at the GitHub repository ([https://github.com/Yefeng0920/advanced\\_animal\\_MA\\_tutorial](https://github.com/Yefeng0920/advanced_animal_MA_tutorial)) and Zenodo repository (Yefeng, & Malgorzata Lagisz. (2022).

Yefeng0920/advanced\_animal\_MA\_tutorial: A tutorial of advanced methods for the meta-analyses of animal models (v1.5.0). Zenodo.

<https://doi.org/10.5281/zenodo.7314683>)

**ORCID**

Yefeng Yang: 0000-0002-8610-4016.

Malcolm Macleod: 0000-0001-9187-9839.

Jinming Pan: 0000-0003-0228-4564.

Malgorzata Lagisz: 0000-0002-3993-6127.

Shinichi Nakagawa: 0000-0002-7765-5182.

**Appendix A. Supporting information**

Supplementary data associated with this article can be found in the online version.

**Reference**

- Aarts, E., Verhage, M., Veenliet, J.V., Dolan, C.V., Van Der Sluis, S., 2014. A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience* 17, 491-496.
- Abbott, K.N., Arnott, C.K., Westbrook, R.F., Tran, D.M., 2019. The effect of high fat, high sugar, and combined high fat-high sugar diets on spatial learning and memory in rodents: A meta-analysis. *Neuroscience & Biobehavioral Reviews* 107, 399-421.
- Aloe, A.M., Becker, B.J., Pigott, T.D., 2010. An alternative to R<sup>2</sup> for assessing linear models of effect size. *Research Synthesis Methods* 1, 272-283.
- Assink, M., Wibbelink, C.J., 2016. Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology* 12, 154-174.

- Augusteijn, H.E., van Aert, R., van Assen, M.A., 2019. The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological methods* 24, 116-134.
- Bahadoran, Z., Mirmiran, P., Kashfi, K., Ghasemi, A., 2020. Importance of systematic reviews and meta-analyses of animal studies: challenges for animal-to-human translation. *Journal of the American Association for Laboratory Animal Science* 59, 469-477.
- Baldez, D.P., Biazus, T.B., Rabelo-da-Ponte, F.D., Nogaro, G.P., Martins, D.S., Kunz, M., Czepielewski, L.S., 2021. The effect of antipsychotics on the cognitive performance of individuals with psychotic disorders: network meta-analyses of randomized controlled trials. *Neuroscience & Biobehavioral Reviews*.
- Bannach-Brown, A., Hair, K., Bahor, Z., Soliman, N., Macleod, M., Liao, J., 2021. Technological advances in preclinical meta-research. *BMJ Open Science* 5, e100131.
- Barha, C.K., Falck, R.S., Davis, J.C., Nagamatsu, L.S., Liu-Ambrose, T., 2017. Sex differences in aerobic exercise efficacy to improve cognition: a systematic review and meta-analysis of studies in older rodents. *Frontiers in neuroendocrinology* 46, 86-105.
- Bates, D., Kliegl, R., Vasishth, S., Baayen, H., 2015. Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.
- Bird, S.M., Sohrabi, H.R., Sutton, T.A., Weinborn, M., Rainey-Smith, S.R., Brown, B., Patterson, L., Taddei, K., Gupta, V., Carruthers, M., 2016. Cerebral amyloid- $\beta$  accumulation and deposition following traumatic brain injury—a narrative review and meta-analysis of animal studies. *Neuroscience & Biobehavioral Reviews* 64, 215-228.

Boca, S.M., Pfeiffer, R.M., Sampson, J.N., 2017. Multivariate meta-analysis with an increasing number of parameters. *Biometrical Journal* 59, 496-510.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* 24, 127-135.

Bonapersona, V., Joels, M., Sarabdjitsingh, R., 2018. Effects of early life stress on biochemical indicators of the dopaminergic system: a 3 level meta-analysis of rodent studies. *Neuroscience & Biobehavioral Reviews* 95, 1-16.

Borenstein, M., Higgins, J.P., Hedges, L.V., Rothstein, H.R., 2017. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research synthesis methods* 8, 5-18.

Burgueno, A.L., Juárez, Y.R., Genaro, A.M., Tellechea, M.L., 2020. Prenatal stress and later metabolic consequences: Systematic review and meta-analysis in rodents. *Psychoneuroendocrinology* 113, 104560.

Chalmers, I., Haynes, B., 1994. Systematic Reviews: Reporting, updating, and correcting systematic reviews of the effects of health care. *Bmj* 309, 862-865.

Chamberlain, S.A., Hovick, S.M., Dibble, C.J., Rasmussen, N.L., Van Allen, B.G., Maitner, B.S., Ahern, J.R., Bell-Dereske, L.P., Roy, C.L., Meza-Lopez, M., 2012.

Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. *Ecology Letters* 15, 627-636.

- Cheung, M.W.-L., 2013. Multivariate meta-analysis as structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal* 20, 429-454.
- Cheung, M.W.-L., 2014. Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods* 19, 211.
- Cheung, M.W.-L., 2015. metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in psychology* 5, 1521.
- Cheung, M.W.-L., 2019. A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review* 29, 387-396.
- Cheung, M.W., 2022. Synthesizing indirect effects in mediation models with meta-analytic methods. *Alcohol and Alcoholism* 57, 5-15.
- Cinar, O., Nakagawa, S., Viechtbauer, W., 2021a. Phylogenetic multilevel meta-analysis: A simulation study on the importance of modeling the phylogeny. *Methods in Ecology and Evolution*.
- Cinar, O., Umbanhowar, J., Hoeksema, J.D., Viechtbauer, W., 2021b. Using information-theoretic approaches for model selection in meta-analysis. *Research Synthesis Methods*.
- Creutzberg, K.C., Sanson, A., Viola, T.W., Marchisella, F., Begni, V., Grassi-Oliveira, R., Riva, M.A., 2021. Long-lasting effects of prenatal stress on HPA axis and inflammation: a systematic review and multilevel meta-analysis in rodent studies. *Neuroscience & Biobehavioral Reviews*.

Currie, G.L., Angel-Scott, H.N., Colvin, L., Cramond, F., Hair, K., Khandoker, L., Liao, J., Macleod, M., McCann, S.K., Morland, R., 2019. Animal models of chemotherapy-induced peripheral neuropathy: a machine-assisted systematic review and meta-analysis. *PLoS biology* 17, e3000243.

de Vries, R.B., Wever, K.E., Avey, M.T., Stephens, M.L., Sena, E.S., Leenaars, M., 2014. The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *ILAR journal* 55, 427-437.

Egan, K.J., Janssen, H., Sena, E.S., Longley, L., Speare, S., Howells, D.W., Spratt, N.J., Macleod, M.R., Mead, G.E., Bernhardt, J., 2014. Exercise reduces infarct volume and facilitates neurobehavioral recovery: results from a systematic review and meta-analysis of exercise in experimental models of focal ischemia.

*Neurorehabilitation and neural repair* 28, 800-812.

England, T.J., Hind, W.H., Rasid, N.A., O'sullivan, S.E., 2015. Cannabinoids in experimental stroke: a systematic review and meta-analysis. *Journal of Cerebral Blood Flow & Metabolism* 35, 348-358.

Fanelli, D., Costas, R., Ioannidis, J.P., 2017. Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences* 114, 3714-3719.

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S.N., Onghena, P., Van den Noortgate, W., 2021. Detecting selection bias in meta-analyses with multiple outcomes: a simulation study. *The Journal of Experimental Education* 89, 125-144.



- Figueiredo, P.R., Tolomeo, S., Steele, J.D., Baldacchino, A., 2020. Neurocognitive consequences of chronic cannabis use: a systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews* 108, 358-369.
- Fisher, Z., Tipton, E., 2015. robumeta: An R-package for robust variance estimation in meta-analysis. arXiv preprint arXiv:1503.02220.
- Frantzias, J., Sena, E.S., Macleod, M.R., Salman, R.A.S., 2011. Treatment of intracerebral hemorrhage in animal models: meta-analysis. *Annals of neurology* 69, 389-399.
- Gasparini, A., Armstrong, B., 2011. Multivariate meta-analysis: A method to summarize non-linear associations. *Statistics in Medicine* 30, 2504-2506.
- Gasparini, A., Armstrong, B., Kenward, M.G., 2012. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in medicine* 31, 3821-3839.
- Gleser, L.J., Olkin, I., 2009. Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York: Russell Sage Foundation.
- Goh, J.X., Hall, J.A., Rosenthal, R., 2016. Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass* 10, 535-549.
- Gomes, D.G., 2021. Including random effects in statistical models in ecology: fewer than five levels? bioRxiv.

