The Usefulness of Systematic Reviews of Animal Experiments for the Design of Preclinical and Clinical Studies

Rob B. M. de Vries, Kimberley E. Wever, Marc T. Avey, Martin L. Stephens, Emily S. Sena, and Marlies Leenaars

Abstract

The question of how animal studies should be designed, conducted, and analyzed remains underexposed in societal debates on animal experimentation. This is not only a scientific but also a moral question. After all, if animal experiments are not appropriately designed, conducted, and analyzed, the results produced are unlikely to be reliable and the animals have in effect been wasted. In this article, we focus on one particular method to address this moral question, namely systematic reviews of previously performed animal experiments. We discuss how the design, conduct, and analysis of future (animal and human) experiments may be optimized through such systematic reviews. In particular, we illustrate how these reviews can help improve the methodological quality of animal experiments, make the choice of an animal model and the translation of animal data to the clinic more evidence-based, and implement the 3Rs. Moreover, we discuss which measures are being taken and which need to be taken in the future to ensure that systematic reviews will actually contribute to optimizing experimental design and thereby to meeting a necessary condition for making the use of animals in these experiments justified.

Key Words: experimental design; systematic review; meta-analysis; animal model; animal ethics; 3Rs; evidence-based preclinical medicine; translation

Introduction

The use of laboratory animals in (biomedical) research keeps provoking moral debates in society, which tend to revolve around two fundamental questions: (1) Are animal experiments morally acceptable? And (2) for which experiments are the expected benefits to humans (or to other animals or the environment) sufficient to outweigh the suffering of the laboratory animals? A third, and often absent, moral question is: How should the experiments that are justified be designed, conducted, and analyzed? Initially, this question appears to be scientific rather than moral, but if animal experiments are not appropriately designed, conducted, and analyzed, the results produced are likely to be unreliable. If the results of the experiments cannot be trusted, the animals used have in effect been wasted (Ioannidis et al. 2014). Such use and suffering of animals not counterbalanced by benefit in terms of science and/or human health is morally unjustifiable.

The aim of this article is to discuss the use of systematic review (SR) to address this third moral question. We will illustrate how the design, conduct, and analysis of future (animal and human) experiments may be optimized through systematic reviews of previously performed experiments.

A systematic review is a literature review to address a specific research question by seeking to identify, select, appraise, and synthesize all available research evidence relevant to that question (Egger et al. 2001). Systematic reviews follow a series of standard steps (see Box 1). This structured process highlights the differences between systematic reviews and classical, narrative reviews (for a summary of these differences, see Table 1). An SR often starts by formulating the research question that the review will try to answer. This research question tends to have a narrow focus (e.g., what is the effect of omega-3 fatty acids on Aβ deposition and cognition in animal models for Alzheimer's disease?). Authors of narrative reviews generally aim to provide an expert opinion on a certain research topic or give an overview of recent

Rob B. M. de Vries, PhD MA, is a scientific researcher at the SYstematic Review Centre for Laboratory animal Experimentation (SYRCLE) at the Radboud University Medical Center, Nijmegen, The Netherlands. Kimberley E. Wever, PhD, is a senior postdoctoral researcher at SYRCLE at the Radboud University Medical Center, Nijmegen, The Netherlands. Marc T. Avey, PhD, is a Canadian Institutes of Health Research (Knowledge Translation) postdoctoral fellow in the Clinical Epidemiology Program at the Ottawa Hospital Research Institute, Ottawa, Ontario, Canada. Martin L. Stephens, PhD, is a senior research associate at the Center for Alternatives to Animal Testing at Johns Hopkins University in Baltimore, Maryland. Emily S. Sena, PhD, is a senior postdoctoral research scientist at the University of Edinburgh, Edinburgh, UK, and a research officer at the Florey Institute of Neuroscience and Mental Health, Melbourne, Australia. Marlies Leenaars, PhD, is an assistant professor at SYRCLE, Radboud University Medical Center, Nijmegen, The Netherlands.

Address correspondence and reprint requests to Dr. Rob B. M. de Vries, SYRCLE, Central Animal Laboratory, Radboud University Medical Center, P.O. Box 9101, 6500 HB, Nijmegen, The Netherlands or email to Rob.deVries@radboudumc.nl.

ILAR Journal, Volume 55, Number 3, doi: 10.1093/ilar/ilu043

Box 1 Steps of a systematic review

1. Formulating a focused research question
2. Preparing a protocol
3. Defining inclusion and exclusion criteria
4. Systematically searching for original papers in at least two databases
5. Selecting the relevant papers
6. Assessing the quality/validity of the included studies
7. Extracting data
8. Data synthesis (if feasible, meta-analysis)
9. Interpreting the results

Table 1 Differences between systematic and narrative reviews

Feature	Narrative review	Systematic review
Research question	Often unclear or broad	Specified and specific
Literature sources and search	Not usually specified	Comprehensive sources (more than one database) and explicit search strategy
Study selection	Not usually specified	Explicit selection criteria and selection by two independent reviewers
Quality assessment included studies	Not usually present or only implicit	Critical appraisal on the basis of explicit quality criteria
Synthesis	Often a qualitative summary	Often also a quantitative summary (meta-analysis)

developments in a particular field. The inclusion of studies in the review is therefore often based on the authors' expert knowledge of the research field. This approach could introduce a risk of subjectivity in the selection of relevant studies. In order to reduce the risk of subjectivity, authors of SRs are encouraged to prespecify the different steps of the review in a protocol. As part of the protocol, the criteria for selecting studies are determined a priori. Thus, studies cannot be included or excluded based on the direction of their findings.

