Journal of Neuroscience Methods 221 (2014) 92–102

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Clinical Neuroscience
Invited review

Meta-analysis of data from animal studies: A practical guide

H.M. Vesterinen[a], E.S. Sena[a,b], K.J. Egan[a], T.C. Hirst[a], L. Churolov[b], G.L. Currie[a], A. Antonic[b], D.W. Howells[b], M.R. Macleod[a,*]

[a] Department of Clinical Neurosciences, The University of Edinburgh, United Kingdom
[b] The Florey Institute of Neuroscience and Mental Health, University of Melbourne, Australia

HIGHLIGHTS

- Meta-analysis is an invaluable tool in the life sciences.
- Methods for the application to clinical data are well documented.
- Consideration is required when applying these methods to preclinical data.
- We describe the application to preclinical data.
- We describe effect size calculations and assessing sources of heterogeneity.

ARTICLE INFO

Article history:
Accepted 16 September 2013

Keywords:
Meta-analysis
Animal studies
Heterogeneity
Meta-regression
Stratified meta-analysis

ABSTRACT

Meta-analyses of data from human studies are invaluable resources in the life sciences and the methods to conduct these are well documented. Similarly there are a number of benefits in conducting meta-analyses on data from animal studies; they can be used to inform clinical trial design, or to try and explain discrepancies between preclinical and clinical trial results. However there are inherit differences between animal and human studies and so applying the same techniques for the meta-analysis of preclinical data is not straightforward. For example preclinical studies are frequently small and there is often substantial heterogeneity between studies. This may have an impact on both the method of calculating an effect size and the method of pooling data. Here we describe a practical guide for the meta-analysis of data from animal studies including methods used to explore sources of heterogeneity.

Contents

* Corresponding author.
E-mail address: malcolm.macleod@ed.ac.uk (M.R. Macleod).

0165-0270
http://dx.doi.org/10.1016/j.jneumeth.2013.09.010

1. Introduction

Systematic review is a type of literature review that aims to identify all relevant studies to answer a particular research question (Greenhalgh, 1997; Cook et al., 1997). Data from these studies are often used in meta-analysis. The Cochrane collaboration has been pivotal in providing a framework for evidence-based health care to guide clinical decisions and healthcare policies. The use of systematic review and evidence-based healthcare is widely accepted by academia, healthcare professionals and funders, and these reviews receive twice as many citations in peer-reviewed journals as non-systematic reviews (Mickenautsch, 2010).

The systematic synthesis of data from the basic sciences is relatively novel. The Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES; www.camarades.info) was established in 2004 to promote and support the use of similar approaches to those used by the Cochrane Collaboration to data from animal studies. Other similar initiatives, such as the SYstematic Review Centre for Laboratory animal Experimentation (SYRCLE; www.umcn.nl/Research/Departments/cdl/SYRCLE) research group also actively promote and train individuals in the conduct of systematic reviews of preclinical studies. Whilst the Cochrane methodology is considered gold-standard, their remit is limited to health care interventions tested in humans, and their activity does not extend to in vitro or in vivo laboratory studies. Crucially, there are fundamental differences in the purposes, design and conduct of systematic review and meta-analysis of preclinical and clinical studies which mean that standard methodologies for systematic overviews and meta-analysis need to be adapted to this new setting.

The objectives of this paper are:

- to outline the rationale for the review and synthesis of preclinical data and to explain why the differences between clinical and preclinical reviews may require different approaches to the conduct of systematic review and meta-analysis;
- to present the methodology for a systematic review of preclinical data in a self-contained tutorial.

Although most of the statistical fundamentals used to review data from preclinical data are not novel, to our knowledge this is the first self-contained tutorial on applying these to the review of preclinical data. Unless otherwise stated, the formulae are adapted from those described by Borenstein (2009) which we recommend for further reading.

This paper is organised as follows: in Section 2 we describe why we perform systematic reviews of preclinical data and what makes them different to clinical systematic reviews; in Section 3 we describe the methodological approach to performing a review of the preclinical data; and in Section 4 we describe further considerations which may be helpful to the reader.

2. Why preclinical systematic reviews and what makes them different to clinical systematic reviews

Systematic reviews of data from preclinical literature are important for a number of reasons. First and foremost, although systematic reviews are not bias free, their purpose is to reduce it by outlining transparent aims, objectives and methodology. This approach enables us to identify all of the published literature to answer a particular research question. In turn this may highlight gaps in our knowledge which can be fulfilled by further preclinical experimentation, or it can help us to avoid unnecessary replication which is unethical and of limited benefit. Secondly, clinical trials of novel interventions should not proceed without a rigorous appraisal of the preclinical data. Systematic reviews can tell us the efficacy of any given intervention as well as the limits to efficacy which may aid in clinical trial design. Additionally, we can assess both the internal and external validity of each included study and assess for publication bias which can help to predict outcome in the clinical setting.

There are fundamental differences in the purposes, design and conduct of systematic review and meta-analysis of preclinical and clinical studies. Clinical reviews are intrinsically confirmatory (see The Cochrane Handbook by Higgins and Green, 2009): the aim of a Cochrane review is to provide evidence to allow practitioners and patients to make informed-decisions about the delivery of healthcare. Because certain aspects of experimental design can introduce bias to the results of relevant studies, a central part of a Cochrane review is to include only those studies meeting a certain threshold of internal validity to allow confidence in the results reported. In contrast, preclinical reviews are typically exploratory. Because the summary estimate of the effectiveness of an intervention in animal models is, of itself, not particularly useful information; the practice has been to include all available data. This is useful for identifying if there are any gaps in the data. One important purpose (and perhaps the single most important impact) of systematic reviews of preclinical studies has been to explore the impact of possible sources of bias, and we recommend that this is carried out in all systematic reviews. The important findings from such reviews are differences between different types of experiments (i.e. sources of heterogeneity) rather than a headline figure for how "good" a drug

is. Thus these analyses have a greater focus on exploring potential sources of heterogeneity. Additionally, reviews of preclinical data are hypothesis generating and can be used to inform the design and conduct of future trials.

Additionally, animal studies are generally small (with a sample size of around 10 per group), and slightly different studies of an individual intervention are often performed across many laboratories. In contrast, clinically trials are generaly larger, with single studies performed across multiple centres. In animal studies there is great emphasis on minimising variance, for instance through the use of inbred strains, pathogen free environments and specific handling conditions. This is not a focus for clincial trial design (and might indeed be considered to limit the generalisability of their findings). Differences between individual animal studies (using different strains, different conditions) are therefore, proportionately, larger. This has important implications for the conduct, analysis and interpretation of meta-analysis of data from preclinical studies.

Finally, conventional meta-analysis assumes effect sizes and their errors are independent when investigating sources of heterogeneity. Correlated error estimates can occur because preclinical studies often report complex experiments where control or treatment groups may be shared (i.e. in multi-armed studies) or use multiple comparisons from one study (such as multiple follow ups or measures of outcome). Correlated effect sizes estimates can occur between, for example, studies from the same laboratory or investigator (Hedges et al., 2010). These correlations between effect sizes, errors, or both, result in dependencies that may confound analyses. However, there may be other sources of correlation between different animal studies, for instance relating to animal husbandry, group housing, source of animals or particular experimental design characteristics shared between different studies; because this is essentially observational research we cannot exclude these factors unless they are reported, and as such this is a limitation to our approach.

