


## RESEARCH ARTICLE

# Assessment of key characteristics, methodology, and effect size measures used in meta-analysis of human-health-related animal studies

Carlijn R. Hooijmans<sup>1,2</sup>  | Rogier Donders<sup>3</sup> | Kristen Magnuson<sup>4</sup> |  
Kimberley E. Wever<sup>1,2</sup> | Mehmet Ergün<sup>2</sup> | Andrew A. Rooney<sup>5</sup> |  
Vickie Walker<sup>5</sup> | Miranda W. Langendam<sup>6,7</sup>

<sup>1</sup>Department of Anesthesiology, pain and palliative care, Radboud university medical center, Nijmegen, The Netherlands

<sup>2</sup>Systematic Review Centre for Laboratory animal Experimentation (SYRCLE), Department for Health Evidence, Radboud Institute for Health Sciences, Radboud university medical center, Nijmegen, The Netherlands

<sup>3</sup>Biostatistics, Department for Health Evidence, Radboud Institute for Health Sciences, Radboud university medical center, Nijmegen, The Netherlands

<sup>4</sup>ICF, Fairfax, Virginia, USA

<sup>5</sup>Division of the National Toxicology Program, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, Durham, North Carolina, USA

<sup>6</sup>Department of Epidemiology and Data Science, Amsterdam UMC location Academic Medical Centre, Amsterdam, The Netherlands

<sup>7</sup>Department of Methodology, Amsterdam Public Health, Amsterdam, The Netherlands

## Correspondence

Carlijn R. Hooijmans, Department of Anesthesiology Pain and Palliative Medicine; Radboud University Medical Center, Geert Grooteplein-Noord 21, Route 126; 6525 GA, Nijmegen, The Netherlands.  
Email: [carlijn.hooijmans@radboudumc.nl](mailto:carlijn.hooijmans@radboudumc.nl)

## Abstract

Since the early 1990s the number of systematic reviews (SR) of animal studies has steadily increased. There is, however, little guidance on when and how to conduct a meta-analysis of human-health-related animal studies. To gain insight about the methods that are currently used we created an overview of the key characteristics of published meta-analyses of animal studies, with a focus on the choice of effect size measures. An additional goal was to learn about the rationale behind the meta-analysis methods used by the review authors. We show that important details of the meta-analyses are not fully described, only a fraction of all human-health-related meta-analyses provided rationales for their decision to use specific effect size measures. In addition, our data may suggest that authors make post-hoc decisions to switch to another effect size measure during the course of their meta-analysis, and possibly search for significant effects. Based on analyses in this paper we recommend that review teams: 1) publish a review protocol before starting the conduct of a SR, prespecifying all methodological details (providing special attention to the planned meta-analysis including the effect size measure and the rationale behind choosing a specific effect size, prespecifying subgroups and restricting the number of subgroup analyses), 2) always use the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) checklist to report your SR of animal studies, and 3) use the random effects model (REM) in human-health-related meta-analysis of animal studies, unless the assumptions for using the fixed effect model (FEM) are all met.

## KEYWORDS

effect size measures, meta-analysis, meta-research, systematic review of animal studies

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

**Funding information**

National Institutes of Health, Grant/  
Award Numbers: GS00Q14OADU417,  
HHSN273201600015U

**Highlights**

- The design and conduct of animal studies differs from clinical studies in a number of aspects, as a consequence meta-analysis methodology and guidelines for the interpretation of the results need to be tailored specifically to animal study data
- There is, little guidance available on when and how to conduct a meta-analysis of human-health-related animal studies
- To gain insight about the methods that are currently used we present an overview of the key characteristics of published meta-analyses of animal studies, with a focus on the choice of effect size measures.
- We show that important details of the meta-analyses are not fully described, and suggest that authors make post-hoc decisions to switch to another effect size measure during the course of their meta-analysis, and possibly search for significant effects.

## 1 | INTRODUCTION

Since Cochran published the first formalized meta-analytic methods in 1954,<sup>1</sup> and Smith and Glass published the first meta-analysis in social sciences in 1977,<sup>2</sup> this method of synthesizing results across studies in order to obtain robust and informative evidence summaries has become widely used. This is no surprise as meta-analyses often provide clear overviews of available evidence as well as guidance for future research on topics with inconclusive data.

One consequence of the widespread use of meta-analysis to summarize evidence is that it has changed how scientists view the results of individual primary studies. A new study is now often seen as a contribution to a body of evidence instead of a conclusive answer to a scientific question.<sup>3</sup> In addition, the process of conducting meta-analyses often revealed various shortcomings of the included original papers that impacted the ability to conduct analyses or sensitivity analyses. Scientists are now, for example, more aware of the need to improve reporting standards of individual studies, the risk of bias caused by poor methodological quality and the value of publishing null or negative results.

Despite these clear advantages of conducting meta-analyses, it took almost two more decades after the publication of Smith and Glass before the first meta-analysis of animal studies was conducted (1994),<sup>4</sup> and two more decades for meta-analyses of animal studies to become relatively common.

One reason for this slow adoption may be that animal researchers are generally unfamiliar with systematic review and meta-analysis methodology, even though there is some tailored guidance available on when and how to conduct meta-analysis of animal studies.<sup>5–8</sup>

The basic principles of the methodology used for meta-analysis of animal studies are largely similar to those used for clinical meta-analysis. However, animal studies differ from clinical studies in a number of aspects: 1) animal studies show far greater diversity in the species studied, the experimental design used, and other study characteristics, 2) animal studies have goals beyond establishing efficacy of an intervention, 3) the goal of meta-analysis of animal studies is generally not to pinpoint the effect estimate (as in clinical meta-analysis), but rather to assess the direction and magnitude of the effect and explore sources of heterogeneity; and 4) it is common for meta-analyses of animal studies to contain dozens or even hundreds of studies, with small sample sizes per study, rather than a small number of comparatively large studies, as is common in clinical meta-analyses. Because of these aspects, both the meta-analysis methodology and guidelines for the interpretation of the results need to be tailored specifically to animal study data.

Meta-analyses of human-health-related animal studies often have one of two purposes: 1) to facilitate healthcare decisions and medical research; and/or 2) to support hazard or risk decisions in toxicology. They may, for example, aid the selection of interventions with therapeutic potential to be tested in clinical trials, inform regulatory decisions limiting human exposure to drugs or toxicants, or guide decisions on the utility of further animal studies.

Many methodological decisions need to be made when conducting a meta-analysis. For example, which effect size measure will be used for each outcome, how the heterogeneity between the studies will be identified and analyzed, which effect model will be used (e.g., fixed effect model (FEM) or random effects model (REM)), whether subgroup analyses or meta-regression

is planned and whether the subgroups are predefined. Importantly, these decisions should be made and documented a priori, before conducting any analyses, in order to prevent bias in the analysis and avoid data dredging.

One important aspect of meta-analysis methodology is the choice of effect size measure. In the majority of animal studies, many outcomes are reported on a continuous scale (e.g., weight gain in grams). For such continuous outcome measures, three effect size measures can be used: the mean difference (MD, the absolute difference between the group means), the standardized mean difference (SMD, the difference between the group means relative to the standard deviation), or the normalized mean difference (NMD, the difference between the group means expressed as a proportion of the mean in the control group).

The MD can be used when all studies report their outcome data in the same unit of measurement, and the reported unit of measurement has the same meaning in all studies. This is not always the case: the units of measurement often differ between studies (e.g., some express weight gain in grams, others in % increase) and sometimes the interpretation of the same unit of measurement differs between populations (e.g., a weight gain of 6 g has a different meaning in studies using mice vs. studies using dogs). The SMD was developed to overcome this problem, and is obtained by dividing the MD by study's pooled standard deviation, to create an effect estimate that is comparable across studies.

The NMD is an effect size measure that is (to our knowledge) only used in meta-analyses of animal studies. It was first described by CAMARADES (Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies) in 2008.<sup>9</sup> The NMD can be used when the measurements for normal, untreated, nonlesioned (sham) animals are known or can be inferred. A sham animal or sham treatment is often a faked (surgical) intervention that omits the step that is thought to be therapeutically necessary. Including the values of a sham treatment in the effect size measure is a useful approach as it relates the magnitude of effect in the treatment group to a normal, healthy animal. Another advantage of the NMD is that the MD is expressed as a proportion of the mean in the control group, which may be easier to interpret.<sup>10</sup> Without sham animals, the interpretation of the NMD changes, and the NMD becomes almost similar to the ratio of means, which is scarcely used in meta-analyses of clinical trials. For more information about the various continuous effect size measures, and the utility and disadvantages of each see Supporting Information, Table S1.

From previous publications and reviews, it is not yet clear whether the rationale used by review authors for selecting the MD, SMD, or the NMD in general is valid, or if there are advantages of the use of one of these effect size measures over the others in a case where multiple effect sizes seem appropriate. In order to obtain more insight into the methods used in meta-analysis of human-health-related animal studies (such as the decision to use a specific effect size measure), we review key characteristics of published meta-analyses of human-health-related animal studies, focussing on important aspects of the methods used and the effect size measures reported (such as the MD, SMD, and NMD). In addition, we aim to obtain insight into the rationale behind the choice in methodology. Insight into the methodology used, will ultimately aid in optimizing the guidance on how to conduct meta-analysis of animal studies, and estimate the reliability of the results presented in the published meta-analysis.

For this overview, we used the recently created database<sup>11</sup> containing all systematic reviews of animal studies until February 2018 as a resource for this assessment.