Because SRs aim to include *all* evidence relevant to the research question, they generally use comprehensive search strategies. To prevent missing relevant studies, authors of SRs are advised to search in at least two databases, for example, PubMed and Embase. Moreover, the search strategy is described in detail, enabling other researchers to replicate the search and to assess its completeness (Leenaars et al. 2012a). From the set of studies identified through the comprehensive search strategy, two independent reviewers select the papers to be included in the review on the basis of the previously defined inclusion and exclusion criteria.

After study selection has been completed, in most SRs, the methodological quality of the included studies is critically appraised (Henderson et al. 2013; van Luijk et al. 2014). This may include an assessment of their risk of bias, usually performed by two independent reviewers (Hooijmans et al. 2014; Krauth et al. 2013). Such an explicit and structured assessment of the reliability of studies included in a review is uncommon in narrative reviews. In addition to the quality assessment, the characteristics of the individual studies (design, species, intervention, outcome measures, etc.) are extracted. Where included studies contain quantitative outcome data and have sufficiently similar characteristics, these data may be statistically pooled using meta-analysis. Such an analysis produces a more precise estimate of the effect of an intervention. For a detailed explanation of the concept of meta-analysis and the added value of including such an analysis in an SR of animal studies, see the article by Hooijmans et al. elsewhere in this special issue.

SRs are common practice in clinical research, particularly for randomized controlled trials. Despite the fact that most animal experiments are performed to inform clinical research, SRs of animal experiments are still rather scarce (Korevaar et al. 2011; Peters et al. 2006; van Luijk et al. 2014). This is unfortunate, because SRs have several scientific advantages from the perspective of both human health and the 3Rs (Hooijmans and Ritskes-Hoitinga 2013; Sena et al. 2014). Here we will illustrate these advantages by showing in which ways SRs may help improve the design of new animal and clinical studies. In the final section, we will discuss measures to promote the implementation of SRs and thereby enhance their contribution to evidence-based preclinical science.

Advantages of Systematic Reviews for Designing Experiments

In order to draw reliable conclusions regarding the causal relationship between the intervention studied and the effects observed, it is vital that experiments are designed appropriately. Experimental design choices may include, but are not limited to, the choice of experimental and control groups, the determination of sample size, the choice of animal or disease model, and the measures taken to reduce the introduction of bias (Festing et al. 2002). In general, SRs of animal studies can guide the design of new experiments by demonstrating the extent of current evidence in the field and providing insight into which questions need to be addressed. In addition, SRs of animal studies can contribute to (1) improving the methodological quality of experiments, (2) an evidence-based choice of animal model, (3) evidence-based translation of animal data to the clinic, and (4) implementing the 3Rs.

Improving the Methodological Quality of Experiments

The term "methodological quality" can refer to the risk of bias in a study as well as to other methodological criteria such as imprecision or lack of power (Krauth et al. 2013). Risk of bias is the risk of systematic errors in the determination of the magnitude or direction of the results (Higgins and Green 2008). In this section, we will focus on risks of bias related to the internal validity of studies and how SRs can contribute to reducing those risks. Our general line of argumentation, however, also applies to other aspects of bias and methodological quality.

The credibility of the inferred causal relationship between an intervention and outcome is, in part, dependent upon the internal validity of the experiment. An experiment is internally valid if the differences in results observed between the experimental groups can, apart from random error, be attributed to the intervention under investigation. This validity is threatened by certain types of bias, where systematic differences between experimental groups other than the intervention of interest are introduced, either intentionally or unintentionally. Where such systematic differences occur, it is no longer clear whether differences in results between experimental groups are caused by the intervention under investigation. Such biases may be introduced at different stages of an experiment: they may be present in the baseline characteristics of the experimental groups (selection bias), in the care for the animals or the administration of the intervention (performance bias), in the way the outcomes are assessed (detection bias), and in the way dropouts are handled (attrition bias). These threats to internal validity can be greatly reduced by a combination of randomization and blinding on three levels: (1) the allocation of animals to experimental groups, (2) the administration of care and interventions during the experiment, and (3) the assessment of outcome (Hooijmans et al. 2014). Several studies have demonstrated that investigators rarely report measures to ensure the internal validity of experiments, such as randomization and blinding (Kilkenny et al. 2009; Mignini and Khan 2006; van Luijk et al. 2014).

SRs of animal studies may be used to demonstrate the importance of maximizing the internal validity of animal experiments to researchers and other stakeholders (e.g., policy makers and funding agencies). Quality assessments in SRs often contain several items related to internal validity (van Luijk et al. 2014). The results of such an assessment give insight into the extent to which the conclusions of the SR may be affected by biases related to internal validity. These results can be depicted per primary study, showing the lack of measures to reduce bias in a particular study, or per risk of bias item, showing the general score of the included studies for particular biases (for examples, see Table 2 and Figure 1). A factor currently hampering the assessment of the internal validity is the poor reporting quality of many animal studies. Because many details of the design and conduct of animal experiments are not reported, it is often unclear whether measures to preserve internal validity are not applied or whether they are applied but their application is not reported. In order to assess the actual risk of bias, the reporting quality of animal studies needs to improve (Hooijmans et al. 2014).