A range of responses to the issue of dependency is possible in the meta-analysis preclinical studies (Hedges et al., 2010). This includes: ignoring the correlation arising due to all or some of the described reasons, creating a single synthetic effect size per sample, modelling dependence with full multivariate analysis, or using recently developed robust methods that estimate empirical standard errors. In our work we typically chose to explicitly address the issue of correlation due to shared control group by appropriately adjusting relevant sample sizes (detailed further in Section 3.3), while largely ignoring other sources of correlation. However, as the software implementations of new robust methodologies become handily available (discussed further in Section 4.1), they should be seriously considered when conducting meta-analysis of preclinical studies (van den Noortgate et al., 2013).

3. Methodological approach

3.1. Research protocol

As with any scientific research the first step should be to produce a detailed protocol describing what will be done, and why. In many cases the summary estimate of efficacy is of minor interest, and it is the heterogeneity between studies, and the differences which account for this heterogeneity, which are much more important. The summary estimate of efficacy should always be presented with, and interpreted in the light of, an analysis of heterogeneity. The protocol should define the aim and objectives, the hypothesis, and the steps that will be taken to meet the objectives. It should include (i) the search strategy used to identify the relevant literature (for details see Leenaars et al., 2012) (ii) criteria for inclusion or exclusion of literature identified by using the search strategy; (iii) the data that will be extracted, (iv) the primary outcome measure of interest. The protocol should define the methodological approach for (v) the calculation of individual effect sizes for each comparison, (vi) the calculation of summary effect sizes, and (vii) whether study design characteristics are going to be assessed as potential sources of heterogeneity, and if so, which characteristics, and by which method; and (viii) the method of assessing the internal validity (that is measures to avoid bias).

Like Cochrane, we encourage investigators to make protocols publicly available to the research community. This provides evidence that analyses are pre-specified, allows others to comment on the approach, provides examples to others planning such reviews and allows potential investigators to identify whether similar reviews are in progress. CAMARADES hosts a repository of protocols at: http://www.camarades.info/index_files/Protocols.html.

3.2. Data extraction

The results of the systematic search are usually downloaded to some form of reference management software. Two investigators independently screen title, abstract and, where necessary, full text, judging the work against the inclusion and exclusion criteria. Disagreements are resolved by discussion or by a third investigator. Disposal of literature thus identified (i.e. exclusions, with reasons given) can helpfully be presented in a flow chart akin to the PRISMA flow chart used in systematic reviews and meta-analysis of health care interventions (Liberati et al., 2009).

Included literature then forms the analysis set. Data should be extracted systematically and consistently from all relevant publications. The two types of information to be extracted are (i) the pre-defined study design characteristics; and (ii) outcome data (including the outcome measure used, the number of animals in which this was assessed, the aggregate value of effect (i.e. mean, median or event data) and where applicable a measure of group variance).

3.3. Meta-analysis

Meta-analysis proceeds through:

(1) calculating an effect size for each comparison;
(2) weighting the effect sizes;
(3) calculating efficacy where more than one relevant outcome is reported in the same cohort of animals;
(4) calculating a summary effect size and
(5) calculating the heterogeneity, and the extent to which the pre-defined study design characteristics explain this heterogeneity.

In the following sections we describe the calculation of effect sizes in the situation where these represent the magnitude of treatment effects; in Section 4.6 we describe how these methods can be applied to other types of animal experiment.

Irrespectively of the nature of the effect size, the first essential step in conducting meta-analysis of preclinical studies is correct estimation of the number of animals used in individual experiments. Since a single experiment can contain a number of comparisons, a control group can serve more than one treatment group. Were this control cohort to be included in more than one comparison, it would be represented more than once in the summary estimates calculated. To avoid this, we recommend to correct the number of animals reported in the control group by dividing the reported number by the number of treatment groups served in order to give a “true number of control animals”. This corrected number can then be used when calculating the total number of animals in the meta-analysis and where the number of animals is used in the weighting of effect sizes (Eqs. (1) and (2) (Table 1)).

This approach to dealing with outcome dependence within individual studies could be overly conservative and whenever there is a possibility to use newly developed robust methods for handling dependencies (see Section 4.1), these should be considered.

3.3.1. Calculating an effect size

For each comparison – where outcome in a cohort of animals receiving treatment is presented, along with that for in a control cohort – we calculate an effect size. A number of methods are available, each with their merits (for example see: Nakagawa and Cuthill, 2007; Baguley, 2009; Durlak, 2009). Here we describe approaches for data measured on a continuous scale (absolute difference in means, Section 3.3.1.1.i; normalised mean differences, Section 3.3.1.1.ii; and standardised mean differences, Section 3.3.1.1.iii); odds ratios (Section 3.3.1.2); and time to event data (e.g. median survival times; Section 3.3.1.3).

3.3.1.1. Calculating effect sizes for continuous data (mean outcome and its variance). Where we have a mean outcome score and a measure of its variance we can calculate an absolute difference in means, a normalised difference in means, or a standardised difference in means. For experiments which report standard error of the mean (*SEM*), these are converted to standard deviations (*SD*; Eq. (3)).

i. Absolute difference in means. Absolute differences in means (MDi) are the simplest measure of effect size and are the difference between the means in the control and treatment groups expressed in the units in which the outcome is measured (Eqs. (4)–(6)). A serious limitation to this approach is that the outcome measure and its scale must be the same across all studies. For instance, a 10 cubic millimetre reduction in mouse brain infarct volume is a much larger effect than the same reduction in infarct volume in a primate. However, where the outcome measures used are very similar, this approach may be used, and we have done so in analyses of self-administration of opioids (Du Sert et al., 2012) (where outcome was assessed as the number of administrations per hour).

ii. Normalised mean difference (NMD). Where data exist on a ratio scale (that is, where the score that would be achieved by a normal, untreated, unlesioned "sham" animal is known or can be inferred), we can express the absolute difference in means as a proportion. This value tells us the direction of the effect (i.e. what direction on the scale is "better" or "worse"), along with the magnitude of the treatment effect. This is a useful approach because it relates the magnitude of effect in the treatment group to a normal, healthy animal. The most common method to calculate NMD effect sizes is as a proportion of the mean in the control group. Typically, effect sizes fall between −100% and +100%.

The effect size is calculated using Eq. (7) with the standard deviations for each group also expressed as a percentage of the control group, normalised to the value in the sham group (Eq. (8)) with standard error calculated as shown in Eq. (9).