## 2 | METHODS

The database of systematic reviews of animal studies contains all systematic reviews of animal studies published in PubMed, Embase, and Web of Science from inception through February 2018. The database contained 2391 systematic reviews covering a broad range of fields including preclinical research and veterinary medicine, toxicology and environmental health research (to see the database used in this manuscript visit: <https://data.mendeley.com/datasets/6fr3nw5mpc/1>, and use version AnimalSR\_V1.0\_12\_27\_2018.accdb).

The methods used to create the database are described in detail elsewhere.<sup>11</sup> Briefly, for studies to be included in the database, the aim of the publication needed to be to systematically review the literature and this had to be stated in the title or abstract. The publication needed to summarize the results of studies in laboratory animals or veterinary patients, report the eligibility criteria for the primary studies, specify search terms and database/electronic sources searched and be available as a full text version.

All included studies were categorized and labeled. One of these labels each reference received was whether or not a meta-analysis was conducted. A study received this label when data from included studies were extracted and combined in a quantitative assessment. This included cluster analysis, pooled risk ratio, pooled odds

ratio, or vote counting (i.e., integrating evidence by counting significant positive, significant negative, and non-significant results).

## 2.1 | Study selection

For this study, we included all references from the above mentioned database that were initially labeled as containing meta-analyses, unless they met one or more of the following exclusion criteria: 1) the study did not present a quantitative summary of the evidence or a summary statistic; 2) the study did not have the aim to investigate or improve human health (i.e., the target populations were not humans such as veterinary studies); 3) the study summarized the evidence of a single group (i.e., no control group present); 4) the study did not investigate the effect of an intervention, defined here as a difference between a control and experimental group controlled by the scientist (i.e., a drug treatment, exposure [toxicological studies], or a specific group characteristic such as sex); 5) the study was not peer reviewed. Studies using vote counting methods were labeled as such.

## 2.2 | Data extraction

A single person extracted the following data elements for each study: study ID from the SR database, health topic (based on the health topics [generally disease-based] created by the Cochrane library), and the number of outcomes assessed (each outcome was considered a separate meta-analysis). In addition, we extracted the following data elements for each meta-analysis: name of outcome, type of intervention, type of control group, number of species, species, number of comparisons included, number of animals in overall analyses, similar units of measurement (y/n), effect size measure used, unit of measurement of effect size, rationale for effect size measure described (y/n), description of rationale (if reported), effect model used in meta-analysis, type of heterogeneity measure(s) used, heterogeneity in overall meta-analysis, pooled summary effect size and confidence interval, subgroup analyses conducted (y/n), number of subgroup analyses, type of subgroup analyses (e.g., stratified or meta-regression), threshold for number of comparisons needed to conduct subgroup analyses, minimal number of comparisons indicated in results for subgroup analyses, test for statistical difference between subgroups (y/n).

The number of comparisons indicates the number of independent experiments included in a meta-

analysis (not publications; as one publication may contain more than one independent experiment). All the extracted data are presented in Supporting Information, File S2.

## 2.3 | Analyses

In order to provide an overview of the various types of meta-analyses and the methods used to conduct meta-analyses of human-health-related laboratory animal studies, we analyzed:

1. the total number of meta-analyses conducted per review
2. the average number of meta-analyses (e.g., outcomes) per review
3. the most common health topics for human-health-related meta-analyses
4. the average number of species, and number of comparisons used per meta-analysis
5. effect size measures used (both continuous and dichotomous)
6. the percentage of meta-analyses showing significant effects
7. heterogeneity measures used and the average level of between-study heterogeneity based on  $I^2$  (either directly presented in the original paper or recalculated based on  $Q$  and degrees of freedom [DF])
8. the proportion of meta-analyses conducting subgroup analyses and the average number of subgroup analyses per outcome
9. minimal number of comparisons per subgroup defined by the author
10. the method used to conduct subgroup analyses (meta-regression versus stratified analyses)

We summarized the rationales (as reported by the authors) for the choice of the effect size measures to obtain insight into the underlying reasons for the methodological decisions to choose to use a specific continuous effect size measure. As we expect the reasoning for the methodological decision to use a specific effect size measure is poorly reported in the included systematic reviews, we also analyzed the existence of a possible relation between the following element by plotting the data: 1) the use of a specific effect size measure and between-study heterogeneity (defined as  $I^2$ ); 2) the use of a specific effect size measure and the proportion of statistically significant findings; and 3) the use of a specific effect size measures, between-study heterogeneity and the proportion of statistically

significant findings. Since we included all human-health-related animal meta-analyses, (e.g., we did not take a small sample but included almost the entire “population” of human-health-related animal meta-analyses), and we do not aim to extrapolate our results to another population or subgroup of studies, significance tests are inappropriate.

We hypothesize that the heterogeneity levels of the meta-analyses using MD are higher compared to the heterogeneity levels of standardized or normalized effect size measures as 1) we expect that the MD is regularly used erroneously in meta-analyses of animal studies when studying multiple species, and 2) the aim of standardizing and normalizing is to make the effects more homogeneous. Subsequently because high heterogeneity levels generally cause wide confidence intervals around the summary effect and decrease the chance of finding significant effects, we also hypothesize that meta analyses using the MD have a lower probability of finding significant effects compared to the SMD and NMD.

Regarding the relation between statistically significant findings, heterogeneity and various effect size measures, we aimed to investigate whether scientists make

post-hoc decisions regarding their decision to use a specific effect size measure.

We also considered the effect of the number of species used in the analyses and the number of comparisons in the meta-analyses.

### 3 | RESULTS

All the extracted data used in this manuscript is presented in Supporting Information, File S2.

#### 3.1 | Overview of the various types of meta-analyses and the methods used to conduct meta-analyses of human-health-related laboratory animal studies

Out of the 2391 systematic reviews included in our database, 840 contained a meta-analysis (35.1%). Fifty-three percent of those reviews containing a meta-analysis ( $n = 443$ ) were included in further analyses, as they aimed to improve human health and studied an

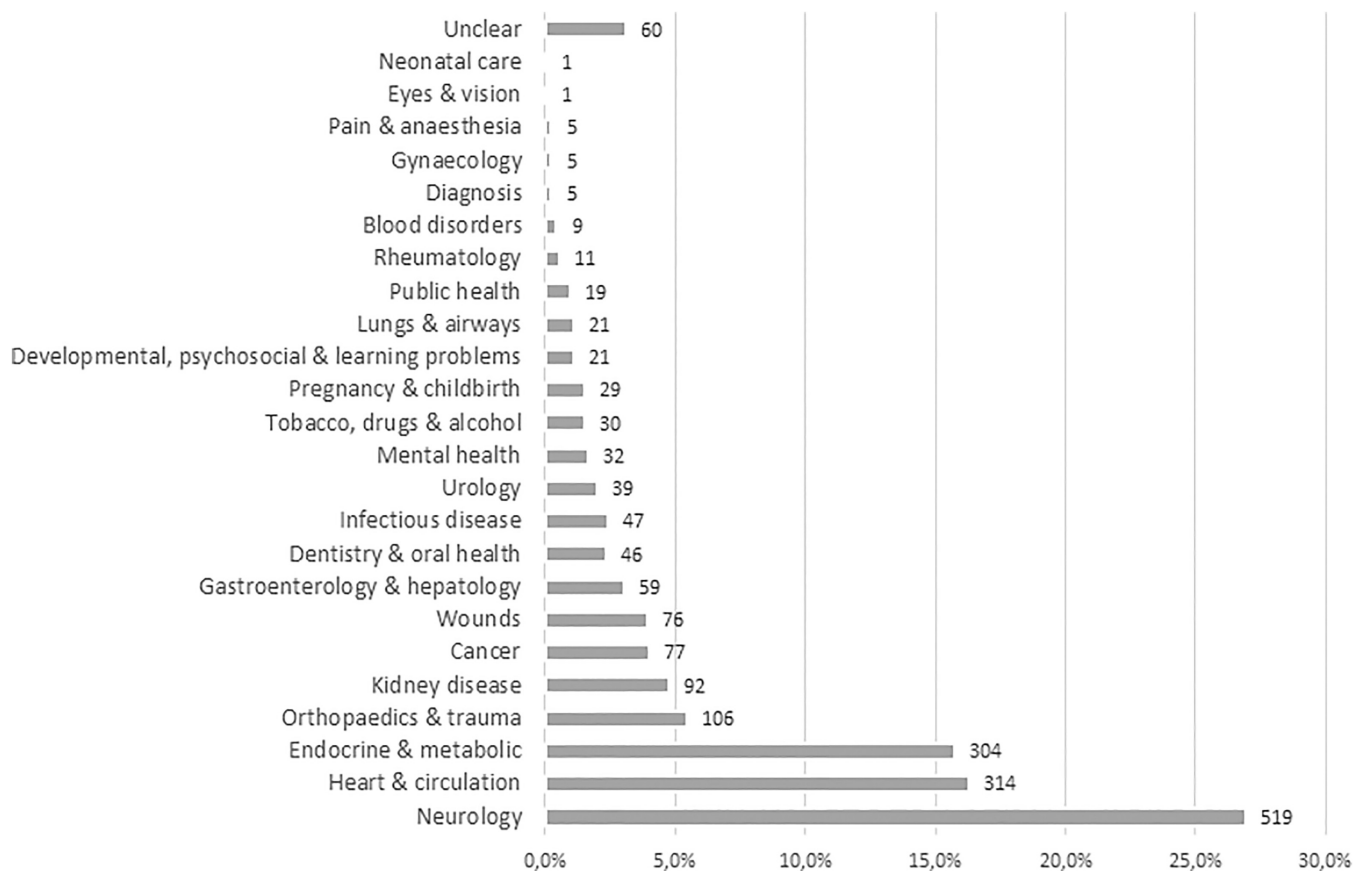


FIGURE 1 Health topics. Number and proportion of meta-analyses per health topic. The health topics are based on the health topics created by the Cochrane library.