Moreover, SRs of animal studies can provide further empirical evidence that a lack of measures to reduce bias can lead to an overestimation or underestimation of the true effect of an intervention (Crossley et al. 2008). The results of the assessment of internal validity per study can be used in meta-analysis to compare subgroups of studies that, for example, did and did not report randomization of the allocation of animals to the experimental groups. Using this approach, an SR of therapeutic hypothermia in experimental models of stroke found that observed treatment effects were 10% larger in studies that did not report randomization and 8% larger in studies that did not report blinding than in those that did take these measures to reduce bias (van der Worp et al. 2007). Similarly, an SR of the drug NXY-059 in experimental stroke showed that the estimate of effect of NXY-059 was reported to be 30% larger in studies that did not report randomization or blinding than in studies that did (Figure 2) (Macleod et al. 2008). It is important to stress that this relationship between these measures to reduce bias and overestimation of effects has not been observed in all cases. In an SR of temozolomide in models of glioma, for instance, greater reductions in tumor volume were observed in blinded studies as compared to studies that did not report blinding (Hirst et al. 2013). However, this finding may be due to the fact that very few studies reported blinding (n = 2 vs. blinding not reported: n = 24), reducing the power of such an analysis. This highlights the limitations of this approach, which should be considered to be hypothesis generating rather than confirmatory.

SRs of clinical trials provided evidence of the number of studies that did not randomize or blind and of the impact of the lack of these measures to reduce bias on outcome. This raised awareness of the importance of preserving internal validity and thereby helped improve the design and reporting of new studies (Mullen and Ramirez 2006). We hope for similar improvements in the conduct and reporting of animal studies.

More Evidence-Based Selection of Animal Models

Another crucial aspect of the design of experiments is the choice of animal model. In this context, the term "animal model" does not only refer to the species or strain of laboratory animal, but also to the way in which a disease or defect is induced. Several studies have shown that the selection of animal models for experiments is not always evidence-based (de Vries et al. 2012; van der Worp et al. 2010a). Firstly, practical reasons—for example, costs of buying and housing the animals, ease of handling, and availability of biochemical

Table 2 Example of risk of bias assessment of individual studies (from Hooijmans et al. 2012)

Question Nr.	Akyol, 2003	Chen, 2007	Deng, 2000	Horst, 2009	Karen, 2010	Lutgendorff, 2008	Mangiante, 2001	Muftuoglu, 2006	Qin, 2006	Sahin, 2007	Tarasenko, 2000	v Minnen, 2006	Yang, 2006
1	yes	yes	yes	yes	yes	yes	no	yes	yes	no	no	yes	yes
2	?	?	?	?	?	?		?	?			yes	?
2							?			?	?		
3	?	?	?	?	?	?	na	?	?	na	na	?	?
4	?	?	?	?	?	yes	?	?	?	?	?	yes	?
5	yes*	?	?	**yes***	**yes***	**yes**	**yes**	**yes***	?	?	?	**yes**	?
6	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
7	?	?	?	?	?	**yes**	**no**	no	?	?	no	**yes**	**no**
8	?	?	?	?	?	?	?	?	?	?	?	?	?
9	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	yes #	**yes**	**yes**	yes	?	**yes**	?
10	**no**	**na**	na	**yes**	**yes ^^**	**na**	**na**	**na**	na	na	no	**yes**	yes

yes = low risk of bias; no = high risk of bias; ? = unclear risk of bias; Abbreviation: na = not applicable. * = assessment of the outcome measure histopathology was blinded, other relevant outcome measures were not blinded.

^^Risk of bias in the analysis because animals were replaced. # solely animals with severe acute pancreatitis are included in the analysis (risk of underestimating the effect of probiotics).

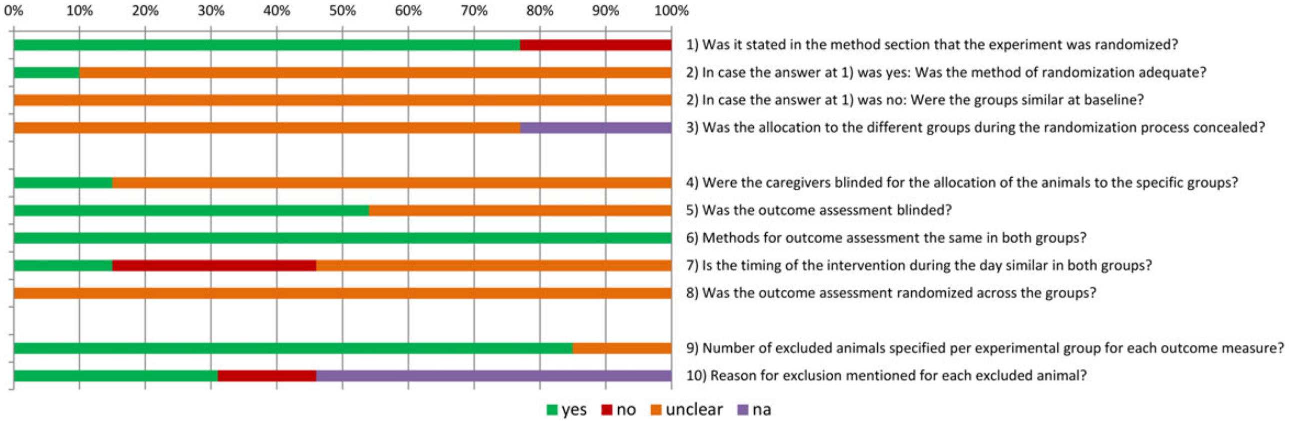

Figure 1 Risk of bias per item (from Hooijmans et al. 2012). Percentages at top refer to percentages of included studies with a particular risk of bias score. Yes = low risk of bias; no = high risk of bias; unclear = unclear risk of bias; na = not applicable.