Grainger, M.J., Bolam, F.C., Stewart, G.B., Nilsen, E.B., 2020. Evidence synthesis for tackling research waste. *Nature ecology & evolution* 4, 495-497.

Greek, R., Menache, A., 2013. Systematic reviews of animal models: methodology versus epistemology. *International journal of medical sciences* 10, 206.

Gurevitch, J., Koricheva, J., Nakagawa, S., Stewart, G., 2018. Meta-analysis and the science of research synthesis. *Nature* 555, 175-182.

Gurka, M.J., 2006. Selecting the best linear mixed model under REML. *The American Statistician* 60, 19-26.

Hadfield, J., Nakagawa, S., 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology* 23, 494-508.

Haggarty, S.J., Karmacharya, R., Perlis, R.H., 2021. Advances toward precision medicine for bipolar disorder: mechanisms & molecules. *Molecular psychiatry* 26, 168-185.

Hedges, L.V., 1982. Estimation of effect size from a series of independent experiments. *Psychological bulletin* 92, 490.

Hedges, L.V., Gurevitch, J., Curtis, P.S., 1999. The meta-analysis of response ratios in experimental ecology. *Ecology* 80, 1150-1156.

Hedges, L.V., Pigott, T.D., 2004. The power of statistical tests for moderators in meta-analysis. *Psychological methods* 9, 426-445.

- Hedges, L.V., Tipton, E., Johnson, M.C., 2010. Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods* 1, 39-65.
- Hentschke, H., Stüttgen, M.C., 2011. Computation of measures of effect size for neuroscience data sets. *European Journal of Neuroscience* 34, 1887-1894.
- Hieronimus, F., Hieronimus, M., Nilsson, S., Eriksson, E., Østergaard, S., 2020. Individual variability in treatment response to antidepressants in major depression: comparing trial-level and patient-level analyses. *Acta Psychiatrica Scandinavica*.
- Higgins, J.P., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine* 21, 1539-1558.
- Higgins, J.P., Thompson, S.G., Deeks, J.J., Altman, D.G., 2003. Measuring inconsistency in meta-analyses. *Bmj* 327, 557-560.
- Hooijmans, C.R., Donders, R., Magnuson, K., Wever, K.E., Ergün, M., Rooney, A.A., Walker, V., Langendam, M.W., 2022. Assessment of key characteristics, methodology and effect size measures used in meta-analysis of human-health-related animal studies. *Research Synthesis Methods*.
- Hooijmans, C.R., IntHout, J., Ritskes-Hoitinga, M., Rovers, M.M., 2014. Meta-analyses of animal studies: an introduction of a valuable instrument to further improve healthcare. *ILAR journal* 55, 418-426.
- Houle, D., Pélabon, C., Wagner, G.P., Hansen, T.F., 2011. Measurement and meaning in biology. *The quarterly review of biology* 86, 3-34.

- Hunniford, V.T., Montroy, J., Fergusson, D.A., Avey, M.T., Wever, K.E., McCann, S.K., Foster, M., Fox, G., Lafreniere, M., Ghaly, M., 2021. Epidemiology and reporting characteristics of preclinical systematic reviews. *PLoS Biology* 19, e3001177.
- IntHout, J., Ioannidis, J.P., Rovers, M.M., Goeman, J.J., 2016. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ open* 6, e010247.
- Jackson, D., Riley, R., White, I.R., 2011. Multivariate meta-analysis: potential and promise. *Statistics in medicine* 30, 2481-2498.
- Jackson, D., White, I.R., Riley, R.D., 2012. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in medicine* 31, 3805-3820.
- Jak, S., Cheung, M.W.-L., 2020. Meta-analytic structural equation modeling with moderating effects on SEM parameters. *Psychological methods* 25, 430-455.
- Jamshidi, L., Declercq, L., Fernández-Castilla, B., Ferron, J.M., Moeyaert, M., Beretvas, S.N., Van den Noortgate, W., 2020. Multilevel meta-analysis of multiple regression coefficients from single-case experimental studies. *Behavior research methods* 52, 2008-2019.
- Joshi, M., Pustejovsky, J.E., Beretvas, S.N., 2022. Cluster wild bootstrapping to handle dependent effect sizes in meta-analysis with a small number of studies. *Research Synthesis Methods* 13, 457-477.
- Kafkafi, N., Agassi, J., Chesler, E.J., Crabbe, J.C., Crusio, W.E., Eilam, D., Gerlai, R., Golani, I., Gomez-Marin, A., Heller, R., 2018. Reproducibility and replicability of

rodent phenotyping in preclinical studies. *Neuroscience & Biobehavioral Reviews* 87, 218-232.

Kara, N., Stukalin, Y., Einat, H., 2018. Revisiting the validity of the mouse forced swim test: Systematic review and meta-analysis of the effects of prototypic antidepressants. *Neuroscience & Biobehavioral Reviews* 84, 1-11.

Kaufmann, E., Reips, U.-D., Merki, K.M., 2016. Avoiding methodological biases in meta-analysis. *Zeitschrift für Psychologie*.

Khorshidi, F., Poljak, A., Liu, Y., Lo, J.W., Crawford, J.D., Sachdev, P.S., 2021.

Resveratrol: A “miracle” drug in neuropsychiatry or a cognitive enhancer for mice only? A systematic review and meta-analysis. *Ageing Research Reviews* 65, 101199.

Knapp, G., Hartung, J., 2003. Improved tests for a random effects meta-regression with a single covariate. *Statistics in medicine* 22, 2693-2710.

Konstantopoulos, S., 2011. Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods* 2, 61-76.

Koricheva, J., Kulinskaya, E., 2019. Temporal instability of evidence base: a threat to policy making? *Trends in ecology & evolution* 34, 895-902.

Kvarven, A., Strømland, E., Johannesson, M., 2020. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour* 4, 423-434.

- Lages, Y., Rossi, A., Krahe, T., Landeira-Fernandez, J., 2021. Effect of chronic unpredictable mild stress on the expression profile of serotonin receptors in rats and mice: a meta-analysis. *Neuroscience & Biobehavioral Reviews*.
- Lagisz, M., Zidar, J., Nakagawa, S., Neville, V., Sorato, E., Paul, E.S., Bateson, M., Mendl, M., Løvlie, H., 2020. Optimism, pessimism and judgement bias in animals: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews* 118, 3-17.
- Lajeunesse, M.J., 2011. On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology* 92, 2049-2055.
- Langan, D., Higgins, J.P., Jackson, D., Bowden, J., Veroniki, A.A., Kontopantelis, E., Viechtbauer, W., Simmonds, M., 2019. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research synthesis methods* 10, 83-98.
- Leffa, D.T., Panzenhagen, A.C., Salvi, A.A., Bau, C.H., Pires, G.N., Torres, I.L., Rohde, L.A., Rovaris, D.L., Grevet, E.H., 2019. Systematic review and meta-analysis of the behavioral effects of methylphenidate in the spontaneously hypertensive rat model of attention-deficit/hyperactivity disorder. *Neuroscience & Biobehavioral Reviews* 100, 166-179.
- Lenth, R.V., 2016. Least-squares means: the R package lsmeans. *Journal of statistical software* 69, 1-33.