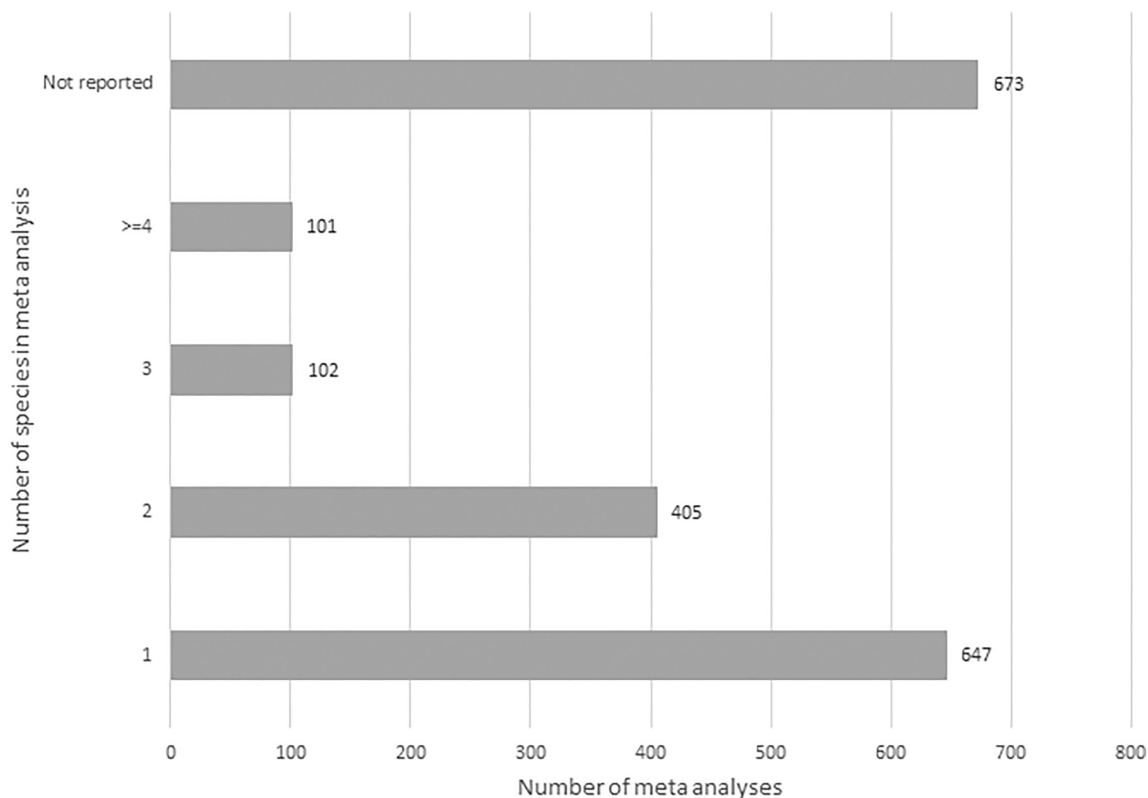


FIGURE 2 Number of meta-analyses relative to the number of species included in meta-analysis

intervention. A total of 1928 meta-analyses were conducted in these 443 references, indicating a mean of 4.4 meta-analyses or outcomes per review.

Sixty percent (60%) of the excluded reviews containing a meta-analysis ( $n = 230$ ) concerned veterinary medicine and were not of direct relevance to human health. Toxicological studies (focused on exposures instead of interventions) were included, as well as all other meta-analyses organized by health topics.

Figure 1 shows that the three most common health topics for human-health-related meta-analyses are: 1) neurology (27%); 2) heart and circulation (16%); and 3) endocrine and metabolic (16%).

Thirty-three percent (33%) of the meta-analyses included only one species, 32% used two or more species, and 35% did not explicitly describe the number of species included (Figure 2). By extracting the type of species used from subgroup analyses and the method 65 different species appear to be used, and the most frequently used species were rats (72% of the meta-analyses) and mice (51% of the meta-analyses).

### 3.1.1 | Number of comparisons

The size of the meta-analyses varied considerably (Figure 3). The mean number of comparisons included in a meta-analysis was 28. Three percent of the meta-

analyses contained only 1 comparison, and in 14% of the meta-analyses, the number of comparisons was unclear.

### 3.1.2 | Meta-analyses models

The majority of studies used a random effects model (REM) (79%). This appears also to be the case analyzed over time (Supporting Information, File S3). A fixed effects model (FEM) was used in 12% of the studies. Five percent of the meta-analyses did not report the type of model they used. The remaining meta-analyses mainly used a vote counting method or a model in which the individual studies were summarized but not weighted (Figure 4).

Out of the 1519 meta-analyses that used a REM, 30% included only one species, 35% did not describe the number of species included in their analyses, and 35% included two or more species.

Out of the 230 meta-analyses that used a FEM, the majority of meta-analyses (60%) included only one species in their analyses. Twenty two percent included two or more species.

### 3.1.3 | Effect size measures

Figure 5 shows the effect size measures used in the meta-analyses. Dichotomous outcomes were used in

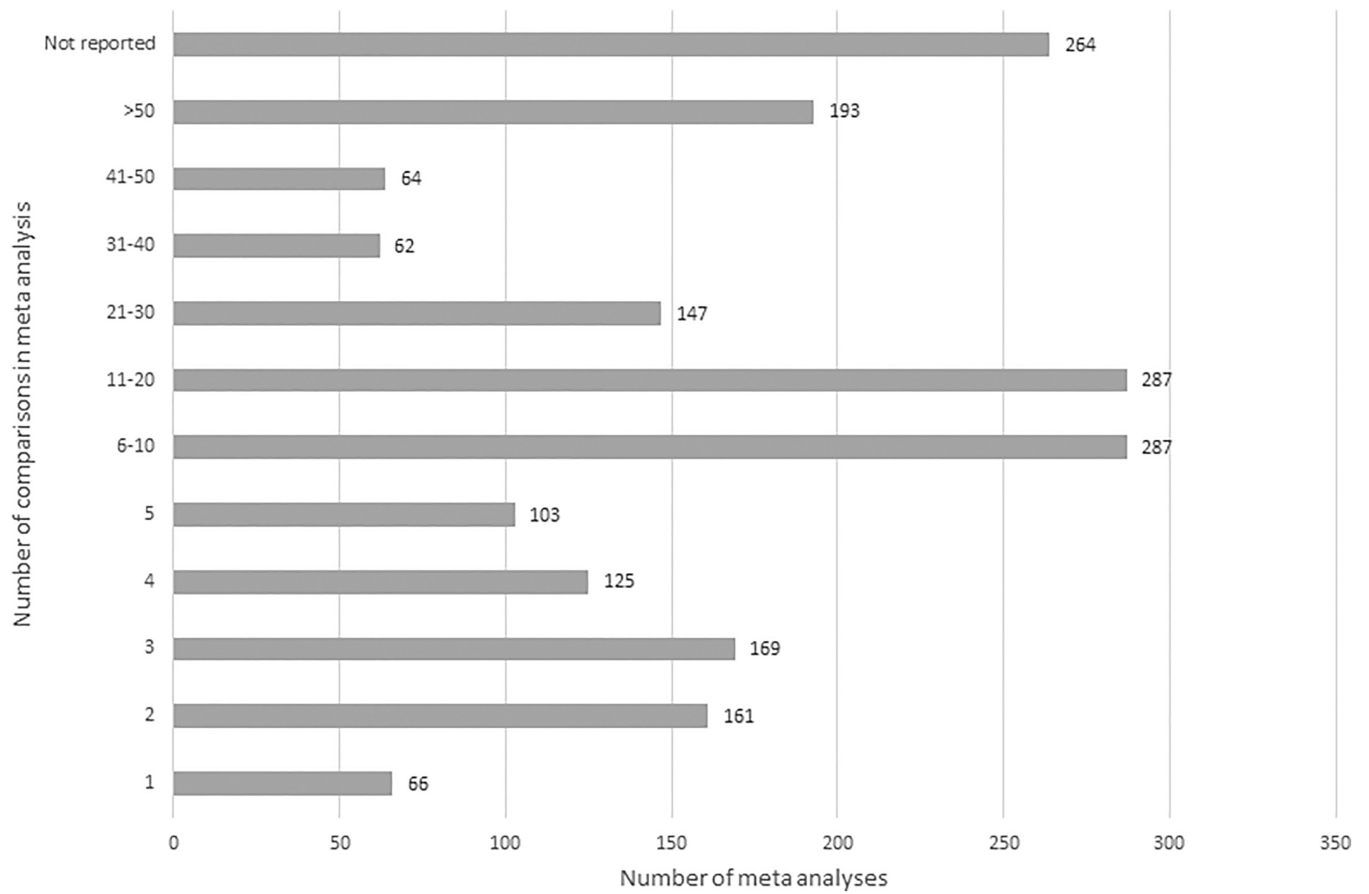


FIGURE 3 Number of meta-analyses relative to the number of comparisons included in meta-analysis