Because animal studies are usually small, and subject to random error, there are times when the observed lesion effect (the difference between sham and control, which serves as the denominator for normalisation) is very small. This can lead to extreme positive or negative calculated effect sizes. To account for this we have developed a second method for calculating a normalised effect size which we use where the absolute value of the effect size, as usually calculated, is more than 100% for any of the comparisons being considered. Under these circumstances we calculate the absolute difference between outcomes for each of the control and treatment groups and outcome in sham animals (Eq. (10)); and we express the effect size as the difference between these two values expressed as a proportion of the larger of the two; thus if $\left|\bar{x}_c - \bar{x}_{sham}\right| > \left|\bar{x}_{rx} - \bar{x}_{sham}\right|$, we use the formula shown in Eq. (11a);

Table 1
Equations used in the meta-analysis of data from preclinical studies.

Equation

$$n'_c = \frac{n_c}{\text{Treatment groups served by one control}} \quad (1)$$

Where n_c refers to the number of animals in the control group and n'_c refers to the true number of control animals.

$$N = n_{rx} + n'_c \quad (2)$$

Where n_{rx} refers to the number of animals in the treatment group and n'_c is calculated as shown in Eq. (1).

$$SD_c = SEM_c \times \sqrt{n_c} \quad \text{and} \quad SD_{rx} = SEM_{rx} \times \sqrt{n_{rx}} \quad (3)$$

Where n_c and n_{rx} refer to the number of animals in the control and treatment group respectively.

$$ES_i = \bar{x}_c - \bar{x}_{rx} \quad (4)$$

Where $\bar{x}_c$ and $\bar{x}_{rx}$ are the mean reported scores in the control and treatment group respectively and i denotes an individual study estimate.

$$SE_i = \sqrt{\frac{N}{n_{rx} \times n'_c} S^2_{pooled}} \quad (5)$$

Where S_{pooled} is calculated as shown in Eq. (6).

$$S_{pooled} = \sqrt{\frac{(n'_c - 1)SD_c^2 + (n_{rx} - 1)SD_{rx}^2}{N - 2}} \quad (6)$$

Where SD_c^2 and SD_{rx}^2 are the reported standard deviations for the control and treatment group respectively, using Eq. (3) to convert from standard errors if necessary.

$$ES_i = 100\% \times \frac{(\bar{x}_c - \bar{x}_{sham}) - (\bar{x}_{rx} - \bar{x}_{sham})}{(\bar{x}_c - \bar{x}_{sham})} \quad (7)$$

Where $\bar{x}_{sham}$ is the mean score for a normal, unlesioned and untreated animal (see Section 3.3.1.1.ii. for details).

$$SD_{c*} = 100 \times \frac{SD_c}{\bar{x}_c - \bar{x}_{sham}} \quad \text{and} \quad SD_{rx*} = 100 \times \frac{SD_{rx}}{\bar{x}_{rx} - \bar{x}_{sham}} \quad (8)$$

Where SD_c and SD_{rx} are the reported standard deviations for the control and treatment group respectively, using Eq. (3) to convert from standard errors if necessary.

$$SE_i = \sqrt{\frac{SD_{c*}^2}{n'_c} + \frac{SD_{rx*}^2}{n_{rx}}} \quad (9)$$

Where SD_{c*}^2 and SD_{rx*}^2 are calculated by squaring the functions calculated as shown in Eq. (8).

$$|\bar{x}_c - \bar{x}_{sham}| \quad \text{and} \quad |\bar{x}_{rx} - \bar{x}_{sham}| \quad (10)$$

$$ES_i = 100\% \times \frac{(\bar{x}_c - \bar{x}_{sham}) - (\bar{x}_{rx} - \bar{x}_{sham})}{(\bar{x}_c - \bar{x}_{sham})} \times direction \quad (11a)$$

The *direction* factor is as described in Table 2.

$$ES_i = 100\% \times \frac{(\bar{x}_{rx} - \bar{x}_{sham}) - (\bar{x}_c - \bar{x}_{sham})}{(\bar{x}_{rx} - \bar{x}_{sham})} \times direction \quad (11b)$$

$$SD_{C*} = 100\% \times \frac{SD_c}{\bar{x}_c - \bar{x}_{sham}} \quad \text{and} \quad SD_{rx*} = 100\% \times \frac{SD_{rx}}{\bar{x}_c - \bar{x}_{sham}} \quad (12)$$

$$SD_{C*} = 100\% \times \frac{SD_c}{\bar{x}_{rx} - \bar{x}_{sham}} \quad \text{and} \quad SD_{rx*} = 100\% \times \frac{SD_{rx}}{\bar{x}_{rx} - \bar{x}_{sham}} \quad (13)$$

$$SE_i = \sqrt{\frac{SD_{c*}^2}{n'_c} + \frac{SD_{rx*}^2}{n_{rx}}} \quad (14)$$

$$ES_i = \frac{\bar{x}_c - \bar{x}_{rx}}{S_{pooled}} \times \left(1 - \frac{3}{4N - 9}\right) \times direction \quad (15)$$

Where S_{pooled} is the pooled standard deviation calculated as shown in Eq. (6). The *direction* factor is as described in Table 2.

$$SE_i \sqrt{\frac{N}{n_{rx} \times n_{c'}} + \frac{ES_i^2}{2(N - 3.94)}} \quad (16)$$

$$OR_i = \frac{a_i \times d_i}{b_i \times c_i} \quad (17)$$

See Table 3 for details.

$$SE(\ln(OR_i)) = \sqrt{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}} \quad (18)$$

Where *ln* is the logarithm to base e (natural logarithm).

$$ES_i = \log\left(\frac{Median_{rx}}{Median_c}\right) \quad (19)$$

Where $Median_{rx}$ and $Median_c$ are the median survival times for the treatment and control group respectively.

Table 1 (*Continued*)

$$W_i = \frac{1}{SE_i^2} \quad (20)$$

Where SE_i^2 is the squared standard error of the effect size calculated as shown in Eq. (5) for absolute differences in means; Eqs. (9) or (14) for normalised mean differences; Eq. (16) for standardised mean differences; and Eq. (18) for odds ratios.

$$W_i ES_i = ES_i \times \frac{1}{SE_i^2} \quad (21)$$

$$W_i = N \quad (22)$$

Where the calculation for N is as shown in Eq. (2).

$$W_i ES_i = ES_i \times N \quad (23)$$

$$ES_{\theta i} = \frac{\sum_{i=1}^{k} W_i ES_i}{\sum_{i=1}^{k} W_i} \quad (24)$$

Where W_i is the measure of weight (e.g. inverse variance; Eq. (20)), $W_i ES_i$ is the weighted effect size, and k denotes the total number of studies included in the meta-analysis.

$$SE_{\theta i} = \sqrt{\frac{N_{comparisons}}{\sum_{i=1}^{k} W_i}} \quad (25)$$

Where $N_{comparisons}$ is the number of observations from the same cohort of animals contributing to the nested estimate of effect size.

$$ES_{fixed} = \frac{\sum_{i=1}^{k} ES_{\theta i} \times W^*}{\sum_{i=1}^{k} W^*} \quad (26)$$

Where W^* is the weight calculated as shown in Eq. (27).