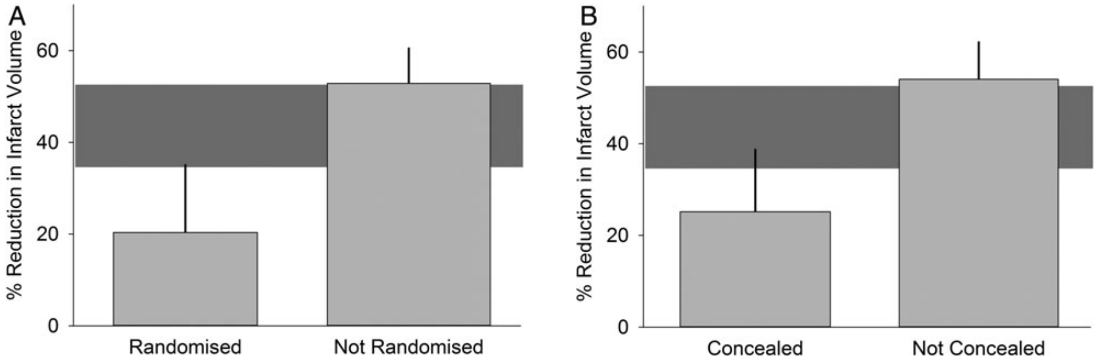

Figure 2 Subgroup analysis based on study quality (from Macleod et al. 2008). Grey horizontal bar depicts 95% confidence interval of overall effect estimate.

tests—tend to play as important a role in the selection of the animal model as the (anticipated) translational value. Secondly, the selection is often not the result of an explicit and extensive comparison between different potentially suitable models. A qualitative interview with researchers in the field of cartilage tissue engineering showed that some of them were only familiar with the characteristics of the model they themselves used and had very limited knowledge about alternative models (de Vries et al. 2012).

SRs support an evidence-based choice of animal models by providing a comprehensive overview of the models used so far, including their respective advantages and disadvantages. Thus, they provide evidence as to which model is likely to be most suitable for a new animal experiment. Alternatively, they might show that none of the available models is adequate and that new models need to be developed.

SRs may explicitly aim to provide the basis for selecting animal models (Ahern et al. 2009; Roosen et al. 2012). In addition, they may produce such evidence as a byproduct of answering another research question. In an SR of adaptive changes of mesenteric arteries in pregnancy, for instance, van Drongelen and colleagues (van Drongelen et al. 2012) found that the pathways involved in the response of mesenteric arteries to pregnancy vary considerably between different strains of rats (Table 3). Moreover, the response in Wistar rats appears to model the response seen during normal pregnancy in healthy women, whereas the response in Sprague-Dawley rats appears to model the response of women showing vascular maladaptation, such as preeclampsia. This finding underlines that researchers designing new animal experiments must be clear about which aspects of a condition they want to study and must be careful in selecting a species or strain of laboratory animal (van der Graaf et al. 2013).

Evidence-Based Translation

The majority of animal experiments are carried out to gather information about human health and disease. In preclinical research, animal experiments explicitly aim to investigate the safety and/or efficacy of interventions intended for use in humans. Moreover, the contribution animal experiments may make to developing new treatments for human diseases is an important reason for their moral justification.

It is becoming increasingly clear, however, that it is not straightforward to translate results found in laboratory animals to patients in clinical trials. In many cases, the

Table 3 Summary of qualitative changes in mesenteric artery adaptation to pregnancy (from: van Drongelen et al. 2012)

	Early gestation		Midgestation		Late gestation	
	WR	SDR	WR	SDR	WR	SDR
Vasodilator						
- Gq$_{EC}$	.	.	.	$=^{2}$	$=^{4}$	$\uparrow^{3}$
- Flow-mediated vasodilation	.	.	$\uparrow^{1}$	.	$\uparrow^{1}$	$\uparrow^{1}$
- Vascular compliance	.	$\uparrow^{1}$	$=^{1}$	$\uparrow^{1}$	$=^{1}$	$\uparrow^{2}$
- Gs$_{SMC}$	.	.	$\uparrow^{1}$	.	$\downarrow^{1}$	$\uparrow^{4}$
Vasoconstrictor						
- Gq$_{SMC}$	.	.	$=^{1}$	$=^{4}$	$=^{6}$	$\downarrow^{12}$
- Myogenic reactivity	.	$=^{1}$	$=^{1}$	$=^{1}$	$=^{2}$	$?^{3}$

Pregnancy-induced vascular function: increase (↑), decrease (↓), no change (=), inconsistent effects (?), no effects reported (.). Superscripted values represent number of responses on which the effect is based.

Abbreviations: WR, Wistar rat; SDR, Sprague-Dawley Rat; Gq/Gs, G-protein coupled receptor pathway; EC, endothelial cell; SMC, smooth muscle cell.

predictive value of animal experiments is low (McGonigle and Ruggeri 2014). One of the most dramatic examples is stroke research. Over the past three decades, over a thousand interventions for stroke have been tested for safety and efficacy in animal experiments. More than 500 of these interventions showed evidence of efficacy in animal tests, but so far only thrombolysis with tissue plasminogen activator (tPA) has proved to be effective in stroke patients (O'Collins et al. 2006).

It is plausible that the striking differences in results between animal and human studies are partly due to fundamental biological/physiological differences between humans and other species. However, other, avoidable factors related to the design, conduct, and reporting of preclinical animal experiments may play an equally important role, notably (1) poor methodological quality, (2) differences in design between experimental animal studies and clinical trials, and (3) publication bias. Avoiding these factors may improve the internal and external validity and thereby the predictive value of animal experiments (Hooijmans and Ritskes-Hoitinga 2013; Sena et al. 2014; van der Worp et al. 2010a).