- López-López, J.A., Van den Noortgate, W., Tanner-Smith, E.E., Wilson, S.J., Lipsey, M.W., 2017. Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: A Monte Carlo simulation. *Research synthesis methods* 8, 435-450.
- Lorenzo-Luaces, L., Peipert, A., Romero, R.D.J., Rutter, L.A., Rodriguez-Quintana, N., 2021. Personalized medicine and cognitive behavioral therapies for depression: Small effects, big problems, and bigger data. *International Journal of Cognitive Therapy* 14, 59-85.
- Luedtke, A., Kessler, R.C., 2021. New Directions in Research on Heterogeneity of Treatment Effects for Major Depression. *JAMA psychiatry*.
- Macaskill, P., Walter, S.D., Irwig, L., 2001. A comparison of methods to detect publication bias in meta-analysis. *Stat Med* 20, 641-654.
- Maire, A., Thierry, E., Viechtbauer, W., Daufresne, M., 2019. Poleward shift in large-river fish communities detected with a novel meta-analysis framework. *Freshwater Biology* 64, 1143-1156.
- Mancini, M., Karakuzu, A., Cohen-Adad, J., Cercignani, M., Nichols, T.E., Stikov, N., 2020. An interactive meta-analysis of MRI biomarkers of myelin. *Elife* 9, e61523.
- Marin-Martinez, F., Sánchez-Meca, J., 2010. Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement* 70, 56-73.

- Maslej, M.M., Furukawa, T.A., Cipriani, A., Andrews, P.W., Sanches, M., Tomlinson, A., Volkmann, C., McCutcheon, R.A., Howes, O., Guo, X., 2021. Individual differences in response to antidepressants: a meta-analysis of placebo-controlled randomized clinical trials. *JAMA psychiatry* 78, 490-497.
- Mauvais-Jarvis, F., Merz, N.B., Barnes, P.J., Brinton, R.D., Carrero, J.-J., DeMeo, D.L., De Vries, G.J., Epperson, C.N., Govindan, R., Klein, S.L., 2020. Sex and gender: modifiers of health, disease, and medicine. *The Lancet* 396, 565-582.
- McCann, S.K., Cramond, F., Macleod, M.R., Sena, E.S., 2016. Systematic review and meta-analysis of the efficacy of interleukin-1 receptor antagonist in animal models of stroke: an update. *Translational stroke research* 7, 395-406.
- Mills, H.L., Higgins, J.P., Morris, R.W., Kessler, D., Heron, J., Wiles, N., Smith, G.D., Tilling, K., 2021. Detecting heterogeneity of intervention effects using analysis and meta-analysis of differences in variance between trial arms. *Epidemiology (Cambridge, Mass.)* 32, 846.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine* 6, e1000097.
- Moreno, S.G., Sutton, A.J., Ades, A.E., Stanley, T.D., Abrams, K.R., Peters, J.L., Cooper, N.J., 2009. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *Bmc Med Res Methodol* 9.



- Moulin, T.C., Amaral, O.B., 2020. Using collaboration networks to identify authorship dependence in meta-analysis results. *Research Synthesis Methods* 11, 655-668.
- Mueller, K.F., Briel, M., Strech, D., Meerpohl, J.J., Lang, B., Motschall, E., Gloy, V., Lamontagne, F., Bassler, D., 2014. Dissemination bias in systematic reviews of animal research: a systematic review. *PloS one* 9, e116016.
- Nakagawa, S., Cuthill, I.C., 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological reviews* 82, 591-605.
- Nakagawa, S., Koricheva, J., Macleod, M., Viechtbauer, W., 2020. *Introducing our series: research synthesis and meta-research in biology*. Springer.
- Nakagawa, S., Lagisz, M., Jennions, M.D., Koricheva, J., Noble, D.W., Parker, T.H., Sánchez-Tójar, A., Yang, Y., O'Dea, R.E., 2021a. Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods in Ecology and Evolution*.
- Nakagawa, S., Lagisz, M., O'Dea, R.E., Rutkowska, J., Yang, Y., Noble, D.W., Senior, A.M., 2021b. The orchard plot: Cultivating a forest plot for use in ecology, evolution, and beyond. *Research Synthesis Methods* 12, 4-12.
- Nakagawa, S., Noble, D.W., Senior, A.M., Lagisz, M., 2017. Meta-evaluation of meta-analysis: ten appraisal questions for biologists. *BMC Biology* 15, 1-14.

- Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., Senior, A.M., 2015. Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution* 6, 143-152.
- Nakagawa, S., Samarasinghe, G., Haddaway, N.R., Westgate, M.J., O'Dea, R.E., Noble, D.W., Lagisz, M., 2019. Research weaving: visualizing the future of research synthesis. *Trends in ecology & evolution* 34, 224-238.
- Nakagawa, S., Santos, E.S., 2012. Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology* 26, 1253-1274.
- Nakagawa, S., Schielzeth, H., 2013. A general and simple method for obtaining  $R^2$  from generalized linear mixed-effects models. *Methods in ecology and evolution* 4, 133-142.
- Nakagawa, S., Senior, A.M., Viechtbauer, W., Noble, D.W., 2021c. An assessment of statistical methods for non-independent data in ecological meta-analyses: Comment. *Ecology*, e03490.
- Neville, V., Nakagawa, S., Zidar, J., Paul, E.S., Lagisz, M., Bateson, M., Løvlie, H., Mendl, M., 2020. Pharmacological manipulations of judgement bias: a systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews* 108, 269-286.
- Noble, D.W., Lagisz, M., O'dea, R.E., Nakagawa, S., 2017. Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses. *Molecular Ecology* 26, 2410-2425.

- Orsini, N., Li, R., Wolk, A., Khudyakov, P., Spiegelman, D., 2012. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. *American journal of epidemiology* 175, 66-73.
- Pound, P., Bracken, M.B., 2014. Is animal research sufficiently evidence based to be a cornerstone of biomedical research? *Bmj* 348.
- Pustejovsky, J.E., Tipton, E., 2018. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics* 36, 672-683.
- Pustejovsky, J.E., Tipton, E., 2022. Meta-analysis with Robust Variance Estimation: Expanding the range of working models. *Prevention Science* 23, 425-438.
- Ramsteijn, A.S., Van de Wijer, L., Rando, J., van Luijk, J., Homberg, J.R., Olivier, J.D., 2020. Perinatal selective serotonin reuptake inhibitor exposure and behavioral outcomes: a systematic review and meta-analyses of animal studies. *Neuroscience & Biobehavioral Reviews* 114, 53-69.
- Richter, S.H., Garner, J.P., Würbel, H., 2009. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nature methods* 6, 257-261.
- Riley, R.D., Higgins, J.P., Deeks, J.J., 2011. Interpretation of random effects meta-analyses. *Bmj* 342.
- Riley, R.D., Jackson, D., Salanti, G., Burke, D.L., Price, M., Kirkham, J., White, I.R., 2017. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *bmj* 358, j3932.

- Riley, R.D., Lambert, P.C., Abo-Zaid, G., 2010. Meta-analysis of individual participant data: rationale, conduct, and reporting. *Bmj* 340.
- Ritz, J., Demidenko, E., Spiegelman, D., 2008. Multivariate meta-analysis for data consortia, individual patient meta-analysis, and pooling projects. *Journal of Statistical Planning and Inference* 138, 1919-1933.
- Rodgers, M.A., Pustejovsky, J.E., 2021. Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological methods* 26, 141.
- Russell, L., 2021. emmeans: estimated marginal means, aka least-squares means. R package version 1.7. 1.
- Sánchez-Meca, J., Marín-Martínez, F., 2008. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological methods* 13, 31-48.
- Sánchez-Tójar, A., Moran, N.P., O’Dea, R.E., Reinhold, K., Nakagawa, S., 2020. Illustrating the importance of meta-analysing variances alongside means in ecology and evolution. *Journal of Evolutionary Biology* 33, 1216-1223.
- Sandercock, P., Roberts, I., 2002. Systematic reviews of animal experiments. *The Lancet* 360, 586.
- Schieffelin, H., 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution* 1, 103-113.
- Schmid, C.H., Stijnen, T., White, I., 2020. *Handbook of Meta-analysis*. CRC Press.

Schork, N.J., 2015. Personalized medicine: time for one-person trials. *Nature News* 520, 609.

Sena, E.S., Currie, G.L., McCann, S.K., Macleod, M.R., Howells, D.W., 2014. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *Journal of Cerebral Blood Flow & Metabolism* 34, 737-742.