$$W^* = \frac{1}{SE_{\theta i}^2} \quad (27)$$

$$SE_{fixed} \frac{1}{\sqrt{\sum_{i=1}^{k} W^*}} \quad (28)$$

$$\tau^2 = \frac{Q - df}{C} \quad (29)$$

Where τ^2 is the *estimation* of between-study variance; Q is the sum of the squared differences in effect sizes between studies and the pooled effect size (as shown in Eq. (30)); df is the degrees of freedom (Eq. (31)); and C is a measure used to convert the heterogeneity value into an average rather than a sum of squared deviations, and to put the value back into its original units (Eq. (32))

$$Q = \sum_{i=1}^{k} W^* \times (ES_{\theta i} - ES_{fixed})^2 \quad (30)$$

Where W^* is calculated as shown in Eq. (27).

$$df = k - 1 \quad (31)$$

Where k is the number of comparisons.

$$C = \sum_{i=1}^{k} W^* - \frac{\sum_{i=1}^{k} W^{*2}}{\sum_{i=1}^{k} W^*} \quad (32)$$

$$ES^*_{rand} = ES_{\theta i} \times W^*_{+\tau^2} \quad (33)$$

Where $W^*_{+\tau^2}$ is calculated as shown in Eq. (34).

$$W^*_{+\tau^2} = \frac{1}{(SE^2_{\theta i} + \tau^2)} \quad (34)$$

Table 1 (*Continued*)

$$ES_{Random} = \frac{\sum_{i=1}^{k} ES^*_{rand}}{\sum_{i=1}^{k} W^*_{+\tau^2}} \quad (35)$$

$$SE_{Random} = \frac{1}{\sqrt{\sum_{i=1}^{k} W^*_{+\tau^2}}} \quad (36)$$

$$95\%\text{CI} = ES_{Random} \pm 1.95996 \times SE_{Random} \quad (37)$$

$$W^*_{+\tau^2} = \frac{1}{1/N + \tau^2} \quad (38)$$

$$\exp ES_{Random} \quad (39)$$

Where ES_{Random} is calculated as shown in Eq. (35).

$$SE_{Random} = \sqrt{\frac{\sum_{i=1}^{k} [W^*_{+\tau^2} (ES_{\theta i} - ES_{Random})^2]}{N^2 \times \sum_{i=1}^{k} W^*_{+\tau^2}}} \quad (40)$$

$$95\%\text{CI} = \exp(ES_{Random} \pm 1.95996 \times SE_{Random}) \quad (41)$$

$$p = CHIDIST(Q_{global} - sum(Q_{strata})), df) \quad (42)$$

Where Q_{global} is the amount of heterogeneity for the global estimate of effect size, Q_{strata} is the amount of heterogeneity within individual components of the strata, and df is the degrees of freedom (the number of components in the strata minus one).

$$I^2 = \frac{Q - df}{Q} \times 100\% \quad (43)$$

$$metareg\ y\ varlist,\ se \quad (44)$$

Where y is the dependant variable, and in this case the effect size; *varlist* are the study covariates that are being assessed; se is the standard error calculated in Eq. (25); the within study variance.

$$\text{Regression weight} = \frac{1}{SE_{\theta_i} + \tau^2} \quad (45)$$

Where $SE_{\theta i}$ is the standard error of the nested effect size for the ith study τ^2 is the residual heterogeneity (Thompson and Sharp, 1999).

$$Adjusted\ R^2 = 1 - \left(\frac{\tau^2_{With\ Covariates}}{\tau^2_{Without\ Covariates}} \right) \quad (46)$$

$$t = \frac{\beta}{SE_\beta} \quad (47)$$

$$AUC = n(\bar{x}) - 0.5(\bar{x}_{FTP} + \bar{x}_{LTP}) \quad (48)$$

Where $\bar{x}$ is the mean of all the individual data points for the treatment or control group; n is the number of data points contributing to the analysis; $\bar{x}_{FTP}$ is the first data point; and $\bar{x}_{LTP}$ is the last data point.

$$SD_{AUC} = \sqrt{\sum_{i}^{k} \left(\bar{x}_i - \bar{X}\right)^2 + \sum_{i}^{k} SD_i^2} \quad (49)$$

Where $\bar{x}_i$ is the mean value in the control or treatment group at the ith time point; SD_i^2 is the squared standard deviation of the mean at the ith time point.

$$\text{Calculated } SD = \frac{SE_{MDi}}{S_i \times \left(1 - \left(\frac{3}{4N-9}\right)\right)} \quad (50)$$

Where SE_{MDi} is the standard error of the difference in means, calculated as shown in Eq. (51); S_i is calculated as shown in Eq. (52); and N is the total number of animals, calculated as shown in Eq. (2).

$$SE_{MDi} = \sqrt{\frac{SD_c^2}{n'_c} + \frac{SD_{rx}^2}{n_{rx}}} \quad (51)$$

Where SD_c^2 and SD_{rx}^2 are the reported standard deviations for the control and treatment group respectively, using Eq. (3) to convert from standard errors if necessary.

$$S_i \frac{(n_c - 1)SD_c^2 + (n_{rx} - 1)SD_{rx}^2}{N} \quad (52)$$

Table 2
The correction factor used to define the direction of the effect size.

Better outcome in?	Higher mean score represents?	Multiply effect size by?
Control group	Better outcome	−1
Treatment group	Better outcome	1
Control group	Worse outcome	−1
Treatment group	Worse outcome	1

and if $\left|\bar{x}_{rx} - \bar{x}_{sham}\right| > \left|\bar{x}_{c} - \bar{x}_{sham}\right|$, we use the formula shown in Eq. (11b). Importantly, in the calculation of this NMD effect size, the value for sham does not provide the direction of the effect (i.e. where a higher score represents a better or worse outcome) and so the effect size needs to be adjusted according to the rules shown in Table 2.

We also normalise the standard deviations of the treatment and control group to the same denominator used in the effect size calculation. Thus if $\left|\bar{x}_{c} - \bar{x}_{sham}\right| > \left|\bar{x}_{rx} - \bar{x}_{sham}\right|$, we use the formulae shown in Eq. (12); or if $\left|\bar{x}_{rx} - \bar{x}_{sham}\right| > \left|\bar{x}_{c} - \bar{x}_{sham}\right|$, we use the formula shown in Eq. (13). Finally, the standard error for the effect size is shown in Eq. (14).

iii. Standardised mean difference. The NMD approach above is relevant to ratio scales, but sometimes it is not possible to infer what a "normal" animal would score – for instance in the number of neurons per high power field, or spontaneous motor activity – and sometimes data for unlesioned animals are not available. In these circumstances we can use standardised mean differences (SMD). The difference in group means is divided by a measure of the pooled variance to convert all outcome measures to a standardised scale with units of standard deviations (*SD*s). This approach can also be applied to data where different measurement scales are reported for the same outcome measure; for example different measures of lesion size such as infarct volume and infarct area.

There are three common methods used (Egger et al., 2001); Cohen's *D* (the difference in means is divided by the pooled standard deviation) Glass's Delta, (the difference in means is divided by the standard deviation of the control group only); and Hedge's *G* (which is based on Cohen's *D* but includes a correction factor for small sample size bias (Hedges and Olkin, 1985)).