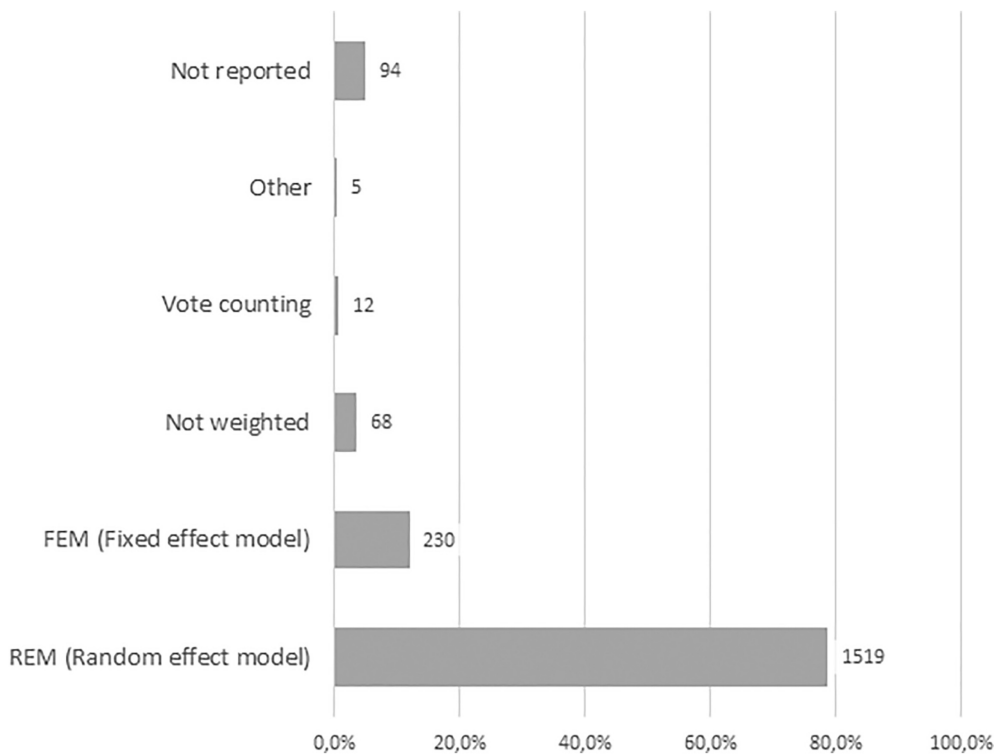
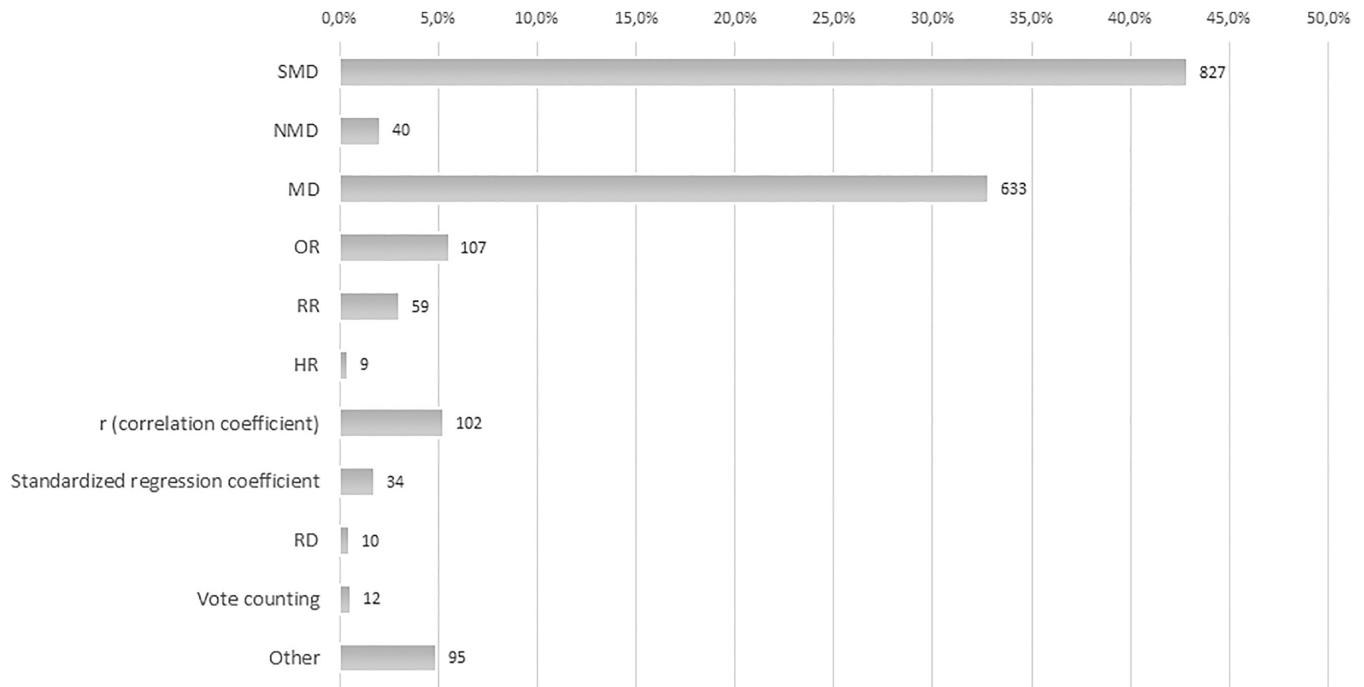


FIGURE 4 Type of model used to summarize the result of individual studies



**FIGURE 5** Type of effect size measures used in human health related meta-analyses. Abbreviations: HR, hazard rate; MD, mean difference; NMD, normalized mean difference; OR, odds ratio; RD, risk difference; RR, risk ratio; SMD, standardized mean difference

10% of the meta-analyses. The odds ratio (OR) was used most often. Most meta-analyses were conducted on continuous outcomes (78%). Those meta-analyses used either the NMD, SMD or MD. The SMD was used in the majority of the meta-analyses (55%) while the NMD was scarcely used (3%; 18 studies, 40 comparisons). In 13% of the studies containing meta-analyses, the rationale for a specific effect size measure was provided.

### 3.1.4 | Significance

Out of 1928 meta-analyses, 1630 (85%) published a confidence interval around the summary effect size. Out of these 1630 meta-analyses, 77% showed significant effects.

### 3.1.5 | Heterogeneity

In 67% of the meta-analyses, a measure for between-study heterogeneity was presented. The level of heterogeneity was described by either  $\chi^2$ ,  $I^2$ ,  $\text{Tau}^2$ , or a combination of those measures. In 87% of the meta-analyses, the  $I^2$  was presented, and the  $\chi^2$  and/or  $\text{Tau}^2$  was presented in 59% and 31% of the meta-analyses, respectively.

Across all meta-analyses that presented the  $I^2$  or from which the  $I^2$  could be calculated based on  $\chi^2$ ,  $\text{Tau}^2$ , and DF), the mean  $I^2$  was 63% (range 0%–100%).

Fourteen percent (14%) of all meta-analysis presenting an  $I^2$  level showed no heterogeneity at all (i.e.,  $I^2 = 0\%$ ).

Altogether, in 24% of meta-analyses, we considered  $I^2$  levels to be low (0%–40%) and in 25% of all meta-analysis the level was considered to be moderate ( $I^2$  41%–75%). Fifty-one percent (51%) of all meta-analyses presented high levels of between-study heterogeneity ( $I^2 > 75\%$ ).

In 58% of the meta-analyses using a FEM and 27% of the meta-analyses using a REM, the authors did not report any heterogeneity measures. In 17% of the meta-analyses using a FEM no between-study heterogeneity was observed, versus 10% for the meta-analyses using a REM. In the remaining 25% of the meta-analyses using a FEM (i.e., the meta-analyses with a between-study heterogeneity above 0%) the average amount of between-study heterogeneity was 60%. This was 72% in the meta-analyses using a REM.

### 3.1.6 | Subgroup analyses

Forty seven percent (47%) of the meta-analyses included one or more subgroup analyses. The mean number of subgroup analyses per meta-analysis was 4.1.



TABLE 1 Rationale for effect size selection

Effect size measure	Rationale provided in reference	Number of references using a similar rational
Mean difference (MD)	Because the outcome parameter was continuous	3
	To overcome the differences of sample size	1
	When sample size is small, weighted mean difference analysis performs better compared to SMD	1
	Data were presented in the same units and it gives a biologically relevant value.	2
	Because outcome remains easy to interpret	1
	comparable units of measurement	3
Normalized mean difference (NMD)	NMD analysis appears to perform better when the size of individual experiments is small compared to SMD, presumably because the observed variance used for weighting is a less precise estimate of the population variance than is the case for larger studies	2
	To account for different scales of measurement and differences between species	1
	This index of effect size enables outcomes that are measured on different scales to be combined in the same meta-analysis. The NMD has the additional advantage that the result is expressed in terms of percent reduction which may be easier to grasp than another common index of effect size, the standardized mean difference.	1
	Allows for correction of baseline values	1
	To compare effects of intervention across multiple studies with widely varying methods	1
Standardized mean difference (SMD)	To avoid heterogeneity (in methodology)	2
	To account for different scales of measurement and differences between species	1
	To account for different scales / units of measurement	25
	To account for different species	5
	Comparison between different studies	1
	More conservative than MD	1
	SMD was used because NMD was not possible, because there was no sham data	1
Hazard ratio (HR)	For survival HR most appropriate as because it allows for differences in sample size and time to an event.	1
Proportions	Because there was no control group	1
Regression coefficient (r)	Specific advantages (not mentioned in paper) over SMD	1

Out of the 913 meta-analyses that included subgroup analyses, 69 (8%) stated a minimum number of comparisons that is required to conduct subgroup analyses in their methods section. The mean minimum number was 3.3 comparisons.

If no threshold for subgroup analysis was explicitly stated, we assessed the size of the smallest subgroup presented in the publication. In 13% of all meta-analyses conducting subgroup analyses, the smallest subgroup contained only 1 comparison.