Senior, A.M., Grueber, C.E., Kamiya, T., Lagisz, M., O'dwyer, K., Santos, E.S., Nakagawa, S., 2016. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. *Ecology* 97, 3293-3299.

Senior, A.M., Viechtbauer, W., Nakagawa, S., 2020. Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio,  $\ln\text{CVR}$ . *Research Synthesis Methods*, e176.

Shadish, W.R., Sweeney, R.B., 1991. Mediators and moderators in meta-analysis: there's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of consulting and clinical psychology* 59, 883-893.

Shields, G.S., Bonner, J.C., Moons, W.G., 2015. Does cortisol influence core executive functions? A meta-analysis of acute cortisol administration effects on working memory, inhibition, and set-shifting. *Psychoneuroendocrinology* 58, 91-103.

Soliman, N., Rice, A.S., Vollert, J., 2020. A practical guide to preclinical systematic review and meta-analysis. *Pain* 161, 1949.

- Song, C., Peacor, S.D., Osenberg, C.W., Bence, J.R., 2020. An assessment of statistical methods for nonindependent data in ecological meta-analyses. *Ecology* 101, e03184.
- Spake, R., Mori, A.S., Beckmann, M., Martin, P.A., Christie, A.P., Duguid, M.C., Doncaster, C.P., 2021. Implications of scale dependence for cross-study syntheses of biodiversity differences. *Ecology Letters* 24, 374-390.
- Stanley, T.D., Doucouliagos, H., Ioannidis, J.P., 2017. Finding the power to reduce publication bias. *Statistics in medicine* 36, 1580-1598.
- Sterne, J.A., Egger, M., Smith, G.D., 2001a. Investigating and dealing with publication and other biases in meta-analysis. *Bmj* 323, 101-105.
- Sterne, J.A., Egger, M., Smith, G.D., 2001b. Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *Brit Med J* 323, 101-105.
- Sterne, J.A., Gavaghan, D., Egger, M., 2000. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology* 53, 1119-1129.
- Stukalin, Y., Lan, A., Einat, H., 2020. Revisiting the validity of the mouse tail suspension test: systematic review and meta-analysis of the effects of prototypic antidepressants. *Neuroscience & Biobehavioral Reviews* 112, 39-47.
- Tannenbaum, C., Ellis, R.P., Eyssel, F., Zou, J., Schiebinger, L., 2019. Sex and gender analysis improves science and engineering. *Nature* 575, 137-146.

Tanner-Smith, E.E., Tipton, E., Polanin, J.R., 2016. Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology* 2, 85-112.

Tanriver-Ayder, E., Faes, C., van de Castele, T., McCann, S.K., Macleod, M.R., 2021. Comparison of commonly used methods in random effects meta-analysis: application to preclinical data in drug discovery research. *BMJ open science* 5, e100074.

Thomas, R.E., Ramsay, C.R., McAuley, L., Grimshaw, J.M., 2003. Unit of analysis errors should be clarified in meta-analyses. *BMJ* 326, 397.

Tipton, E., 2013. Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods* 4, 169-187.

Tipton, E., 2015. Small sample adjustments for robust variance estimation with meta-regression. *Psychological methods* 20, 375-393.

Tipton, E., Pustejovsky, J.E., 2015. Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics* 40, 604-634.

Usui, T., Macleod, M.R., McCann, S.K., Senior, A.M., Nakagawa, S., 2021. Meta-analysis of variation suggests that embracing variability improves both replicability and generalizability in preclinical research. *PLoS Biology* 19, e3001009.

van Aert, R.C., 2022. Analyzing Data of a Multilab Replication Project With Individual Participant Data Meta-Analysis. *Zeitschrift für Psychologie*.

- van Aert, R.C., Schmid, C.H., Svensson, D., Jackson, D., 2021. Study specific prediction intervals for random-effects meta-analysis: A tutorial: Prediction intervals in meta-analysis. *Research synthesis methods* 12, 429-447.
- Van Aert, R.C., Van Assen, M.A., Viechtbauer, W., 2019a. Statistical properties of methods based on the Q-statistic for constructing a confidence interval for the between-study variance in meta-analysis. *Research synthesis methods* 10, 225-239.
- Van Aert, R.C., Wicherts, J.M., Van Assen, M.A., 2019b. Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PloS one* 14, e0215052.
- Van den Noortgate, W., López-López, J.A., Marín-Martínez, F., Sánchez-Meca, J., 2013. Three-level meta-analysis of dependent effect sizes. *Behavior research methods* 45, 576-594.
- Vendl, C., Pottier, P., Taylor, M.D., Braeunig, J., Gibson, M.J., Hesselson, D., Neely, G.G., Lagisz, M., Nakagawa, S., 2021. Thermal processing reduces PFAS concentrations in blue food—A systematic review and meta-analysis.
- Vesterinen, H., Sena, E., Egan, K., Hirst, T., Churolov, L., Currie, G., Antonic, A., Howells, D., Macleod, M., 2014. Meta-analysis of data from animal studies: a practical guide. *Journal of neuroscience methods* 221, 92-102.
- Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. *Journal of statistical software* 36, 1-48.



- Viechtbauer, W., López-López, J.A., Sánchez-Meca, J., Marín-Martínez, F., 2015. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods* 20, 360-374.
- Voelkl, B., Altman, N.S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., Karp, N.A., Kas, M.J., Schielzeth, H., Van de Castele, T., 2020. Reproducibility of animal research in light of biological variation. *Nature Reviews Neuroscience* 21, 384-393.
- Volkman, C., Volkman, A., Müller, C.A., 2020. On the treatment effect heterogeneity of antidepressants in major depression: A Bayesian meta-analysis and simulation study. *PloS one* 15, e0241497.
- Wang, Q., Liao, J., Hair, K., Bannach-Brown, A., Bahor, Z., Currie, G.L., McCann, S.K., Howells, D.W., Sena, E.S., Macleod, M.R., 2018. Estimating the statistical performance of different approaches to meta-analysis of data from animal studies in identifying the impact of aspects of study design. *Biorxiv*, 256776.
- Welz, T., Viechtbauer, W., Pauly, M., 2023. Cluster-robust estimators for multivariate mixed-effects meta-regression. *Computational Statistics & Data Analysis* 179, 107631.
- Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I., Nakagawa, S., 2022a. Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Global Change Biology* 28, 969-989.

Yang, Y., Lagisz, M., Foo, Y.Z., Noble, D.W., Anwer, H., Nakagawa, S., 2021.

Beneficial intergenerational effects of exercise on brain and cognition: a multilevel meta-analysis of mean and variance. *Biological Reviews*.

Yang, Y., Sánchez-Tójar, A., O'Dea, R.E., Noble, D., Koricheva, J., Jennions, M.D.,

Parker, T.H., Lagisz, M., Nakagawa, S., 2022b. Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology.

Zajitschek, S.R., Zajitschek, F., Bonduriansky, R., Brooks, R.C., Cornwell, W.,

Falster, D.S., Lagisz, M., Mason, J., Senior, A.M., Noble, D.W., 2020. Sexual dimorphism in trait variability and its eco-evolutionary and statistical implications. *eLife* 9, e63170.

Zoerle, T., Ilodigwe, D.C., Wan, H., Lakovic, K., Sabri, M., Ai, J., Macdonald, R.L.,

2012. Pharmacologic reduction of angiographic vasospasm in experimental subarachnoid hemorrhage: systematic review and meta-analysis. *Journal of Cerebral Blood Flow & Metabolism* 32, 1645-1658.