It is suggested that "small" samples are those of less than 10 subjects per group, and because most animal experiments use fewer than this (Rooke et al., 2011) we have used Hedge's *G* effect sizes for SMD analyses. Hedges *G* introduces a correction factor between 0 and 1, and for larger sample sizes this tends towards 1 and therefore the effect size tends towards Cohen's *D* (Cooper et al., 2009).

The formulae used to calculate Hedge's *G* standardised effect size are shown in Eqs. (15) to (16). Again, the calculations need to take into account the direction of effect.

3.3.1.2. Calculating an effect size for event data (odds ratio). For binary outcomes such as the number of animals that developed a disease or died, data can be represented in a 2 × 2 table (Table 3) and the odds ratio and its standard error calculated as described (Egger et al., 2001). Note that where the value in any cell is zero, 0.5 is added to each cell to avoid problems with the computation of the standard error. For each comparison the odds ratio (OR) is calculated using Eq. (17) (Egger et al., 2001). Odds ratios are normally combined on a logarithmic scale therefore the standard error of the log OR measure is calculated as shown in Eq. (18) (Egger et al., 2001);

Table 3
Summary table for events data, where *i* denotes the individual comparison.

$Study_i$	Event	No event	Group size
Treatment group	a_i	b_i	n_{rx}
Control group	c_i	d_i	n_c

3.3.1.3. Calculating an effect size for median survival data/time to event data. Where data are presented as median survival (for example in animal models of glioma), we divide the median survival in the treatment group by the median in the control group and take the logarithm of this factor (Eq. (19)). This approach does not allow for a calculation for the variance of the effect size, and this problem is addressed in Section 4.4.

3.3.2. Weighting effect sizes

In meta-analysis it is usual to attribute different weights to each study in order to reflect relative contributions of individual studies to the total effect estimate. This is done according to the precision of that study, so that more precise studies are given greater weight in the calculation of the pooled effect size. In the first stage of meta-analysis we recommend to use the inverse variance method, where individual effect sizes are multiplied by the inverse of their squared standard error (*SE*). This gives a weighted effect size $W_i ES_i$, where ES_i is the individual effect size and W_i is the weight ($1/SE_i^2$) (Eqs. (20) and (21)). For median survival or other time to event data we weight effect sizes according to the total number of animals (the true number of control animals plus the number of treated animals) in that comparison (Eqs. (22) and (23)).

3.3.3. Combining effect sizes from similar outcome measures in the same cohort of animals

Where multiple similar outcomes are reported from the same cohort of animals we must choose either to extract a single outcome or to combine more than one outcome. Separate meta-analyses of each individual outcome measure are sometimes appropriate where there are enough data; however it is often preferable to take all available data, particularly when the data are limited, unless a primary outcome measure has been pre-specified. For instance, four different neurobehavioural tests might be reported from the same experimental groups. If we wanted to use a single outcome we might select the smallest effect size, or have a hierarchy of preferred outcome measures, or only include data for one specific outcome measure. Alternatively, we could combine outcomes as appropriate to give a single outcome statistic (the "nested" outcome), representing a global measure of the behavioural outcome in that comparison. To do this we take each outcome, weight it by multiplication by the inverse of the variance for that outcome, sum these weighted values for all outcomes and divide by the sum of the weights (Eq. (24)). The standard error of this effect size is given by the square root of the number of comparisons divided by the sum of the weights (Eq. (25)).

3.3.4. Pooling effect sizes

Effect sizes can be combined using fixed- or random-effects model (Borenstein, 2009). The fixed effects model is used when it can be assumed that the different studies each give an estimate of the same effect, which is assumed to be *fixed* across all comparisons. Thus, *observed* effect sizes vary due to random sampling error alone. The random effects model is used when it can be assumed that the underlying effect size differs between studies, perhaps due to different doses or routes of administration. Random effects meta-analysis therefore takes into account both the within-study (sampling error) and between-study (differences in the true effect size) variance. The distribution of effect sizes has a weighted mean (the summary estimate), a weighted sum of the square of the deviations from that mean (the heterogeneity), and an estimate of the variance of the effect sizes beyond that expected by chance (tau-squared, τ^2).

i. Calculate a fixed effects summary estimate. For each comparison, a weight is calculated from the inverse of the square of the standard error ("inverse variance"). Where pooled data from a single comparison are used, the standard error is calculated as described above (Eq. (25)). Each effect size is multiplied by its weight and the resulting products are summed, and then divided by the sum of the weights to give the summary estimate (Eqs. (26) and (27)). The 95% CI for the fixed effects estimate is the same as that shown in Eq. (37) for the random effects estimate, replacing ES_{Random} and SE_{Random} with ES_{Fixed} and SE_{Fixed} respectively. The standard error of the fixed effects estimate is the square root of the sum of the weights (Eq. (28)). Tau-squared (τ^2) is a measure of excess between-study variation, reflecting the difference between the observed treatment effects across different studies beyond that which would be expected if the assumptions of fixed effects modelling (that all studies measured the same underlying effect) held. It is used to refine the weighting used in the random effects model, which uses both within-the study variance (the variance of the individual studies) and the between-study variance (τ^2, constant across all studies being pooled; Eqs. (29)–(32)). If τ^2 is large compared to the within study variance, the random effects estimate will tend towards a simple average, and if τ^2 is zero the random effects estimate will be the same as the fixed effects estimate.

Because the true effect size for an intervention is unknown, τ^2 cannot be known, but it can be estimated using the method of moments (Dersimonian and Laird, 1986).

ii. Calculate a random effects estimate. We now calculate the random effects estimate as we did for fixed effects, except the studies are weighted by the inverse of the sum of within study variance and τ^2 rather than by within study variance alone (Eqs. (33)–(35)). From the standard error (Eq. (36)) we can calculate 95% confidence intervals (Eq. (37)).

3.3.4.1. Median survival data. Different approaches have been described for the meta-analysis of median survival or time to event data (Michiels et al., 2005). In animal studies we have the special circumstance that cohort size is often orders of magnitude smaller than the clinical studies for which these techniques were developed, limiting their validity. We calculate effect sizes for individual studies by dividing the median survival in the treatment group by the median survival in the control group and then taking the logarithm of the quotient (Eq. (19); Simes, 1987). The precision of survival studies is related to the number of animals included so we use this to weight studies, giving a fixed effects weight of N (rather than inverse variance) (Eq. (22)). τ^2 is calculated as previously (Eqs. (29)–(32)) and for the random effects analysis, studies are weighted using the formula shown in Eq. (38). The random effects estimate is calculated according to Eq. (35) to which we use the exponential function to convert the estimate to a linear scale, providing a figure which is representative of a median survival ratio (Eq. (39)). This provides a more intuitive summary, as one can use it to estimate, by simple multiplication, what the survival in the treatment group would be under different assumptions of control group survival. Finally the standard error and 95% confidence interval are calculated according to Eqs. (40) and (41).