The majority of studies (45%) used stratified meta-analyses to analyze differences between subgroups. Meta-regression techniques were used in 20% of the meta-

analyses that conducted subgroup analyses. In 11% of the meta-analyses, both stratified and meta-regression techniques were used. In 22% of the meta-analyses that conducted subgroup analyses, it was either unclear or not reported which techniques were used.

### 3.2 | Reasons for the methodological decisions to use a specific effect size measure

Table 1 provides an overview of the rationales provided by the authors concerning their choice for a specific effect

size measure. The rationale was provided in only 8% of the included studies ( $n = 55$ ).

The MD appears to be used when the continuous outcomes are presented in the same units of measurements. One publication mentioned that when sample size is small, weighted mean difference analysis performs better compared to SMD.<sup>12</sup>

For the relatively new effect size measure, NMD (used in 18 studies), a rationale was provided in 33% ( $n = 6$ ) of the publications. Generally, authors indicated that the NMD was chosen to account for different scales of measurement and differences between species. Some authors stated that the NMD appears to perform better compared to SMD when the sample size of primary studies is small.<sup>9,13</sup> Last but not least some authors believe the interpretation of the NMD is easier to grasp compared to the SMD.<sup>14</sup>

Table 1 also shows that similar to the NMD, the SMD is also often used to account for different units of measurement and differences between species. A rationale for using the SMD was provided in 15% of the publications.

In 10% of the meta-analyses, dichotomous outcomes were analyzed, using either an odds ratio (OR), relative risk (RR), hazard ratio (HR), or risk difference (RD). The most commonly used effect size measure for dichotomous outcomes was the OR, which was used in 58% of the meta-analyses of dichotomous outcomes. The rationale provided by the authors for choosing a specific effect size measure for dichotomous outcomes was only provided in 3 publications and is listed in Table 1.

Because the rationale for choosing a specific effect size measure was reported in only a very limited number of publications, we attempted to gain more insight in the reasoning behind the choice of effect size measure by assessing the relationship between: 1) the use of a specific effect size measure and between-study heterogeneity (defined as  $I^2$ ); 2) the use of a specific effect size measure and the proportion of statistically significant findings; and 3) the use of a specific effect size measures between-study heterogeneity and the proportion of statistically significant findings. Of note, these analyses could only be performed for meta-analyses using the MD or SMD, because of insufficient data for all other effect size measures.

### 3.2.1 | Effect size measures and heterogeneity

In Figure 6, the distribution of  $I^2$  levels as a percentage of the total number of meta-analyses using a specific effect size measure (SMD, NMD, and MD) is plotted. Based on the definition for high/considerable heterogeneity in the Cochane handbook,  $I^2$  above 75%, meta-analyses using the NMD and the MD show higher between-study heterogeneity levels compared to the SMD. For the SMD 47% of the meta-analyses show an  $I^2$  above 75%, whereas for the MD and NMD this is 55% and respectively 62% of the meta-analyses. This tendency is also observed when calculating the average level of heterogeneity per measure. The average  $I^2$  level

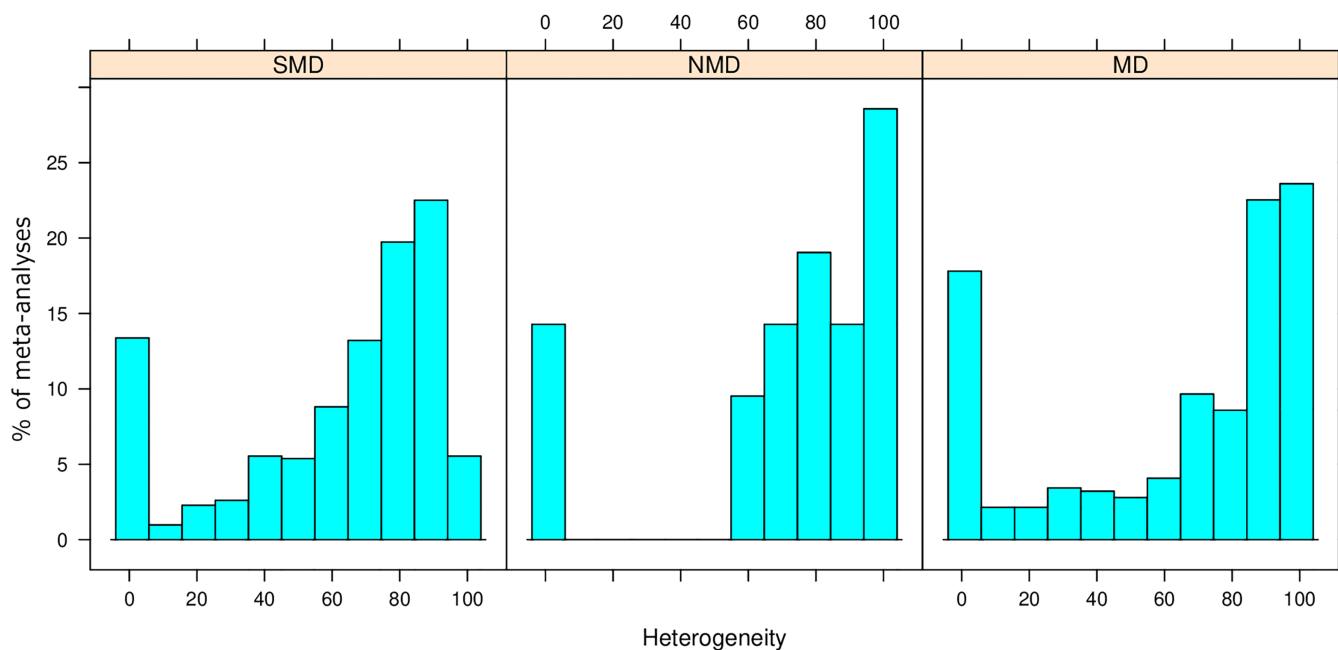
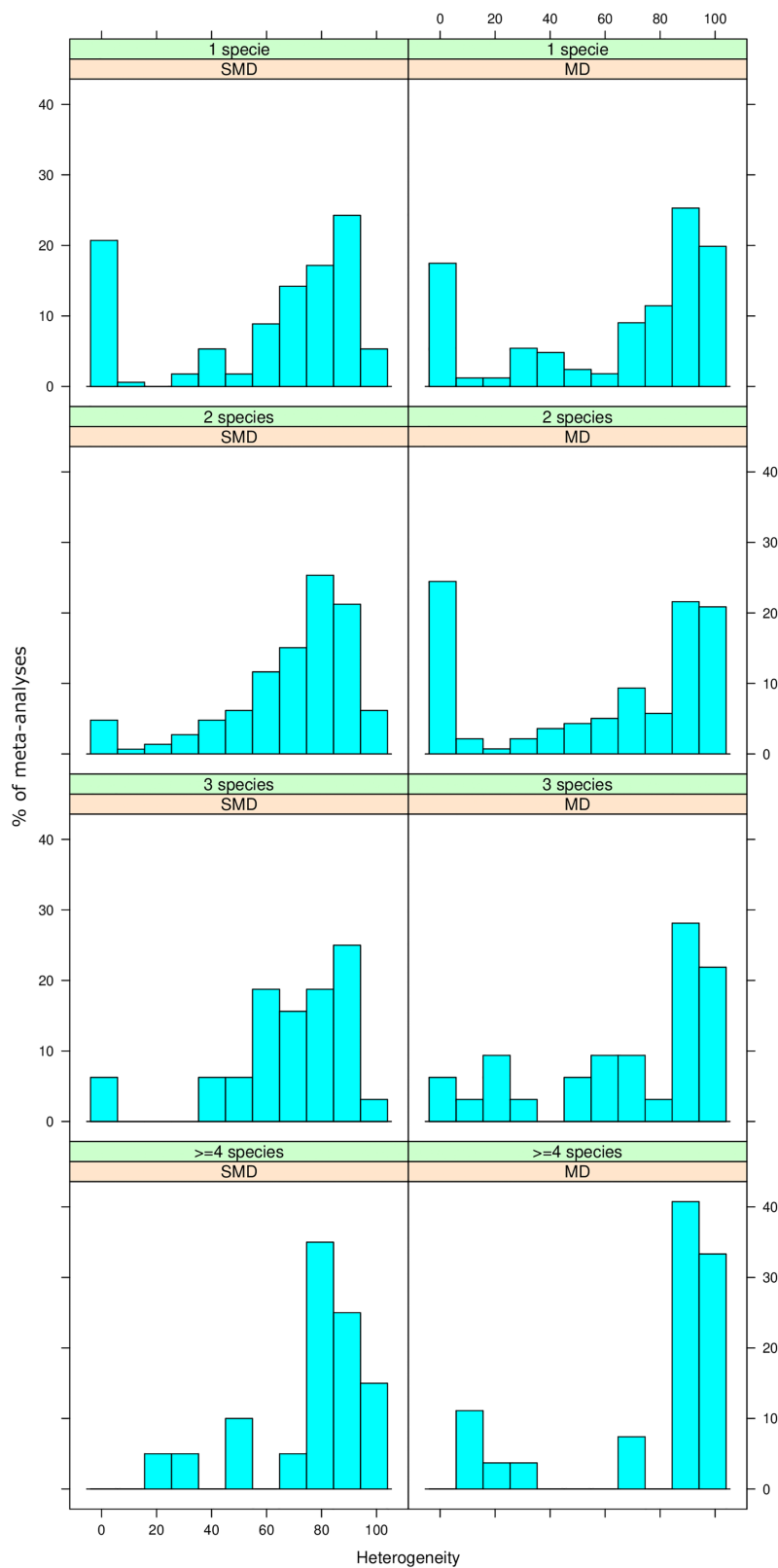