## Tables

**Table 1** Comparison of different methods dealing with non-independence of effect sizes in terms of model coefficient estimates, confidence intervals (CIs), corresponding hypothesis tests and variance component estimates. *ICC* = intra-class correlation denoting the degree of dependence/correlation within clustering groups

(i.e., study).  $\rho_s$  = correlation of sampling errors (sampling level). RE = fit a random-effects meta-analytic model ignoring the dependency among effect sizes and treating them as if they were statistically independent (assuming  $ICC = 0$ ). Average-FE = fit a random-effects model but after aggregating effect sizes and sampling variance (assuming sampling correlation  $\rho_s = 0.5$  in this example) for each study (assuming  $ICC = 1$ , namely, homogeneous effect sizes within studies). Sample-FE = fit a random-effects model but after randomly sampling a single effect size from each study. ML = fit a multilevel meta-analytic model directly modelling the dependency among effect sizes. ML-VCV = fit a multilevel meta-analytic model with a VCV matrix accounting for correlation  $\rho_s$  (assumed to be 0.5) of sampling errors or observed effect size estimates (sampling level). ML-VCV-RVE = use robust variance estimation (RVE) to guard against model misspecification of ML-VCV. SE = standard error of the pooled SMD or cluster-robust standard error if applying RVE. LCI = 95% lower confidence interval (CI) or lower CI based on cluster-robust standard error if applying RVE. UCI = 95% upper CI or upper CI based on cluster-robust standard error if applying RVE.  $\sigma_b^2$  = Between-study variance.  $\sigma_w^2$  = Within-study variance.  $I_b^2$  = Between-study heterogeneity.  $I_w^2$  = Within-study heterogeneity.

Parameter estimates	Conventional practices			Default practices	Optional practices	
	RE	Average-FE	Sample-FE	ML	ML-VCV	ML-VCV-RVE
Pooled SMD	-0.38	-0.37	-0.52	-0.39	-0.39	-0.39
SE	0.163	0.211	0.243	0.20	0.195	0.193
<i>p</i> -value	0.03	0.11	0.06	0.06	0.06	0.07
LCI	-0.72	-0.84	-1.05	-0.80	-0.80	-0.82

UCI	-0.03	0.09	0.02	0.03	0.03	0.04
$\sigma_b^2$	0.287	0.397	0.543	0.146	0.055	0.055
$\sigma_w^2$				0.264	0.352	0.352
$I_b^2$	69%	79%	82%	27%	10%	10%
$I_w^2$				49%	66%	66%
ICC				0.356	0.136	0.136

### Figure legends

**Figure 1.** Schematic of nested experimental designs and clustered data structures in animal studies. Neurobiology and behavioural sciences often involve nested experimental designs, in which multiple traits are measured from a single experimental unit. For example, in an animal experiment where independent mice are randomly allocated to drug-treatment groups with different drug doses and a control group (placebo), multiple traits can be measured from one mouse for each group. Further, measurements can be taken repeatedly over time (longitudinal measures) or from the same body parts. All these can lead to multiple effect sizes per study/paper (effect sizes are correlated with each other within the same studies), which violates the critical assumption of statistical independence between effect sizes.

**Figure 2.** A diagram showing the computation and interpretation of three important but underappreciated effect sizes in the meta-analysis on animal data. Imagine we aim to screen potential anti-dementia drugs using a fear-conditioning test. To do so, we need to answer two questions. The first is “what is the average treatment effect of a drug?” The log-transformed response ratio (lnRR) can quantify the average treatment effect by comparing the ratio of means between the drug group and the placebo group.

The second is “do all animals respond in a similar way to the drug?” Here, the log-transformed variability ratio (lnVR) can be used to detect heterogeneous treatment effects. The log-transformed coefficient of variation ratio (lnCVR) is a mean-adjusted version of lnVR, in which the indirect effect from the mean is controlled for.  $\ln VR > 0$  or  $\ln CVR > 0$  indicate that the tested drug shows a selective efficacy (i.e., high inter-individual variability).

**Figure 3.** The summary of the main results of a survey of 62 “formal” meta-analyses using animal models mimicking human diseases, physiology, and behaviour (2011 – 2021). Summarised methodological and reporting practice regarding (A) effect size used, (B) statistical models employed, (C) heterogeneity indices, (D) sign of statistical non-independence (multiple effect sizes per animal study; the medium number of effect size per animal study  $k = 56$ , the medium number of animal studies per meta-analysis  $N = 25$ ,  $k/N$  ratio = 2.2), (E) addressing non-independent effect sizes, and (F) publication bias test methods. Barplots indicate numbers of papers with a given methodological and reporting characteristic. Plural answers were allowed (i.e. one paper could fit into more than one category/option, for example, both the random-effects model and meta-regression were used in Leffa et al., 2019, such papers were used more than once when tallying counts). We also surveyed other questions, such as whether and how the study accounted for heterogeneity and whether the authors of the meta-analysis acknowledged the presence of statistical non-independence. For the

detailed survey methods, questions, and complete results, see Supplementary

Materials file 1.

**Figure 4.** Four common drivers of statistical non-dependence in meta-analyses on animal data (A) – (D). Statistical non-independence means that effect sizes have a multilevel/nested structure and are correlated within a ‘cluster’ variable, where effect sizes may be more similar than those across ‘cluster’ variables. Ignoring non-independence might lead to a biased estimate of model coefficients, and underestimated standard errors and, consequently, hypothesis testing with incorrect Type I error rates and confidence intervals with incorrect coverage levels. ES = effect size (e.g., lnRR; Figure 2). Double-headed arrows = correlations in effect sizes. (A) Shared study identity where a single construct of interest (i.e., single outcome, response variable) is measured using different assays, instruments, or scales within the same study. For example, when three assays are simultaneously performed on animals from an antedementia – placebo comparison to quantify the magnitude of dementia, correlations occur among ES 1 to ES 3. It should be noted that multiple constructs within the same study (e.g., dementia and anxiety) are also possible. In such a case, (multivariate) effect sizes are better modelled by a multivariate than a multilevel model (see Section 9.2). (B) Shared animal identity where multiple effect sizes can be derived from a single animal cohort. For example, brain morphology can be measured for different parts of the same brain. Another example, fear condition test can be conducted on an animal at follow-up times (1, 3, and 5 days). Note that

shared animal identity also leads to correlated sampling errors because the same cohort animal will be used multiple times when computing effect sizes. (C) Shared control where multiple treatment groups are compared to the same control group. For example, when comparing a single placebo (control) to two doses of the antidementia drug, correlations in effect sizes occur because data from the placebo group is repeatedly used to calculate effect sizes. Note that shared control also contributes to correlated sampling errors. (D) Shared species or evolutionary history – effect sizes taken from the same species may be more similar. For example, if three effect sizes were calculated from rabbits, they would be correlated with one another because of the similarity of individuals within the species. Moreover, ES 1 to ES 3 shown in panel D are not independent because of the phylogenetic relatedness (where mice and rabbits are more similar to each other than to dogs). See the main text for real examples for each scenario.

**Figure 5.** Common approaches used to handle statistical non-independence among effect sizes in meta-analyses on animal data. ES = effect size (SMD or lnRR;  $ES_{[i]}$  means  $i$ th effect size in  $j$ th study).  $v$  = sampling variance ( $v_{[i]}$  means sampling variance of  $i$ th effect size in  $j$ th study). (A) Following the hypothetical data set with six non-independent effect sizes from three studies (panel (C) in Figure 2), effect sizes are not independent in Studies 1 and 3 due to shared control (i.e., multiple doses). (B) Researchers completely ignore non-independence by using a fixed- and random-effect model to fit these non-independent effect sizes, treating them as if they were

statistically independent. (C) Researchers first compute the simple average (weighted average) of multiple effect sizes for Studies 1 and 3 and subsequently use it in a fixed- or random-effects model. (D) Researchers first select one random effect size from Studies 1 and 3 and subsequently use it in a fixed- or random-effects model. As clearly shown in panels C and D, eliminating non-independence could lead to the loss of information (in this case, within-study moderator: doses of the antidepressant). See the main text for real examples for each approach.