3.3.5. Heterogeneity

It is sometimes interesting to know if there are important differences between groups of studies, or study characteristics (such as delays to treatment) which may influence outcome. The differences between studies can also give some indication of whether they are drawn from the same (i.e. measure the same thing) or different populations. To identify heterogeneity, visual inspection of individual effect sizes (e.g. funnel plotting) or overall effect size estimations and their 95% confidence intervals (CI) can give an informal indication of the presence of heterogeneity. However although 95% CIs which do not overlap indicate statistical significant at the $p < 0.05$ level, overlapping confidence intervals do not necessarily indicate a non-significant difference. To empirically assess heterogeneity we calculate heterogeneity using Cochran's Q (hereafter referred to as Q (Cochran, 1954); and I^2 (Higgins et al., 2003). There are two approaches to assessing differences between studies or the impact of study characteristics, stratified meta-analysis by partitioning of heterogeneity (Borenstein, 2009), and meta-regression (Thompson and Higgins, 2002).

3.3.5.1. Estimating the amount of heterogeneity. Q is an estimate of the between study heterogeneity which is independent of the units in which the effect size is expressed. Q is calculated from the effect sizes in the fixed effect model. If the studies are drawn from the same population of studies which measure the same thing, then any variation is due to sampling error and the expected value of Q is simply the degrees of freedom. Under this assumption the values of Q follow a chi-squared distribution with [k (comparisons) minus one] degrees of freedom. Therefore the significance of differences between Q and the expected variation can be tested using the chi-squared statistic (Eq. (42)). Importantly, a non-significant value for Q does not necessarily indicate that the studies are drawn from the same population, as low power within studies (small sample size for the comparisons) and between studies (a small number of comparisons contributing to the meta-analysis) may yield a falsely neutral result.

While Q is very useful it is not easily understood and is sensitive to the number of comparisons. To address this issue Higgins and Thompson (2002) defined I^2 as the proportion of total variance between studies that is due to true differences in effect sizes as opposed to chance (Eq. (43)). I^2 lies between 0% (all variation being due to chance alone) and 100% (all variation reflects real differences between the true effect sizes between studies) and does not depend on the number of comparisons in the meta-analysis. Guidance for interpreting the I^2 value is provided by Higgins et al. (2003); 0–25% is considered to reflect very low heterogeneity; 25–50% reflects low heterogeneity; 50–75% reflects moderate heterogeneity; and >75% reflects high heterogeneity. The decision to use a fixed effects or random effects model based on these statistics is subjective; however, we would consider using a random effects model on I^2 values greater than 50%.

3.3.5.2. Exploring sources of heterogeneity. Here we describe two methods to explore sources of heterogeneity; stratified meta-analysis and meta-regression.

i. Stratified analysis. The principle underlying stratified meta-analysis is that, if certain study characteristics are important, effect sizes from studies which share those characteristics will be more similar to each other than they will to studies which do not share those characteristics. The heterogeneity is partitioned into that within groups of similar studies and that between groups of studies. For each group of studies (or stratification) we calculate a random-effects effect size and heterogeneity Q. The heterogeneity statistics for each grouping are added together and subtracted from the total heterogeneity to give the residual heterogeneity between groups (Eq. (42); Excel function, version 2003–2007). The extent of heterogeneity between these groups (that is, are they significantly different?) can then be tested as before using the chi squared distribution.

ii. Meta-regression. Meta-regression extends the random effects meta-analysis model by taking into account one or more study-level characteristics (covariates) and determines how much heterogeneity can be explained by taking into account both within- and between-study variance. Meta-regression can be conducted using *Stata*/*SE* with the linear function, *metareg* (Eq. (44); Thompson and Higgins, 2002).

Meta-regression is a weighted linear regression and describes a line of best fit between covariates and effect size. Unless it can be assumed that the covariate in question explains all between study heterogeneity, a random effects meta-regression is used (Egger et al., 2001), which weights on both within-study variance and between-study variance (Eq. (45)).

The measure of between study variance is again termed τ^2, and there are a number of ways of calculating this. The moment estimator calculation of τ^2 is that used in DerSimonian and Laird random effects meta-analysis but is less suitable when covariates are included (Thompson and Sharp, 1999). Other methods are iterative, and the choice of method directly influences both coefficient estimates and standard errors (Thompson and Sharp, 1999). We recommend using the restricted maximum likelihood estimate (REML) approach to estimate τ^2 because it is less likely to underestimate or produce biased estimates of variance (Thompson and Sharp, 1999). Both univariate (to assess the impact of a single covariable on effect size) or multivariate analyses (to assess the impact of multiple variables) are possible. Where covariates are categorical rather than continuous, dummy variables are required. This converts categorical variables with n potential values into $n-1$ dichotomous variables (value 0 or 1), with the final value for the category serving as a reference value with value 0 across all dichotomous variables.

3.3.5.3. Output of model. In conventional linear regression, the adjusted R^2 measures the variance in the dependant variable which is accounted for by different values of the independent variable. In meta-regression, the estimated between study variance τ^2 is a measure of the residual heterogeneity. Therefore the change in τ^2, following inclusion of covariables represents the change in residual heterogeneity, and the variance in the dependant variable which is accounted for by covariates is used to calculate an adjusted R^2 (Eq. (46)), a measure of how much heterogeneity is explained by the model.

The *F*-ratio is a measure of how much the addition of covariates has improved the prediction of outcome with larger *F-ratios* indicating better prediction. The *F*-ratio is expressed with the *df* of both the number of covariates and the number cases given in subscript, and significance is tested against the *F* distribution, commonly used in analysis of variance.

For each covariate a coefficient (β) is calculated, which represents the change in y with each unit change in the covariate, along with a standard error for β; its 95% confidence interval; and a *t*-statistic testing the null hypothesis that the value of β is zero (Eq. (47)).

Predictive multivariate regression models can be built using any of the standard backward elimination, forward selection, or stepwise approaches. Such models can then be validated using training and validation sets, or other approaches such as leave-one-out validation or *k*-fold validation (Efron, 1983).

4. Further considerations

Here we provide further considerations which might be helpful.

4.1. Software

Although other software packages (e.g. R statistical software) may be suitable, we use the following: (i) the CAMARADES Microsoft Access (2003 version) data-manager and Microsoft Excel (any version) for stratified meta-analysis; (ii) *Stata*/*SE* using the linear function, metareg, for conventional meta-regression in which effect sizes and errors are assumed to be independent. When this is not the case, i.e. when effect sizes, errors, or both, are expected to be correlated (see Section 2 for details), a more recently developed "robumeta" function in Stata (Hedges et al., 2010) can be used.

We have developed the CAMARADES data-manager, access to which is available upon request (www.camarades.info), and which can be used to record data and perform analyses. Other free software program such as RevMan have been specifically developed for the meta-analysis; however the reader should be aware that these were developed for the collation of data from clinical trials. Comprehensive Meta-Analysis is proprietary software developed for data entry and analysis in meta-analysis.