SRs are useful to identify these design-related factors in the short run and to promote their avoidance in the long run. In 2008, the results were published of a clinical trial that was testing probiotics as an alternative to antibiotics for the treatment of acute pancreatitis (Besselink et al. 2008). No differences between experimental groups were found for any of the primary endpoints. Moreover, mortality in the probiotics group was significantly higher than in the placebo group. This outcome was unexpected in light of the results of the animal studies preceding the start of the trial. However, an SR of the animal data by Hooijmans and colleagues (2012) revealed that, prior to the start of the clinical trial, no animal experiment with a design similar to the trial had been carried out. None of the animal experiments used the same probiotics as were used in the trial (Ecologic 641), the probiotics were often administered before the induction of pancreatitis (prophylactically rather than therapeutically), and none of the animal studies used an intrajejunal administration route, as was used in the clinical trial (Hooijmans et al. 2012). Similar discrepancies were found in an animal SR of the antioxidant tirilazad (Sena et al. 2007), which appeared to be effective in animal models of acute ischemic stroke, but which increased the risk of death and dependency in patients. Sena and colleagues found that time to treatment was substantially longer in the clinical studies (median 5 h) than in the animal studies (median 10 min). Moreover, only a small number of studies used comorbid animals, whereas many comorbidities such as hypertension, diabetes, or hyperlipidemia are common in stroke patients.

Further, the conduct of an SR of animal studies before the start of a clinical trial can help establish whether there is sufficient evidence of sufficient quality to justify the trial and can inform the design of the trial. An example is the SR by Wever and colleagues of ischemic preconditioning (IPC) as a therapy against renal damage (Wever et al. 2012). IPC is a strategy in which brief periods of ischemia and reperfusion are used to induce protection against subsequent ischemia-reperfusion injury, for example, after kidney transplantation or cardiovascular surgery. Wever and colleagues showed that IPC protocols studied in animal models were diverse in terms of timing, duration, and the number of ischemic and reperfusion periods used. In the animal studies, IPC protocols applied 24 hours or more in advance of the prolonged ischemic insult reduced renal injury more effectively than protocols performed within 24 hours of the index ischemia. Interestingly, remote IPC (where the brief ischemic stimuli are not applied to the kidney itself, but to another organ or tissue) was not studied extensively in animals, even though it is the preferred method of IPC in human patients. Meta-analysis suggested that these two types of IPC might be equally effective in animal models of renal reperfusion injury. Up to 2012, all clinical trials applied nearly identical IPC protocols: three or four cycles of five minutes ischemia and reperfusion, applied directly before the index ischemia.

Thus, in light of the animal data, clinical trials of therapeutic IPC for renal injury may have been suboptimally designed. Based on the results of their animal SR, Wever and colleagues designed a clinical trial in which IPC will be applied either directly before renal damage, 24 hours in advance, or both. Furthermore, since the publication of this SR, several clinical trials have been registered on Clinicaltrials.gov that apply IPC 24 hours or more before ischemic injury (e.g., trial numbers

NCT01903161, NCT01658306, and NCT01739088). Similarly, an SR of therapeutic hypothermia for animal models of ischemic stroke was used to inform the design of the EuroHYP-1 clinical trial (van der Worp et al. 2010b; van der Worp et al. 2007).

Unfortunately, the use of SRs of animal studies to help improve the translational value of animal experiments is likely to be further hampered by the presence of publication bias. Publication bias is bias caused by the phenomenon that studies reporting statistically significant data are much more likely to get published than studies reporting neutral or negative data (Higgins and Green 2008; Song et al. 2010). There are strong indications that publication bias is far more abundant in the field of animal studies than in clinical research (Korevaar et al. 2011; ter Riet et al. 2012). Meta-analyses have suggested that publication bias may lead to major overstatements of treatment effects (Sena et al. 2010b). Funnel plot inspection, Egger regression, and trim-and-fill analysis are common tools in SRs to assess the likelihood of publication bias and its impact on the conclusions drawn (see also the article by Hooijmans et al. in this special issue).

An essential role in the long-term solution to minimizing publication bias may be played by registries, similar to the ones established in the field of clinical research (for example, ClinicalTrials.gov) (Dickersin 1990). However, there are likely to be more impediments to establishing such registries in the preclinical field. Major objections to the registration of preclinical studies are that such registries might (1) threaten the strategic advantage (private) drug developers have over their competitors; (2) create an undue administrative burden for researchers, especially in basic science; and (3) prove too costly (Kimmelman and Anderson 2012). These objections can largely be met, however, by appropriately streamlining the content of such registries and limiting access to sensitive information.

Implementation of the 3Rs

Apart from their potential use in designing new animal (and human) studies, SRs may also contribute to the development of alternatives to animal experiments. The term “alternatives” is taken here to refer to the 3Rs of replacement, reduction, and refinement (Russell and Burch 1959).

Replacement

SRs alone will rarely replace animals or animal experiments directly, although they can prevent unnecessary duplication of experiments or show that proposed experiments do not add substantially to current knowledge. However, they can contribute indirectly to replacement by supporting the development or validation of replacement alternatives. For instance, an SR of replacement alternatives based on tissue engineering (de Vries et al. 2013) demonstrated that the potential for the development of these alternatives is broader than use of tissue-engineered skin for toxicological applications. Previous, narrative reviews had either discussed only a few examples or addressed only one area of application, for example, safety testing. By providing a comprehensive overview, the SR helped to make both tissue engineers and alternative experts aware of the full range of possibilities of using tissue-engineered constructs as a replacement of laboratory animals.