**Figure 6.** Schematic illustrations of the fixed-effect model (A), random-effects model (B), and hierarchical structure of a multilevel model (C). Imagine we aim to assess the efficacy of a new antedementia drug. To properly estimate the generalizability of the efficacy, we employ a multilevel meta-analytic model to aggregate effect sizes derived from different animal models. In the sampling level, we derived mean, standard deviations, and sample size to calculate effect sizes  $ES_{[i]}$ . In this level, the only deviation between the estimate of  $ES_{[i]}$  and true effect is the sampling error effect  $e_{[i]}$ . In the within-study/effect size level, effect sizes are not independent (see Figure 1). The multilevel model uses a random-effects term – effect-size specific effect  $u_{w[i]}$  to account for within-study variability.  $u_{w[i]}$  allows  $ES_i$  varies within studies and follows a normal distribution with mean  $E[u_{w[i]}] = 0$  and variance  $\text{Var}[u_{w[i]}] = \sigma_w^2$ . In the between-study level, the true effect of each study is aggregated to generate the overall efficacy of the antedementia drug (overall mean effect/pooled SMD:  $E[ES_{[i]}] = \beta_0$ ). In this level, the multilevel model uses another



random-effects term – study-specific effect  $u_{b[j]}$  to account for between-study variability.  $u_{b[j]}$  allows  $ES_i$  varies between studies and follows a normal distribution with mean  $E[u_{b[j]}] = 0$  and variance  $\text{Var}[u_{b[j]}] = \sigma_b^2$ .

**Figure 7.** A comparison of variance components of a random-effects model and a multilevel model. In a random-effects model, the total variance ( $\sigma_{total}^2$ ) can be partitioned into two components: “typical” sampling variance ( $\sigma_{sampling}^2$ ; see Equation 7 for formula) and between-study variance ( $\tau^2$ ). In the multilevel model, the  $\sigma_{total}^2$  can be decomposed into three components:  $\sigma_{sampling}^2$ , within-study variance ( $\sigma_w^2$ ) and between-study variance ( $\sigma_b^2$ ). Using a random-effects model to fit non-independent data could misassign  $\sigma_w^2$  to  $\sigma_b^2$  (see Supplementary Materials file 2 for a real example). Importantly, a random-effects model can only use between-study moderator variables to explain the heterogeneity in true effect, while a multilevel model can use both within- and between-study moderator variable to explain the heterogeneity in true effect.

**Figure 8.** A diagram showing a typical example of an interactive effect between two moderator variables in meta-analyses on animal data. (A) No interactive effect between the dose of an antedementia drug and animal sex. (B) Interactive effect between the dose of an antedementia drug and animal sex. When an interactive effect exists, the dose-response to an antedementia drug is dependent on animal sex (i.e., the effect of the dose is confounded by animal sex). The steep slope of the regression line

for females indicates a strong dose-response relationship of the antedementia drug.

The near-flat slope for males indicates that there is no dose-response relationship.