FIGURE 6 Proportion of meta-analyses relative to the amount of between-study heterogeneity per effect size measures. The amount of between-study heterogeneity is expressed as  $I^2$ . Abbreviations: MD, mean difference; NMD, normalized mean difference; SMD, standardized mean difference [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



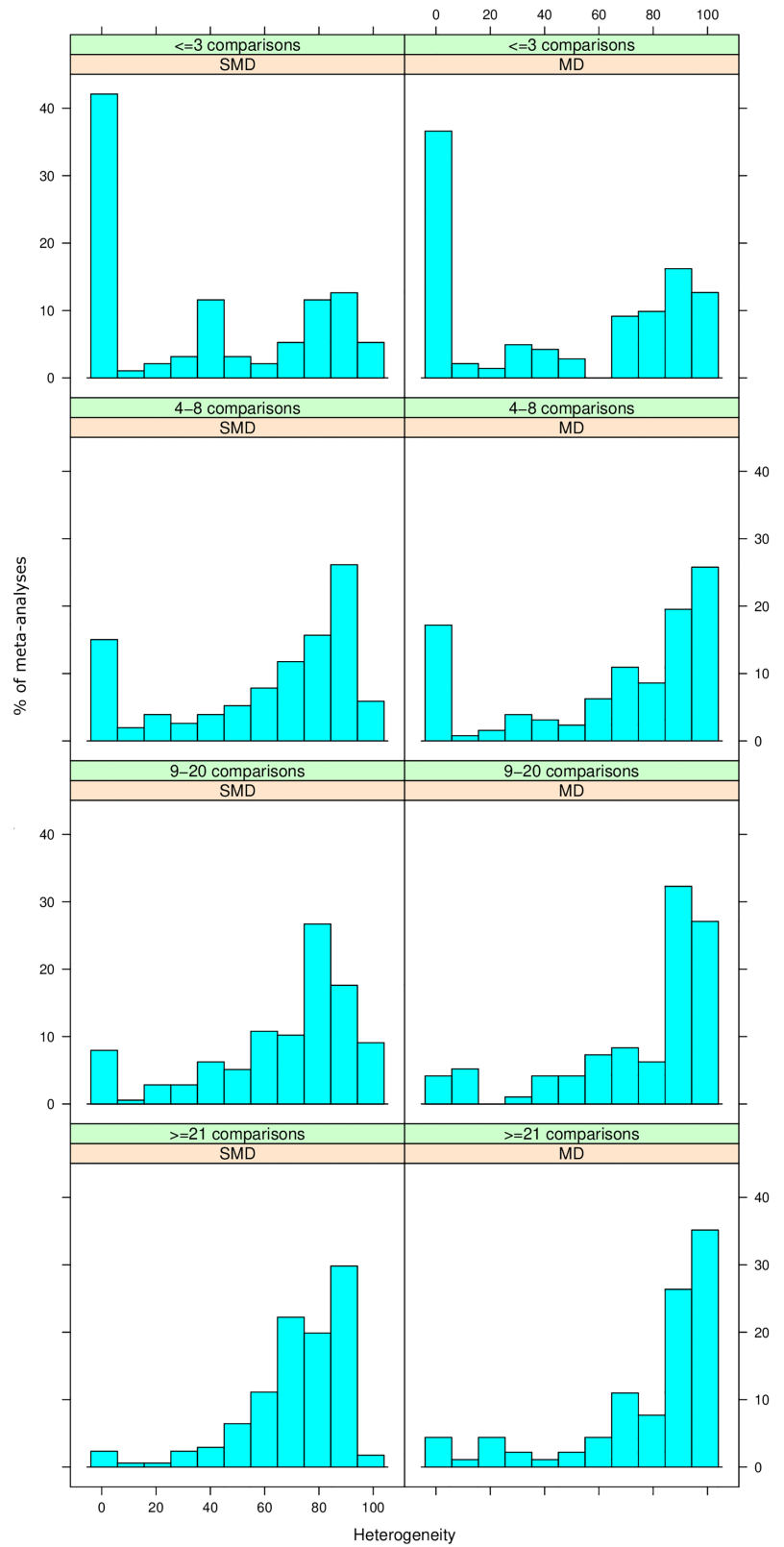
**FIGURE 7** Proportion of meta-analyses relative to the amount of between-study heterogeneity per effect size measures and per number of species group. The amount of between-study heterogeneity is expressed as  $I^2$ . Abbreviations: MD, mean difference; SMD, standardized mean difference [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

for the SMD was 62% and for the MD and respectively the NMD 72% and 65%.

When the  $I^2$  distribution is grouped per “number of species used in the analyses” we show another difference between the SMD and MD. For the SMD, the  $I^2$  levels

increase as soon as more than 1 species are included in the meta-analyses (the percentage of meta-analyses with  $I^2 = 0$  decreases). For the MD, a difference in distribution of  $I^2$  levels (e.g., the percentage of meta-analyses with  $I^2 = 0$  decreases) seems to occur when  $\geq 3$  species are used.

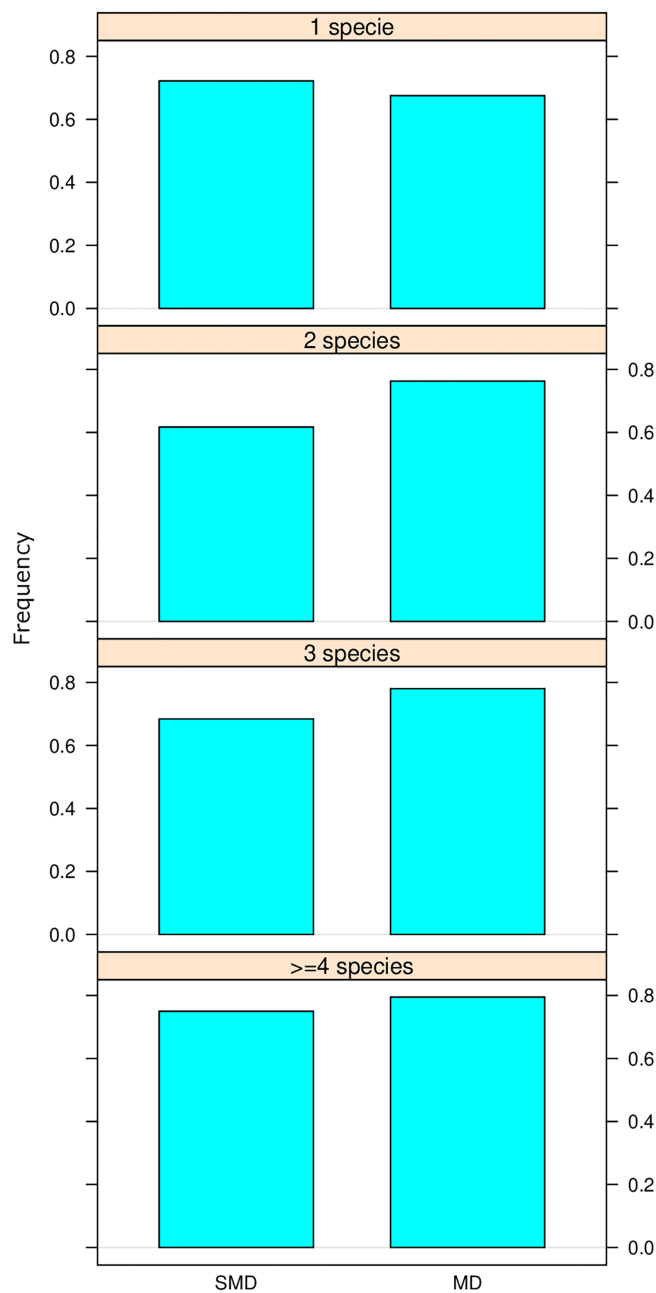
**FIGURE 8** Proportion of meta-analyses relative to the amount of between-study heterogeneity per effect size measures and per number of comparisons group. The amount of between-study heterogeneity is expressed as  $I^2$ . Abbreviations: MD, mean difference; SMD, standardized mean difference [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



In other words; the distribution of  $I^2$  levels shifts to the right (e.g., higher  $I^2$  levels) whenever multiple species are included. For the SMD this distribution shift starts as soon as 2 or more species are used, and for the MD when data from 3 or more species are included (Figure 7). The NMD

is not depicted in these figures because the number of meta-analyses using the NMD stratified for number of species was too small for reliable interpretation.