Additionally, SRs of animal studies can be incorporated into assessments of how well alternative test methods perform in comparison to the animal-based methods they are intended to replace. Such comparative reviews could be loosely likened to retrospective validation (Hartung 2010). The Evidence-based Toxicology Collaboration (EBTC; www.ebtox.com) is pioneering this application of SRs in an assessment of the performance of the Zebrafish Embryo Test (ZET) in predicting the results of prenatal developmental toxicity tests (see Test Guideline 414 of the Organisation for Economic Co-operation and Development [2001]). The goal of such developmental toxicity tests is to assess the effects of chemical exposure on the growth and development of the fetus. The current mammalian-based tests have significant limitations; they are animal use-intensive, costly, and time-consuming (Selderslaghs et al. 2012).

The goals of the EBTC’s SR are to assess the ZET’s performance compared to that of the established mammalian tests and, more broadly, to serve as a methodological case study to assess the feasibility of using an SR-based approach for the evaluation of (alternative) test methods. The ZET is currently considered a screening tool for prenatal developmental toxicity (Adler et al. 2011; Basketter et al. 2012). A good test performance would underscore the ZET’s use as a screening tool and possibly provide evidence to extend its use as a partial substitute for mammalian testing. The EBTC also plans to assess in a similar way the performance of the high-throughput cellular and biochemical assays at the heart of “21st century toxicology” (Stephens et al. 2013). Such assays typically do not correspond, one to one, to existing methods, making their validation especially challenging.

Reduction and Refinement

SRs contribute to reduction by making further use of the data already available, thereby producing new scientific information without the use of new animals. Moreover, SRs may improve the design and therefore the relevance and reliability of new experiments, leading to more reliable information from the same number of animals.

It is still unclear, however, whether the large-scale application of SRs of animal studies will reduce the absolute number of animals used. An SR of animal models of multiple sclerosis (Vesterinen et al. 2010) showed that many animal experiments in the field are underpowered and that adequately powered experiments would have required more, rather than fewer, animals. However, such adequately powered experiments would produce more reliable data that would require

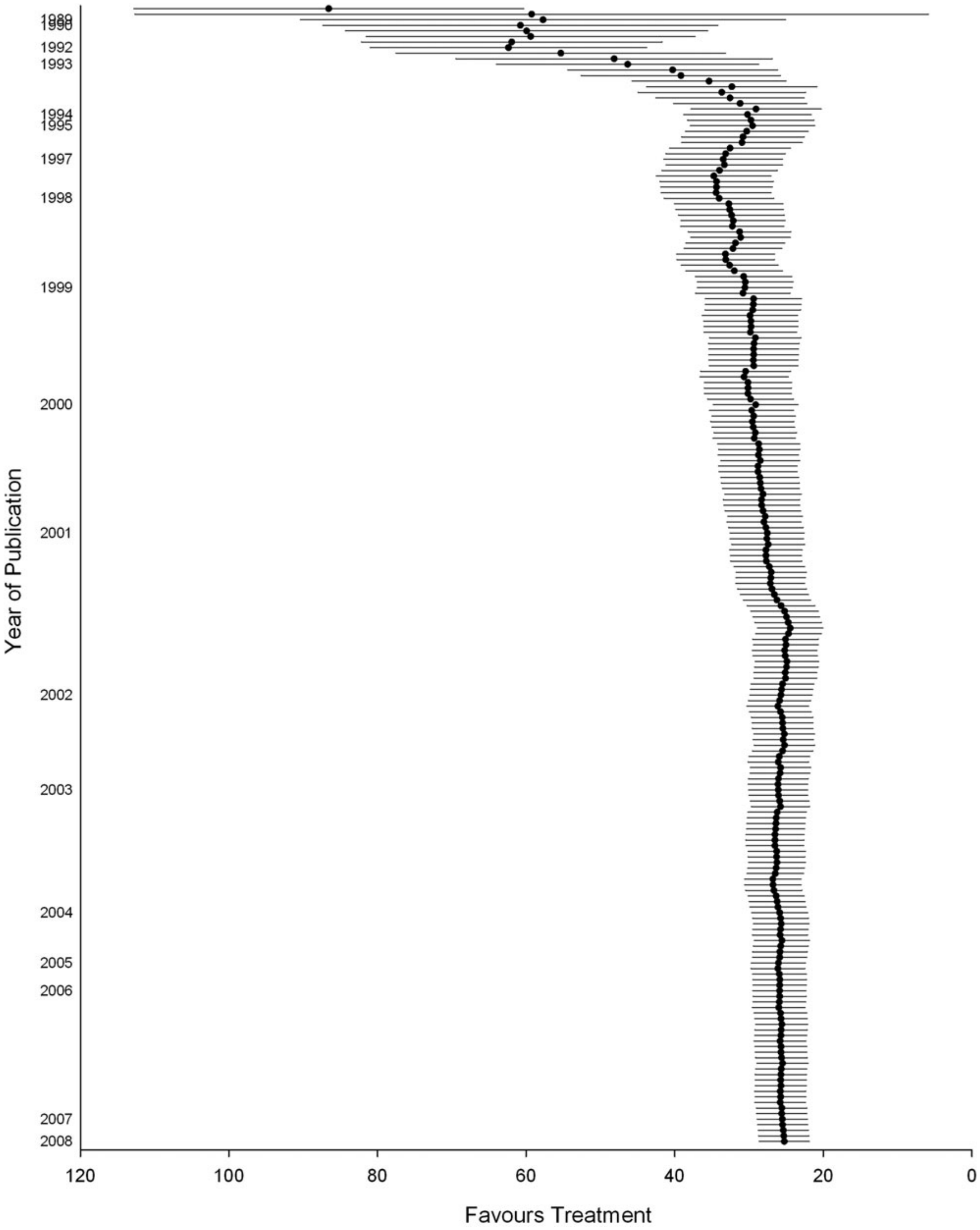

Figure 3 Cumulative meta-analysis of the effect of tPA on stroke (each time a new animal study is published, the overall effect size is recalculated for all studies available at that time, resulting in an increasingly more precise estimate of the effect of the intervention) (from Sena et al. 2010a). Values expressed as effect size + 95% confidence intervals.

larger but probably fewer experiments and that would at least prevent waste of animals.