4.2. Multiple testing – correcting p values and confidence intervals?

Meta-analyses of in vivo animal data will often involve large numbers of contrasts being specified in the study protocol, and the statistical analysis plan should account for this. We routinely group contrasts according to the broad hypothesis being tested (e.g. that study quality has an impact) or to the category of outcome measure (structural or functional), and within these groups of contrasts partition a Type 1 error rate of 5% among the contrasts tested using Bonferroni correction.

4.3. Missing data

Meta-analyses are based on data available in the public domain, typically in peer-reviewed journals, or on unpublished data which has been sought out. The reporting of data is not always adequate (Sena et al., 2007), and it is our experience that the number of animals per group or the variance or both are not always reported. In these situations we make attempts to contact authors for the information, or (if many studies are missing the variable of interest) use a method to calculate and pool effect sizes that does not require these data, or (if only a small number of studies are missing the variable of interest) we exclude the data. We report the prevalence of inadequate reporting in study publications in a flow chart of the disposal of publications identified in the review. Additionally if data for sham animals are missing we cannot calculate normalised mean difference effect sizes. In these circumstances, if greater than 10% of the data for sham animals are missing we would use an alternative approach such as calculating standardised mean difference effect sizes.

4.4. Data on a continuous scale where variance is not reported

Sometimes studies report mean outcomes without reporting variance. If there are substantial other data which do report variance, we can simply use these and exclude the others. However on occasion as many as 80% of publications within a review do not report variance. In these circumstances it may be possible to calculate a summary estimate using absolute difference in means or normalised difference in means; however, because the weighting given to individual studies is usually based at least in part on inverse variance we must in these circumstances either not weight (i.e. use a simple average) or weight according to some other factor such as the number of animals in each comparison, with the variance of the summary estimate as the square root of the sum of the squares of the deviations from the pooled mean.

4.5. Difficulties with certain data values

In some circumstances the calculation of effect size or standard error, or both, cannot proceed – if the group sizes are toosmall to

Table 4
A description of circumstances in which an effect size cannot be calculated.

Cause	Consequence	Affects which method?
SD_c and SD_{rx} are both zero	The calculation of the effect size includes a division by zero for SMD; the *SE* for the individual effect size will equal zero and so the weighting cannot be calculated.	SMD; NMD
Small group sizes	This can introduce a requirement to take the square root of a negative number.	SMD
Numbers cancel each other out	A calculated denominator equals zero, introducing a divide by zero instruction.	SMD; NMD
$\bar{x}_c$ is equal to $\bar{x}_{sham} - \bar{x}_c$	Where lesioning has no effect it is not possible to calculate the relative effect of an intervention	NMD

allow Hedges *G* to be calculated, or the variance is zero – and these comparisons are excluded from further analysis. Some of these circumstances are described in Table 4.

4.6. Other types of data presentation

The term *effect size* is often understood as a *treatment effect*, the impact of a treatment intended to improve outcome. However meta-analyses are not restricted to data from such studies and are useful tools in understanding disease models as a whole. For example we have conducted a meta-analysis on behavioural and macroscopic data from studies of animal models of bone-cancer pain. For this we use the value for a normal animal as our control, and the value in the animals with bone cancer as the "treatment group". For this, consideration simply needs to be taken in ensuring the direction of effect is the same for all comparisons. However, it is not always possible to determine the direction of effect size; for example, some biochemical markers are reported, but it is not always clear – or known – whether an increase is a beneficial or negative effect. In this situation we reported these separately, stating simply whether the value was higher or lower in the animals with bone cancer.

4.7. Choosing between multiple control groups

In some situations the choice of the most relevant control group is not clear. For instance, in studies involving stem cells, data may be presented for stem cells; for another cell type not thought to have certain characteristics; for dead cells; for conditioned culture medium; for unconditioned culture medium; for saline; or for no treatment. The preferred choice, and if necessary a hierarchy of preferred choices, should be addressed in the protocol.

4.8. Median survival data

The median survival time is the time of the first event at which the Kaplan–Meier estimator is below 0.5. This is calculated by drawing a horizontal line at 50% on the *y*-axis and estimating the intercept with the curve. If the curve is horizontal at $y = 50\%$, the average of the first and last time point of the horizontal line can be considered the best estimate of the median. One problem with this approach is that if more than half the animals in a group (usually the treatment group) survive to the end of the experiment a median survival time cannot be calculated. If we exclude these data our summary estimate will be overly conservative so in these circumstances we consider median survival as the last time point of assessment and noted that more than 50% of animals survived at this time. This will still underestimate efficacy, but not to the same extent as if the data were excluded completely. There are alternative methods to calculate a pooled median survival estimate, including the mean survival time; however, survival times tend to be highly skewed and so the median is generally a better measure of the central tendency.

4.9. Co-treatments

Sometimes publications report the effect of drugs in combination – for instance control (C), A, B and AB. In a review of the efficacy of A it is reasonable to extract data for A v C and AB v B. However, in a review of the efficacy of all treatments the comparisons would be A v C, B v C, and AB v C. Unfortunately, if in a review of the efficacy of A we are only provided with data for AB v C then these should not be included in the analysis, as any effect may be due to B rather than A.

4.10. Using ordinal scale data as continuous

These approaches require the assumption that data lie on an interval scale (that is, differences between different points on the scale are of the same magnitude); and that they are normally distributed. These assumptions do not always hold, particularly for functional outcomes. However, when datasets are large (as they usually are in such reviews) parametric manipulations do have some validity (Lord, 1953). This is however a potential limitation of the methodology and can usefully be discussed in study reports.

4.11. Including multiple time points

Where differences in the change of outcome over time are of interest (for instance the acquisition of learning in the Morris water maze) we can include these data by calculating the area under the performance–time curve (AUC) for different cohorts. Using the data extracted regarding mean and variance point estimates, all time points are used to calculated one overall comparison (Eq. (48)) with standard deviation (Eq. (49))

4.12. Assessing the relationship between outcome measures

Where more than one outcome measure is reported for the same cohort of animals we can assess the extent to which these outcomes measure the same or different effect of treatment using meta-regression, using the same approach described above.

4.13. Publication bias

Funnel plotting, Egger regression and "trim and fill" can each be applied to data from systematic reviews of in vivo data. Where different outcomes have been measured in the same cohort of animals (see Section 3.3.3) we recommend using each of these outcomes rather than the pooled estimate, since to do otherwise would in effect be suppressing these studies from the publication bias analysis. For funnel plotting and Egger regression of SMD effect sizes where studies are small certain symmetries arise because the standardised effect size is constrained to a certain set of values by its sample size, and this becomes apparent with small sample sizes, as is the case for in vivo studies. We therefore recommend using a measure of pooled standard deviation in the formula for precision (1/variance) in Egger regression, shown in Eqs. (50)–(52).

4.14. Alternative effect size calculations, and choice of measure

Alternatives to normalised and standardised mean difference analyses include the ratio of means. The performance of this approach has been compared to mean difference and standardised mean difference approaches but not to the normalised mean difference approach (Friedrich et al., 2008). It is reported to perform less well where variance is more than 70% of the effect size, or when standardised effect sizes are large, as is often the case in reviews of in vivo data.