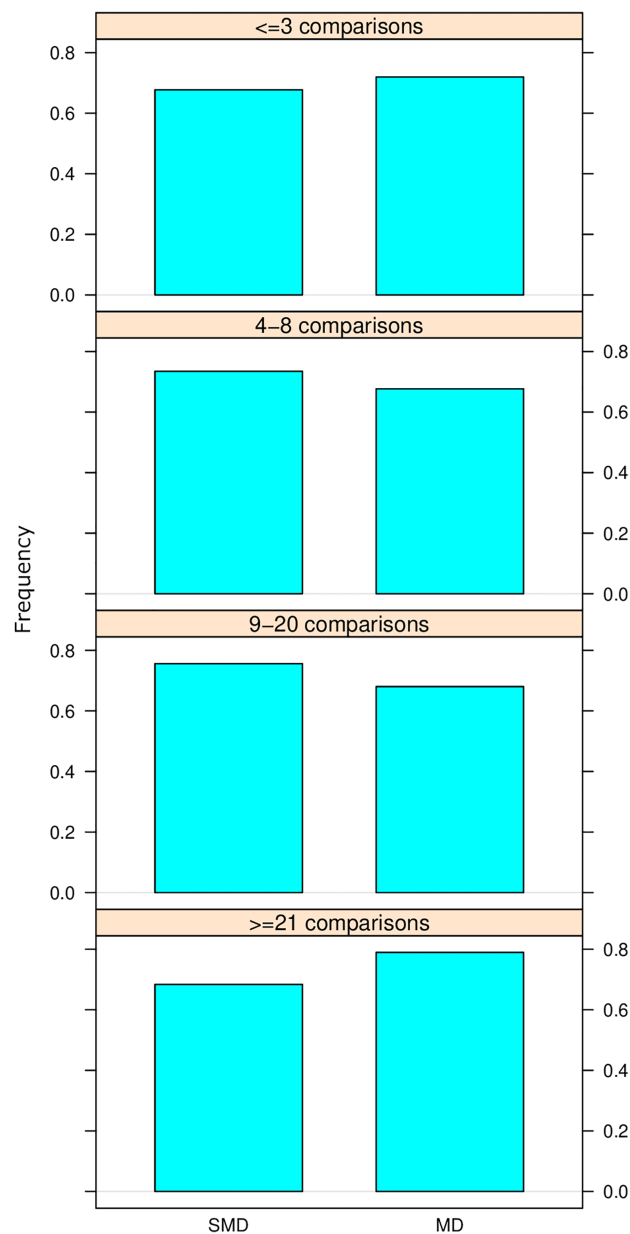
In addition, in meta-analyses including  $\geq 3$  species there is a larger proportion of studies showing very high



**FIGURE 9** The proportion of significant effects per effect size measure and per number of species group. Abbreviations: MD, mean difference; SMD, standardized mean difference [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

heterogeneity levels compared to the meta-analyses with a MD using one or two species. These proportions of studies with very high heterogeneity levels are also larger for meta-analyses using the SMD when including 4 or more species.

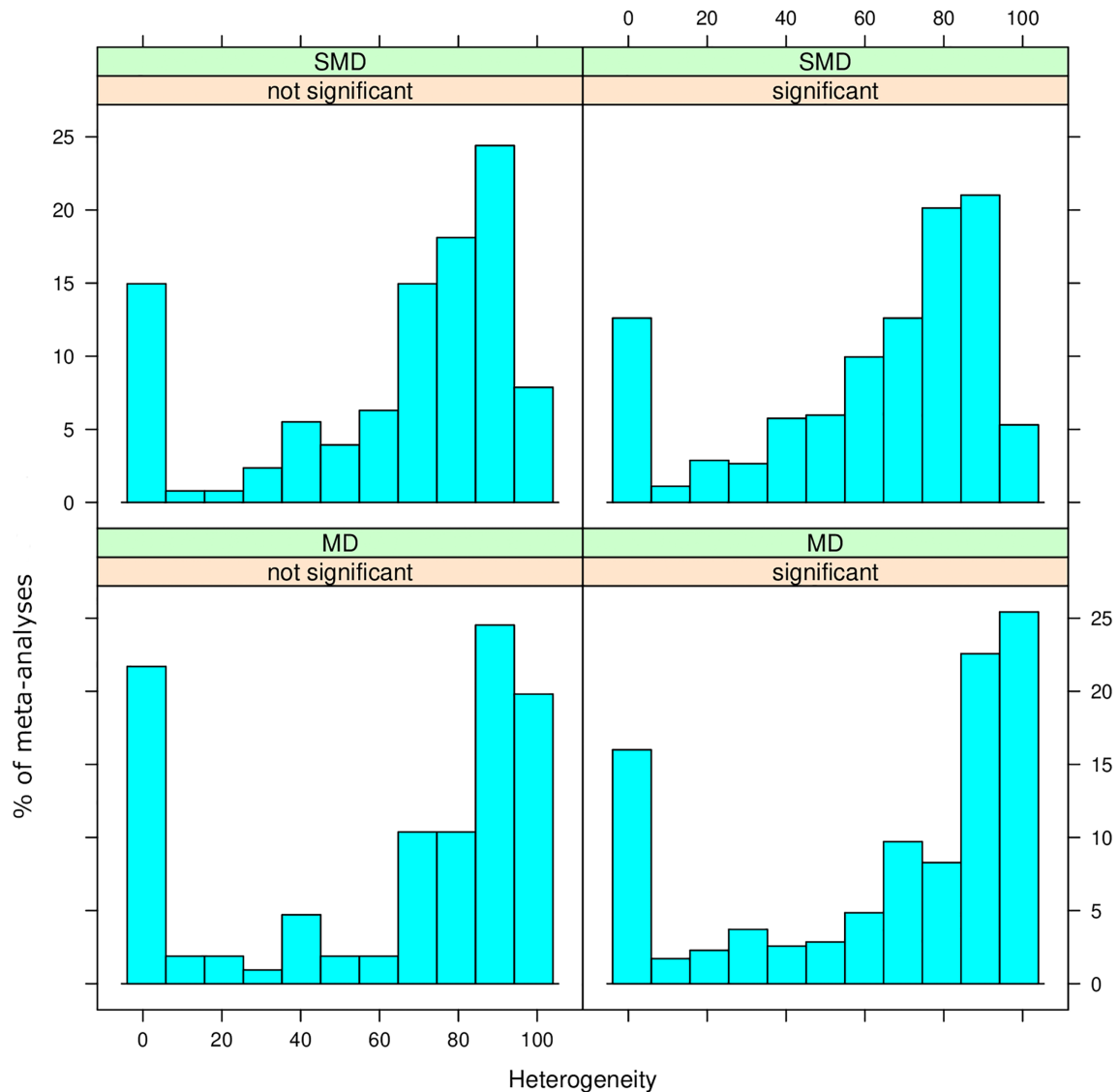
Figure 8 shows that the amount of heterogeneity seems to increase when the number of comparisons in the analyses increases. The same trend is seen for the SMD and MD.



**FIGURE 10** The proportion of significant effects per effect size measure and per number of comparison group. Abbreviations: MD, mean difference; SMD, standardized mean difference [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

### 3.2.2 | Effect size measures and statistical significance

The NMD shows the highest proportion of significant effects (SMD 79%, NMD 85%, MD 75%). Taking into account the number of comparisons and the number of species used in the analyses did not change this conclusion (Figures 9 and 10).



**FIGURE 11** The distribution of between-study heterogeneity analyses using either a MD or SMD that show significant or non-significant effects. Abbreviations: MD, mean difference; SMD, standardized mean difference [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

### 3.2.3 | Effect size measures, between-study heterogeneity, and the proportion of statistically significant findings

Figure 11 presents the distribution of between-study heterogeneity analyses using either a MD or SMD that show significant or nonsignificant effects. Figure 11 also shows that for the SMD the distribution of the proportion of meta-analyses across the various levels of between-study heterogeneity follows a similar pattern for significant and nonsignificant results. The distribution of the proportion of meta-analyses across the various levels of between-study heterogeneity for the MD do seem to differ between significant and nonsignificant results. There are fewer studies with high heterogeneity levels in the nonsignificant group compared to the group of meta-analyses with significant findings.

## 4 | DISCUSSION

In order to advance the methodology of meta-analysis of animal studies, we provided an overview of the key characteristics of published meta-analyses of human-health-related animal studies, focussing on important aspects of the methodology used and the effect size measures reported. In addition, we aimed to obtain insight into the rationale behind the choice of a specific continuous effect size measure, as the decision to use a specific effect size measure is not always straightforward but does have consequences for the reliability of the results and interpretation. To create this overview we used the recently published database of systematic reviews of animal studies (<https://data.mendeley.com/datasets/6fr3nw5mpc/1>).



## 4.1 | Overview of published meta-analyses of animal studies

Our overview shows 12% of the meta-analyses used a FEM to estimate the pooled effect. This is a reason for concern, as a FEM assumes that all studies estimate the same real effect, and differences between studies are only a consequence of random error (sample size). In the case of animal studies, there is often a lot of heterogeneity between the studies, as animal studies are often exploratory by nature and therefore show far greater diversity in the species and strains studied, the experimental design used, and other study characteristics. We show that in 30% of the meta-analyses that used a FEM, multiple species were involved, and that the average between-study heterogeneity levels were considerable (60%) in a quarter of the meta-analyses using a FEM. We therefore believe that a FEM was inappropriately used in part of the meta-analyses included in our database, and we recommend to always use the REM in human-health-related meta-analysis of animal studies (unless the assumptions for using the FEM are all met). Although it was beyond the scope of the current analyses, it would be interesting to investigate if the assumptions for using a FEM were correct in those reviews. For example, when significant small study effects are observed and a specific animal trial was dominating the meta analyses of a very homogeneous set of studies, a FEM may be appropriate. Nevertheless, we postulate that a FEM is rarely appropriate in meta-analyses of animal studies aimed at improving human health. The REM has the advantage of yielding almost identical results to a FEM when between-study heterogeneity is low, whereas it will safeguard against overly precise effect estimates when between-study heterogeneity is high.

We showed that in almost half of all meta-analyses subgroup analyses were conducted, and that the average prespecified minimum number of comparisons needed to perform subgroup analysis was relatively low (i.e.,  $n = 3$ ). In 13% of the subgroup analyses, the smallest subgroup contained only one comparison. Although we do not know the actual average number of comparisons across all subgroup categories, we assume that the studies that specified a low minimum number of comparisons in their subgroup categories, also included some small subgroups in their analyses. The statistical power for detecting differences between subgroups is therefore suggested to be low in many meta-analyses of animal studies, and conclusions based on subgroup analyses cannot be any more than hypothesis-generating. It is important to emphasize that failure to obtain statistical difference between those subgroups should never be interpreted as evidence that the covariate is not related to the effect size.