On a smaller scale, SRs can reduce the number of animals by preventing the conduct of experiments not necessary to establish a certain effect of an intervention. Cumulative meta-analysis has great potential for this purpose. In a cumulative meta-analysis, studies are sequentially included in the meta-analysis, and the point at which sufficient data exist to show stability of a treatment effect can be observed. This technique has been used in clinical studies to identify at which point there are sufficient data to refute or support drug efficacy and whether further trials are required (Lau et al. 1992). An SR of tPA in models of stroke (Sena et al. 2010a) showed that a total of 3,388 animals had been used. A cumulative meta-analysis included in this SR suggested that the estimate of efficacy was stable from around the inclusion of 1500 animals (see Figure 3). It is important to note that many of the experiments in this SR were using tPA as a positive control (as the only clinically effective treatment for ischemic stroke). However, this technique has the potential for novel interventions to ascertain where sufficient data exist for evidence of a stable treatment effect so that further animal studies are not required to demonstrate efficacy.

An example of the contribution SRs can make to refinement is provided by an SR of the cisplatin-induced ferret model of emesis (Percie du Sert et al. 2011). Their meta-analysis on the effects of ondansetron provided evidence that the observation period in studies of anti-emetics could be reduced from 24 hours to 4 hours. Similarly, an SR by Percie du Sert and colleagues (2012) suggested that refinement could be attained by using rats instead of nonhuman primates in self-administration studies to determine the reinforcing properties of opioid drugs. Currie and colleagues are currently conducting an SR of animal models of neuropathic pain (see www.camarades.info for their SR protocol). One of the objectives of this SR is to establish whether refinements are possible by using tests with a lower burden of pain/distress, avoiding multiple tests, and shortening the duration of the tests.

Progress in Implementing Systematic Reviews

From a scientific and moral perspective, animal experiments should be appropriately designed, correctly performed, thoroughly analyzed, and transparently reported. In this final section, we want to highlight a number of recent and future initiatives/activities that have been, are, or will be undertaken in order to ensure that SRs of animal studies are conducted and thereby actually contribute to achieving that ideal situation (Hooijmans and Ritskes-Hoitinga 2013).

Two major research groups involved in the promotion of SRs of animal studies are (1) the Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES; www.camarades.info) group and (2) the SYstematic Review Centre for Laboratory animal Experimentation (SYRCLE; www.SYRCLE.nl). CAMARADES is routinely performing systematic reviews of preclinical animal models of disease and has formed a worldwide network. SYRCLE has focused on the development of methodology and guidelines and offers teaching and training internationally, in addition to performing collaborative systematic reviews (Ritskes-Hoitinga et al. 2014).

In 2011, the Montréal Declaration on the synthesis of evidence to advance the 3Rs principles in science was initiated to make people aware of the need for a change in animal research. It was adopted by the participants of the 8th World Conference on Alternatives and Animal Use in the Life Sciences (Leenaars et al. 2012b). This Declaration is calling for a change in the culture of planning, executing, reporting, reviewing, and translating animal research via the promotion and coordination of synthesis of evidence, including systematic review of animal studies.

The number of SRs performed is increasing (Korevaar et al. 2011; van Luijk et al. 2014), and many examples demonstrate its value. However, most animal researchers presently receive no or only limited training in SR methodology and are unaware of its availability and potential. To achieve more general awareness of the availability of the methodology and its added value, and to have high-quality SRs performed on a large scale, education and training in SR methodology are needed. The Dutch parliament recently accepted a motion stating that education and training in systematic reviews of animal studies should be part of the course on laboratory animal science for animal researchers (comparable with Federation of European Laboratory Animal Science Associations [FELASA] category C or EU2010/63: article 23.2.b functionary) in The Netherlands. In the context of continuing professional development, SYRCLE is currently providing education and training to researchers in The Netherlands funded by The Netherlands Organisation for Health Research and Development (ZonMw). Introduction of the SR methodology into the curriculum of BSc and MSc students in relevant fields such as biology and biomedical sciences is ongoing and can be a great opportunity to introduce the next generation of researchers to the concept.

In order to facilitate the conduct of SRs of animal studies, a number of useful tools have been developed, such as the step-by-step guide to find all animal studies (Leenaars et al. 2012a); search filters (de Vries et al. 2011; 2014; Hooijmans et al. 2010); a risk-of-bias tool (Hooijmans et al. 2014); a practical guide to meta-analysis (Vesterinen et al. 2014); and suggested guidance for the conduct, reporting, and critical appraisal of SRs (Sena et al. 2014). Nevertheless, there is a need for further development of methodology tailored to the conduct of high-quality SRs in preclinical animal studies. In this respect, much can be learned from the Cochrane Collaboration, an international organization of more than 30,000 scientists, which has been collaborating on methodology, guidelines, education, and conduct of SRs of clinical trials for more than 20 years. The first steps toward establishing a Preclinical Animal Study Methods Group, in close cooperation with the Cochrane Collaboration, are currently being taken (Ritskes-Hoitinga et al. 2014).