However, performance of each of these approaches has not to our knowledge been compared either in simulation or in reanalysis of existing datasets. The optimal approach for different circumstances is therefore not known. On the basis that SMD is more conservative than NMD analysis, and meta-regression is probably more conservative than the partitioning of heterogeneity, we tend to use NMD with meta-regression and SMD with partitioning heterogeneity, with the alternative approach used to provide sensitivity analysis.

5. Discussion

Here we have outlined the main steps to meta-analysis of data from animal studies. It should be noted that there a number of alternative methods, for instance as described by Borenstein (2009). However in our experience the methods we have described here are practical and appropriate for a wide range of circumstances and in particular where there are large numbers of small studies with substantial heterogeneity in study design and outcomes reported, as common in the preclinical sciences. In this section we discuss some of the limitations to the approaches described here, and outline some of the questions which remain to be answered.

The choice of whether to use SMD or NMD analysis is not always clear. Because group size is often small, the measured variance is an imprecise estimate of the population variance, and therefore the calculation of a standardised effect size introduces a measurement error. However, the outcome for sham (unlesioned) animals may be neither reported nor obvious, and in those circumstances NMD analysis is not possible. The investigator may therefore be faced with the choice of an SMD analysis involving an entire dataset, or an NMD analysis involving a proportion of the dataset. This will depend on a judgement about whether the benefits of NMD analysis outweigh the loss of data; where possible it is preferable to establish the criteria for this judgement in advance, and whatever the decision, to use the alternative approach as a secondary analysis.

In addition the choice of whether to use stratified meta-analysis or meta-regression to assess the significance of associations between study design characteristics with effect sizes is not always clear. In preliminary work applying both approaches to the same large dataset we have found that meta-regression is substantially more conservative, and further analysis should provide better guidance of the most appropriate method in different circumstances.

Meta-analysis is an evolving methodology, and one recent advance has been in the handling of dependencies between effect sizes, variance, or both. This is an important consideration and we are in the process of merging this into our approach to meta-analysis of pre-clinical data. Importantly, the nature of preclinical experimentation means that the issue of dependencies may be more pronounced than in the clinical literature; we have observed that control groups can serve more than twenty treatment groups; one laboratory can produce more than ten research articles on a particular topic; and there can be over five behavioural endpoints reported for a single cohort of animals. To account for this we now recommend using the robust variance estimate which is described in more detail by Hedges et al. (2010).

A limitation to meta-analysis in general is the risk of spurious findings due to statistical artefact rather than true associations between study design characteristics with effect sizes. Although this is an important consideration, the use of a correction factor (e.g. Bonferroni) will reduce the likelihood of this.

6. Conclusions

Animal studies are crucial to our understanding of disease mechanisms and for testing interventions for safety and efficacy. Animal studies are inherently heterogeneous, and more so than a typical clinical trial. Successfully translating findings to humans diseases depends largely upon an understanding these sources of heterogeneity, and their impact on effect size. Meta-analysis is a useful tool for this purpose when the data are systematically identified. Here we have summarised the main methods which can be used to meta-analyse data from animal studies. All of the methods described have been used previously across a range of preclinical data, some of which are referred to here. Further information and guidance on conducting systematic reviews and meta-analyses of data from preclinical studies is available from the CAMARADES collaboration (www.camarades.info) or SYRCLE (http://www.umcn.nl/research/departments/cdl/syrcle/Pages/default.aspx); for general background reading on systematic review and meta-analysis (more focused on the clinical perspective) we recommend textbooks by Higgins and Green (2009) and Borenstein (2009).

Conflict of interest

The authors declare that there is no conflict of interest.

References

Baguley T. Standardized or simple effect size: what should be reported? British Journal of Psychology (London, England: 1953) 2009;100:603–17.

Borenstein M. Introduction to meta-analysis. Chichester, U.K.: John Wiley & Sons; 2009.

Cochran WG. The combination of estimates from different experiments. Biometrics 1954;10:101–29.

Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Annals of Internal Medicine 1997;126:376–80.

Cooper HM, Hedges LV, Valentine JC. The handbook of research synthesis and meta-analysis. New York: Russell Sage Foundation; 2009.

Dersimonian R, Laird N. Metaanalysis in clinical-trials. Controlled Clinical Trials 1986;7:177–88.

Du Sert NP, Chapman K, Sena E. Systematic review and meta-analysis of the self-administration of opioids in rats and non-human primates to provide evidence for the choice of species in models of abuse potential. In: Safety Pharmacology Society: 12th Annual Meeting; 2012.

Durlak JA. How to select, calculate, and interpret effect sizes. Journal of Pediatric Psychology 2009;34:917–28.

Efron B. Estimate the error rate of a prediction rule: some improvements on cross-validation. Journal of the American Statistical Association 1983;78:316–31.

Egger M, Smith GD, Altman DG. Systematic reviews in health care: meta-analysis in context. London, UK: BMJ Publishing Group; 2001.

Friedrich JO, Adhikari NKJ, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. BMC Medical Research Methodology 2008;8., http://dx.doi.org/10.1186/1471-2288-8-32.

Greenhalgh T. How to read a paper: papers that summarise other papers (systematic reviews and meta-analyses). British Medical Journal 1997;315:672–5.

Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando: Academic Press; 1985.

Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods 2010;1:39–65.

Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2. [updated September 2009]. The Cochrane Collaboration; 2009, Available from www.cochrane-handbook.org

Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 2002;21:1539–58.

Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. British Medical Journal 2003;327:557–60.

Leenaars M, Hooijmans CR, van Veggel N, Ter Riet G, Leeflang M, Hooft L, van der Wilt GJ, Tillema A, Ritskes-Hoitinga M. A step-by-step guide to systematically identify all relevant animal studies. Laboratory Animals 2012;46:24–31.

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Medicine 2009;6:e1000100.

Lord FM. On the statistical treatment of football numbers. American Psychologist 1953;8:750–1.

Michiels S, Piedbois P, Burdett S, Syz N, Stewart L, Pignon JP. Meta-analysis when only the median survival times are known: a comparison with individual patient data results. International Journal of Technology Assessment in Health Care 2005;21:119–25.

Mickenautsch S. Systematic reviews, systematic error and the acquisition of clinical knowledge. BMC Medical Research Methodology 2010;10:53.

Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews 2007;82:591–605.

Rooke EDM, Vesterinen HM, Sena ES, Egan KJ, Macleod MR. Dopamine agonists in animal models of Parkinson's disease: a systematic review and meta-analysis. Parkinsonism & Related Disorders 2011;17:313–20.

Sena E, van der Worp HB, Howells D, Macleod M. How can we improve the pre-clinical development of drugs for stroke? Trends in Neurosciences 2007;30:433–9.

Simes RJ. Confronting publication bias – a cohort design for metaanalysis. Statistics in Medicine 1987;6:11–29.

Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 2002;21:1559–73.

Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in Medicine 1999;18:2693–708.

van den Noortgate W, Lopez-Lopez JA, Marin-Martinez F, Sanchez-Meca J. Three-level meta-analysis of dependent effect sizes. Behavior Research Methods 2013;45:576–94.