The relatively large number of subgroup analyses conducted per meta-analyses suggests that many of the meta-analyses were hypothesis generating (aim to explore which factors may influence the overall effect).

It would be interesting to know how often the subgroup analyses were pre-specified in a protocol, as pre-specifying the methodological details of the meta-analysis reduces the risk of inappropriate post-hoc analyses and selective outcome reporting (i.e., reporting only the results of subgroup analyses that show significant effects). Although one of the aims of meta-analysis of animal studies is to identify sources of heterogeneity, there is a risk of performing too many subgroup analyses, resulting in multiple comparison problems and loss of power. We therefore recommend to prespecify subgroups and restrict the number of subgroup analyses.

For this overview it was not possible to assess how often the subgroup analyses were prespecified in a protocol as the first protocol format for SRs of animal studies was published by SYRCLE in 2015<sup>15</sup> and before that SR protocols were not often prepared, and seldom publicly available. Since 2017, the protocol format for SRs of animal studies is implemented in PROSPERO, the international database of prospectively registered systematic reviews where there is a health-related outcome. In updates of the database we will be able to investigate if pre-specifying methodological details regarding the meta-analyses reduces reporting bias.

Pre-specifying methodological details in a protocol will probably also improve the reporting quality of essential details of the meta-analysis methods and results section in the ultimate publication. This is very much needed, as we show that important details regarding the methodology are often not reported. For example, the number of species included was not mentioned in 35% of the meta-analyses, the rationale for using a specific effect size measure was only provided in 13% of the meta-analyses, the effect model used was not described in 5% of the meta-analyses, and the number of comparisons included in analyses was unclear in 14% of all meta-analyses.

We strongly recommend to use the PRISMA checklist for reporting systematic reviews and meta-analyses of animal studies.<sup>16,17</sup> This checklist is not tailored to human-health-related animal studies but is, until a PRISMA extension for this type of studies becomes available, a very good alternative.

## 4.2 | Effect size measures and rationale

Only 13% of the meta-analyses reported a rationale for their decision to use a specific effect size measure. This percentage is surprisingly low, as the decision to use a

specific effect size measure is not always straightforward. We recommend to always present a rationale for the choice of the effect measure.

For the NMD, the percentage of studies providing a rationale was highest (36%,  $n = 6$ ). This is, however, not very surprising as a CAMARADES member was listed on the author list in the majority of papers using a NMD, and this leading center working on systematic reviews of animal studies developed the NMD.

Based on the rationales provided in the included papers, both the NMD and SMD generally appear to be used to account for different scales of measurement and differences between species. Using the SMD to take into account differences between species seems to be a relatively new application of the SMD, as the rationale in the Cochrane handbook reads: the SMD is used as a summary statistic in meta-analyses when the studies all assess the same outcome but measure it in a variety of ways. However, using the SMD also when studying different species/ populations is in a way not so different to regular use of the SMD because in both applications, the absolute effect of a specific outcome calculated in an individual study may mean something different across studies and therefore needs to be standardized.

Generally, the NMD was also chosen when authors needed to account for different scales of measurement and differences between species. However, some authors state that the NMD appears to perform better compared to SMD when the sample size of primary studies is small. However, we cannot investigate this in our review as the amount of studies using the NMD was too low ( $n = 14$ ).

Because the rationale for choosing a specific effect size measure was reported in only a very limited number of publications, and we wanted more insight in the reasoning behind the decision to use a specific effect size measure, we also assessed the existence of a possible relationship between the use of a specific effect size measure and between-study heterogeneity and the proportion of statistically significant findings. This is of interest because high heterogeneity levels generally cause wide confidence intervals around the summary effect and decrease the chance of finding significant effects, and the decision to use a specific effect size measure may influence these factors. It is also possible that it works the other way around, and the decision to use a specific effect size measure is based on the observed heterogeneity levels and the presence of statistically significant findings.

With regard with the SMD, we showed that the between-study heterogeneity levels seem to increase as soon as more than 1 species is used, whereas for the MD, the distribution starts to change when 3 or more species are used (Figure 7).

Including more species into the analysis would logically increase between-study heterogeneity levels, as observed for the SMD. It is surprising that this relationship is, not as obvious as for the SMD, seen for the MD.

One possible explanation might be that scientists decide post-hoc on which effect size measure to use. In other words, in cases where the between-study heterogeneity is low, scientists stick to the MD, but when the heterogeneity levels are higher they might search for other possibilities to summarize the effects (e.g., change effect size measures, and for example use the SMD).

In this meta research we identified another finding which suggests post-hoc decision making. We observed different distributions of the proportion of meta-analyses across the various levels of between-study heterogeneity for significant and non-significant results regarding the MD (Figure 11). More specifically, there are fewer studies with high heterogeneity levels in the nonsignificant group of the MD compared to the significant findings.

This may indicate that in case of high heterogeneity levels and no significant outcomes, authors make a post-hoc decision to switch to another effect size measure, and possibly search for a significant effect.

Further investigation into whether scientists make post-hoc decisions based on heterogeneity levels and/or significant findings is needed, but if scientists do make such decisions, it clearly advocates for preregistering a protocol before starting the conduct of a SR. regardless, we recommend to prespecify the effect size measure and the rationale behind choosing a specific effect size measure.

Future studies are needed to investigate if one effect size measure performs better than another. The database used as a resource in this paper is not appropriate for such assessments as the included data suffer from confounding. In other words: the decision of authors to choose an effect size measure could have been dependent on the observed results.

There is, however, one study that shows that the use of SMD estimates in stratified meta-analysis has a low statistical power to detect the effect of a variable of interest (compared to the NMD). This indicates that a SMD may not be very powerful in detecting an effect.<sup>18</sup> However, this study is a simulation study and might be prone to bias, as the assumptions underlying the simulation study might influence the result found. (e.g., the scientist decides upon which data to use).

In order to really compare the SMD, NMD, and MD, and get insight into whether or not a particular effect size measure performs better or is more reliable compared to another, we plan to conduct reanalyses of the data used in some published meta-analyses in the near future.

## 5 | CONCLUSION

Based on all analyses in this paper we recommend to 1) publish a review protocol before starting the conduct of a SR in which all methodological details are prespecified (providing special attention to the effect size measure and the rationale behind choosing a specific effect size, prespecifying subgroups and restricting the number of subgroup analyses), 2) always use the PRISMA checklist to report your systematic review of animal studies and 3) always use the REM in human-health-related meta-analysis of animal studies, unless the assumptions for using the FEM are all met.

### FUNDING INFORMATION

This work was supported by the National Institutes of Health under contract GS00Q14OADU417; HHSN273201600015U.

### CONFLICT OF INTEREST

The authors declare no potential conflicts of interest.

### DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the Supporting Information, File S2.

### ORCID

Carlijn R. Hooijmans  <https://orcid.org/0000-0001-6435-5714>

### REFERENCES

- Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10(1):101-129.
- Smith ML, Glass GV. Meta-analysis of psychotherapy outcome studies. *Am Psychol*. 1977;32(9):752-760.
- Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. *Nature*. 2018; 555(7695):175-182.
- Freedman LS. Meta-analysis of animal experiments on dietary fat intake and mammary tumours. *Stat Med*. 1994;13(5-7):709-718.
- Hooijmans CR, de Vries RBM, Ritskes-Hoitinga M, et al. Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLoS One*. 2018;13(1):e0187271.
- Hooijmans CR, Int'Hout J, Ritskes-Hoitinga M, Rovers MM. Meta-analyses of animal studies: an introduction of a valuable instrument to further improve healthcare. *ILAR J*. 2014;55(3): 418-426.
- Vesterinen HM, Sena ES, Egan KJ, et al. Corrigendum to 'Meta-analysis of data from animal studies: a practical guide': (*J Neurosci Methods* 221 (2014) 92-102). *J Neurosci Methods*. 2016;259:156.

- Zwetsloot PP, Van Der Naald M, Sena ES, et al. Standardized mean differences cause funnel plot distortion in publication bias assessments. *elife*. 2017;6.
- Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke*. 2008;39(10):2824-2829.
- Vesterinen HM, Sena ES, Egan KJ, et al. Meta-analysis of data from animal studies: a practical guide. *J Neurosci Methods*. 2013;221C:92-102.
- Langendam MW, Magnuson K, Williams AR, V.R. W, K.L. H, A.R. R, et al. Developing a database of systematic reviews of animal studies. *Regul Toxicol Pharmacol*. 2021;123: 104940.
- van der Worp HB, Sena ES, Donnan GA, Howells DW, Macleod MR. Hypothermia in animal models of acute ischaemic stroke: a systematic review and meta-analysis. *Brain*. 2007;130(Pt 12):3063-3074.
- Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of experimental stroke studies: a meta-epidemiologic approach. *Stroke*. 2008;39(3):929-934.
- Archer DP, Walker AM, McCann SK, Moser JJ, Appireddy RM. Anesthetic neuroprotection in experimental stroke in rodents: a systematic review and meta-analysis. *Anesthesiology*. 2017; 126(4):653-665.
- de Vries RBM, Hooijmans CR, Langendam MW, et al. A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies evidence-based preclinical medicine. *Evid Based Preclin Med*. 2015;1(2):1-9.
- Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.
- Page MJ, McKenzie JE, Bossuyt PM, et al. Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *J Clin Epidemiol*. 2021;134:103-112.
- Wang Q, Liao J, Hair K, et al. 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*. 2018;256776.

### SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Hooijmans CR, Donders R, Magnuson K, et al. Assessment of key characteristics, methodology, and effect size measures used in meta-analysis of human-health-related animal studies. *Res Syn Meth*. 2022;13(6): 790-806. doi:10.1002/jrsm.1578