## Figures

Figure 1

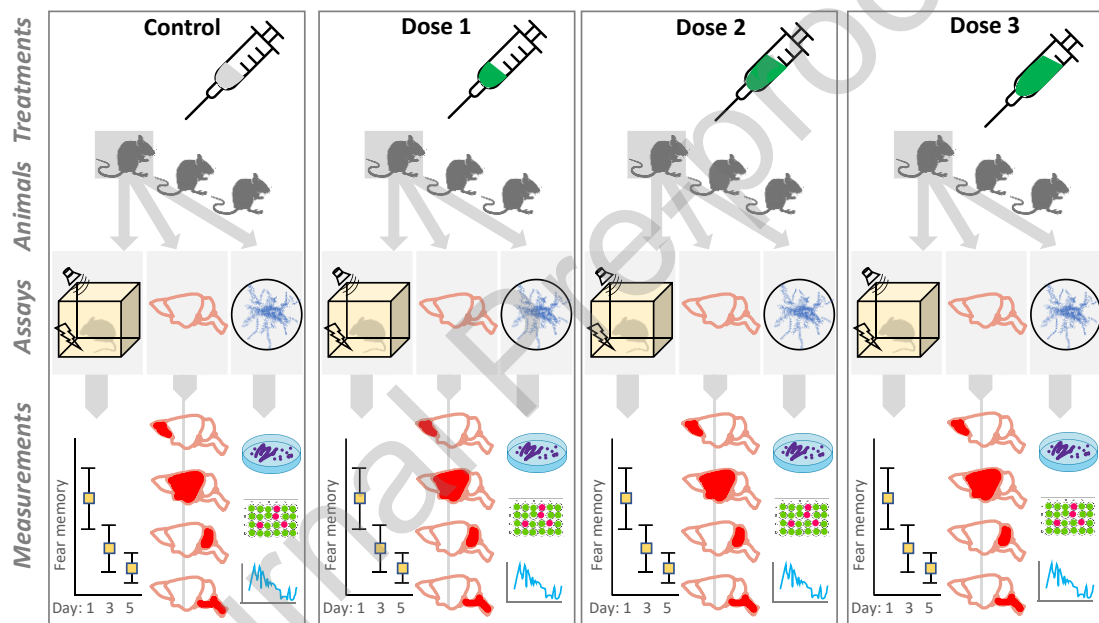


Figure 2

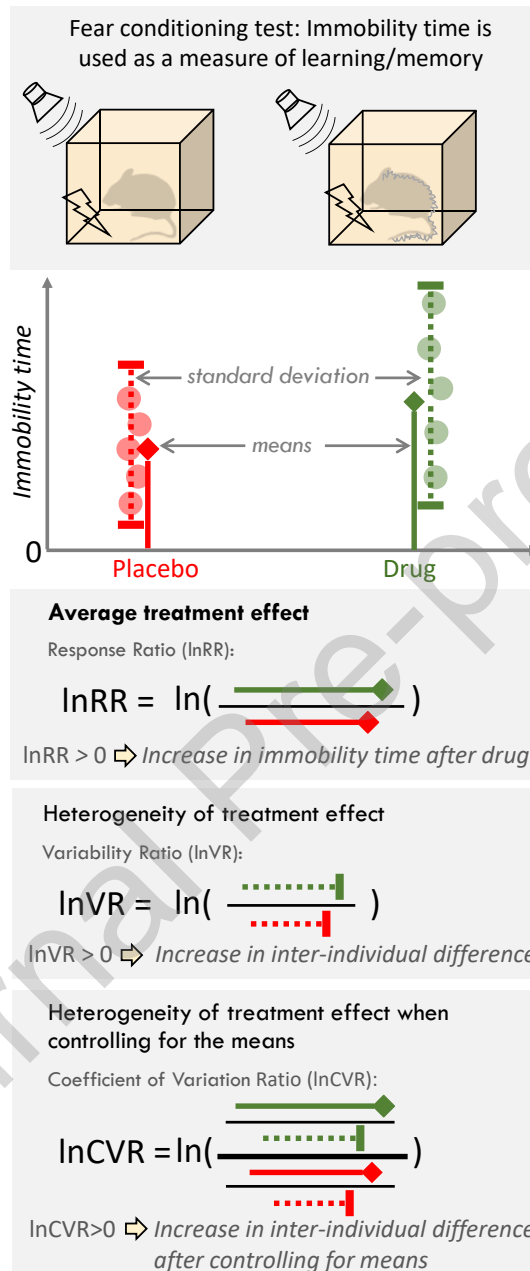


Figure 3

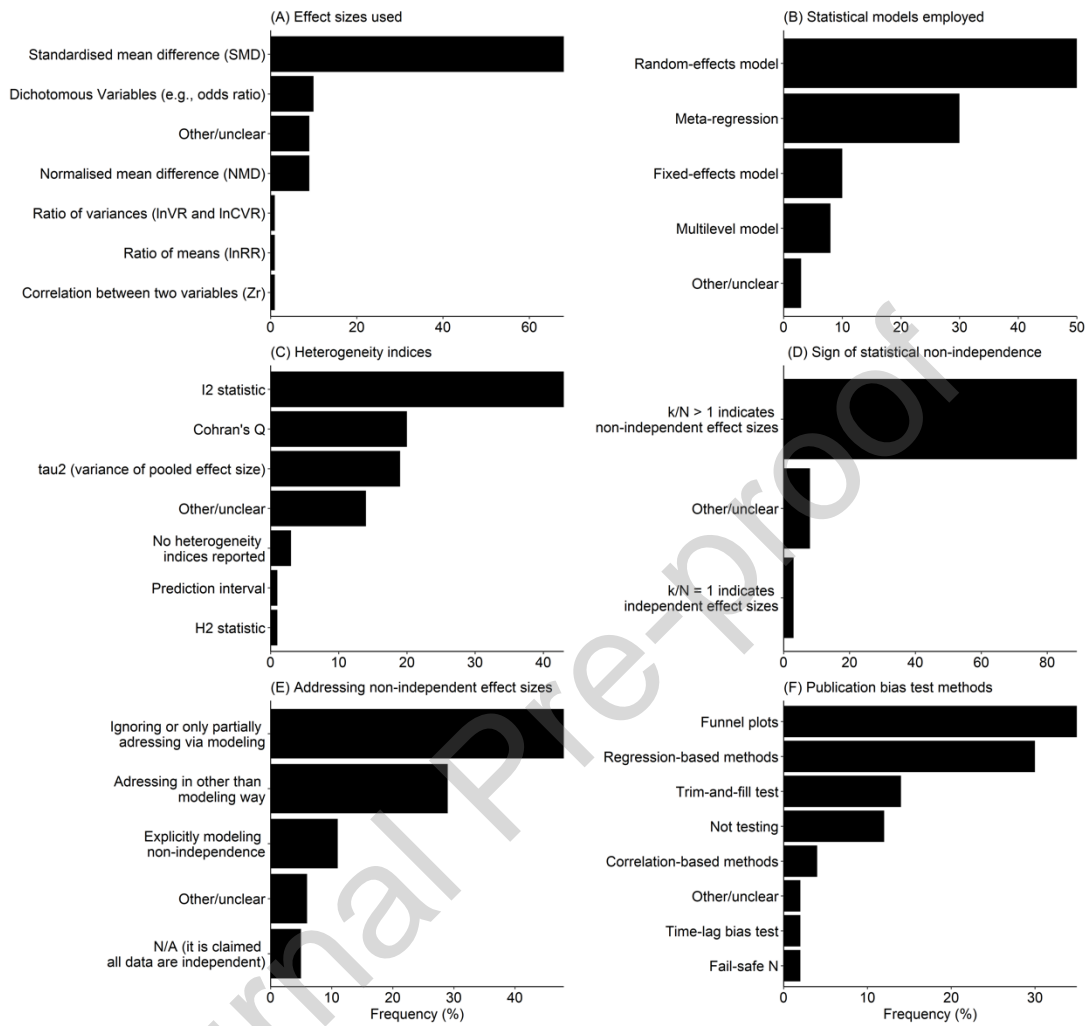


Figure 4

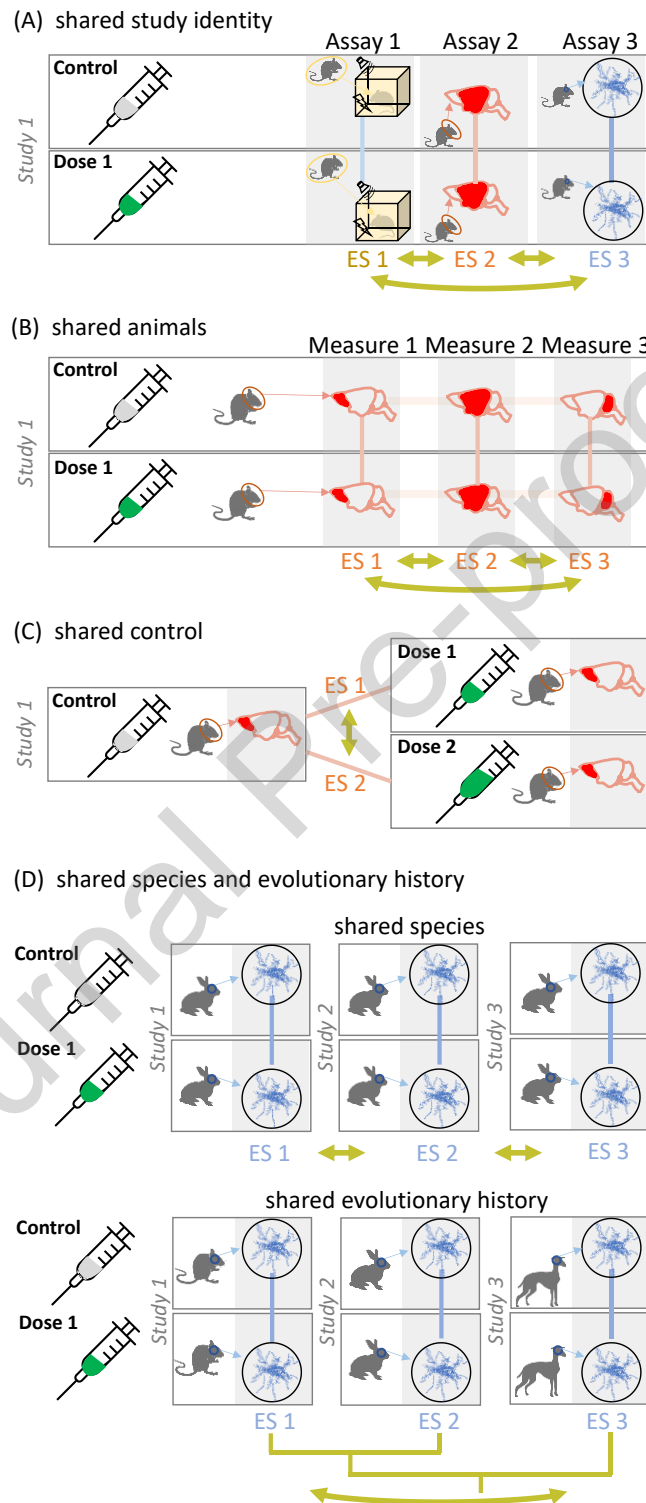
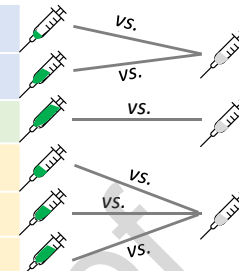


Figure 5

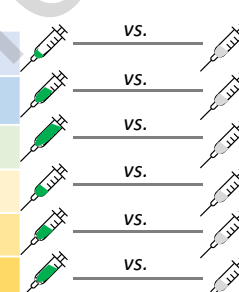
(A) original dataset with non-independent effect sizes

Study ID	Effect size	Sampling variance	Within-study moderator: dosages
Study 1	$ES_{11}$	$v_{11}$	0.1
Study 1	$ES_{12}$	$v_{12}$	0.3
Study 2	$ES_{21}$	$v_{21}$	0.5
Study 3	$ES_{31}$	$v_{31}$	0.2
Study 3	$ES_{32}$	$v_{32}$	0.3
Study 3	$ES_{33}$	$v_{33}$	0.4



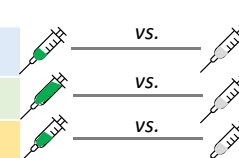
(B) ignoring statistical non-independence of effect sizes

Study ID	Effect size	Sampling variance	Within-study moderator: dosages
Study 1	$ES_1$	$v_1$	0.1
Study 2	$ES_2$	$v_2$	0.3
Study 3	$ES_3$	$v_3$	0.5
Study 4	$ES_4$	$v_4$	0.2
Study 5	$ES_5$	$v_5$	0.3
Study 6	$ES_6$	$v_6$	0.4



(C) averaging effect sizes to eliminate non-independence

Study ID	Effect size	Sampling variance	Within-study moderator: dosages
Study 1	$\overline{ES}_1$	$\overline{v}_1$	0.2
Study 2	$\overline{ES}_2$	$\overline{v}_2$	0.5
Study 3	$\overline{ES}_3$	$\overline{v}_3$	0.3



(D) sampling one from each study to eliminate non-independence

Study ID	Effect size	Sampling variance	Within-study moderator: dosages
Study 1	$ES_{11}$	$v_{11}$	0.1
Study 2	$ES_{21}$	$v_{21}$	0.5
Study 3	$ES_{31}$	$v_{31}$	0.3

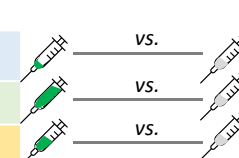
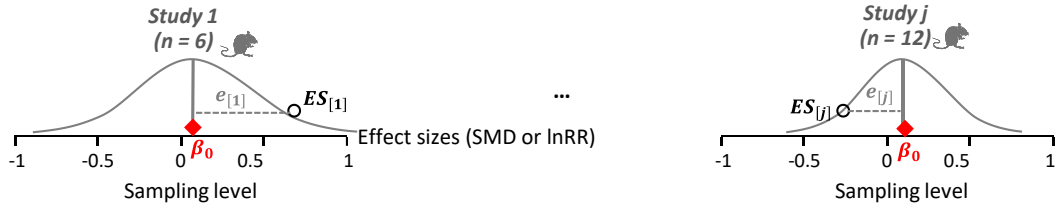
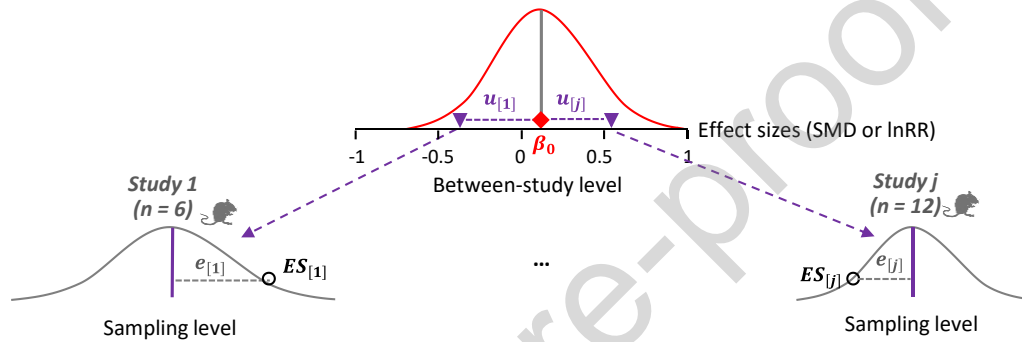


Figure 6

(A) Fixed-effects model:  $ES_{[j]} = \beta_0 + e_{[j]}$



(B) Random-effects model:  $ES_{[j]} = \beta_0 + u_{[j]} + e_{[j]}$



(C) Multilevel model:  $ES_{[i]} = \beta_0 + u_{b[j]} + u_{w[i]} + e_{[i]}$

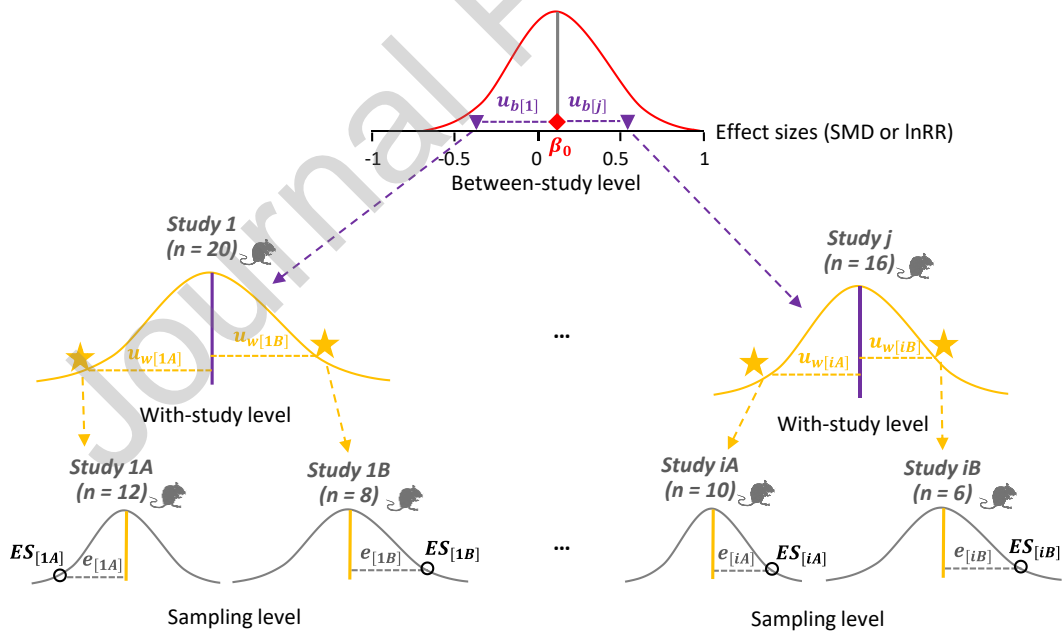


Figure 7

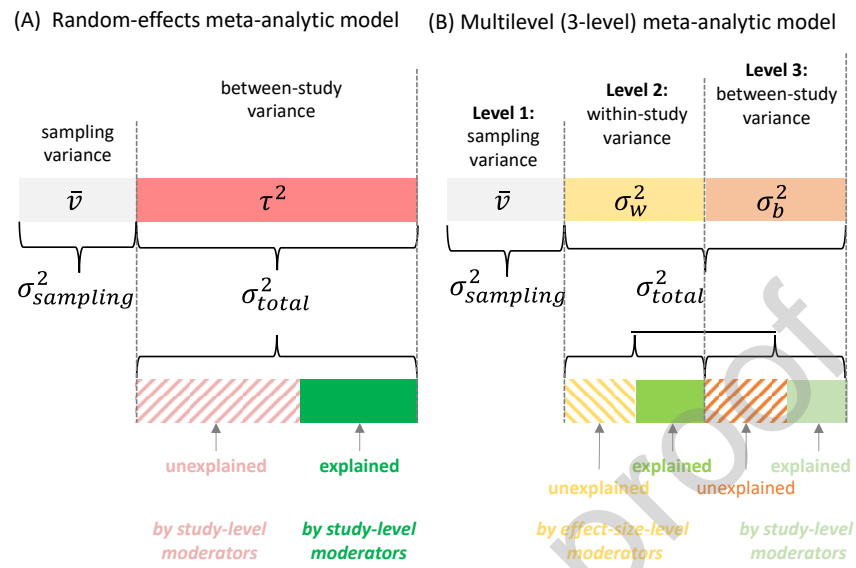
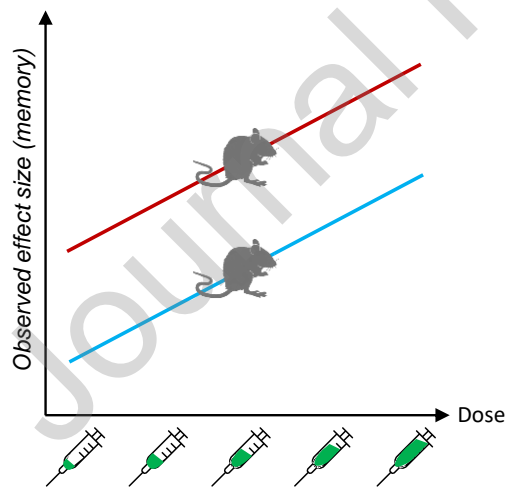
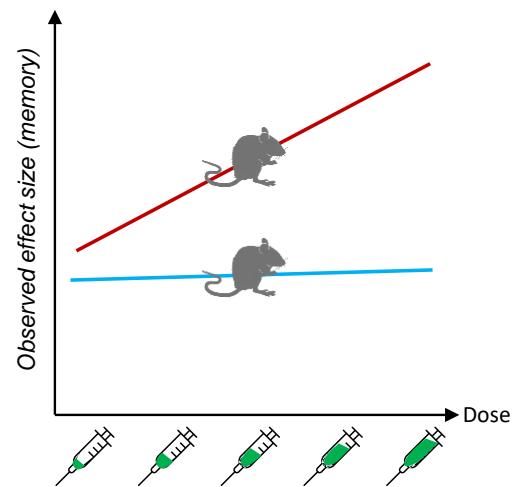


Figure 8

(A) No interaction



(B) Interaction





## Highlights

- Animal meta-analyses often involve non-independent and heterogeneous effect sizes
- Ignoring these issues leads to unreliable and less-informative evidence
- These issues have not been properly addressed in current animal meta-analyses
- Multilevel meta-analysis is introduced to solve the issues and is recommended
- A tutorial is provided to facilitate the application of advanced meta-analyses