Through these activities, SRs will become common practice in the field of animal studies and thereby help improve the design of future animal and human studies. So far, SRs have made a major contribution to the growing body of evidence that improper conduct of animal studies generates unreliable data and is unlikely to lead to clinical benefit. By demonstrating the consequences of poor experimental design, SRs have initiated the first steps in a culture shift toward an improved standard of practice in the field of animal experimentation. During this process, SRs act and have acted in synergy with other initiatives, such as reporting guidelines (e.g., Animal Research: Reporting of in Vivo Experiments [ARRIVE] guidelines, Gold Standard Publication Checklist, and ILAR Guidance) and education on experimental design (e.g., FRAME's training schools). As awareness continues to grow (through publication of SRs, international meetings, and education), researchers, journals, ethics committees, and funding bodies will be motivated to strive for better preclinical science. This in turn encourages the use and enforcement of reporting guidelines and other initiatives concerning optimal registration, conduct and reporting of animal studies. Such measures will directly improve the design of experiments, but also facilitate the conduct of more high-quality SRs (which, for instance, can assess the actual risk of bias in the included studies or draw more reliable conclusions regarding the influence of study characteristics such as the sex of animals). Through awareness, evidence, and education, the much-needed new global standard of practice for animal experiments will hopefully be achieved in the near future. This new standard will ensure that the laboratory animals used are not wasted and their suffering is counterbalanced by maximal human benefit.

Acknowledgments

Rob de Vries and Kimberley Wever received funding from The Netherlands Organisation for Health Research and Development (ZonMw; grant nr. 104024065) and the Dutch Ministry of Health (grant nr. 321200). Marc T. Avey is funded by a postdoctoral fellowship from the Canadian Institutes of Health Research (Knowledge Translation). We are grateful to Lisa A. Bero, PhD, for helpful comments on an earlier version of the manuscript. This article was published Open Access, thanks to the support of the Netherlands Organisation for Scientific Research (NWO).

References

Adler S, Basketter D, Creton S, Pelkonen O, van Benthem J, Zuang V, Andersen KE, Angers-Loustau A, Aptula A, Bal-Price A, Benfenati E, Bernauer U, Bessems J, Bois FY, Boobis A, Brandon E, Bremer S, Broschard T, Casati S, Coecke S, Corvi R, Cronin M, Daston G, Dekant W, Felter S, Grignard E, Gundert-Remy U, Heinonen T, Kimber I, Kleinjans J, Komulainen H, Kreiling R, Kreysa J, Leite SB, Loizou G, Maxwell G, Mazzatorta P, Munn S, Pfuhler S, Phrakonkham P, Piersma A, Poth A, Prieto P, Repetto G, Rogiers V, Schoeters G, Schwarz M, Serafimova R, Tähti H, Testai E, van Delft J, van Loveren H, Vinken M, Worth A, Zaldivar JM. 2011. Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. Arch Toxicol 85:367–485.

Ahern BJ, Parviz J, Boston R, Schaer TP. 2009. Preclinical animal models in single site cartilage defect testing: a systematic review. Osteoarthr Cartilage 17:705–713.

Basketter DA, Clewell H, Kimber I, Rossi A, Blaauboer B, Burrier R, Daneshian M, Eskes C, Goldberg A, Hasiwa N, Hoffmann S, Jaworska J, Knudsen TB, Landsiedel R, Leist M, Locke P, Maxwell G, McKim J, McVey EA, Ouedraogo G, Patlewicz G, Pelkonen O, Roggen E, Rovida C, Ruhdel I, Schwarz M, Schepky A, Schoeters G, Skinner N, Trentz K, Turner M, Vanparys P, Yager J, Zurlo J, Hartung T. 2012. A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing - t4 report. ALTEX 29:3–91.

Besselink MG, van Santvoort HC, Buskens E, Boermeester MA, van Goor H, Timmerman HM, Nieuwenhuijs VB, Bollen TL, van Ramshorst B, Witteman BJ, Rosman C, Ploeg RJ, Brink MA, Schaapherder AF, Dejong CH, Wahab PJ, van Laarhoven CJ, van der Harst E, van Eijck CH, Cuesta MA, Akkermans LM, Gooszen HG; Dutch Acute Pancreatitis Study Group. 2008. Probiotic prophylaxis in predicted severe acute pancreatitis: a randomised, double-blind, placebo-controlled trial. Lancet 371:651–659.

Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, Macleod M, Dirnagl U. 2008. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. Stroke 39:929–934.

de Vries RB, Buma P, Leenaars M, Ritskes-Hoitinga M, Gordijn B. 2012. Reducing the number of laboratory animals used in tissue engineering research by restricting the variety of animal models. Articular cartilage tissue engineering as a case study. Tissue Eng Part B Rev 18:427–435.

de Vries RB, Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. 2011. A search filter for increasing the retrieval of animal studies in Embase. Lab Anim 45:268–270.

de Vries RB, Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. 2014. Updated version of the Embase search filter for animal studies. Lab Anim 48:88.

de Vries RB, Leenaars M, Tra J, Huijbregtse R, Bongers E, Jansen JA, Gordijn B, Ritskes-Hoitinga M. 2013. The potential of tissue engineering for developing alternatives to animal experiments: a systematic review. J Tissue Eng Regen Med doi 10.1002/term.1703.

Dickersin K. 1990. The existence of publication bias and risk factors for its occurrence. JAMA 263:1385–1389.

Egger M, Davey Smith G, Altman DG. 2001. Systematic reviews in health care. Meta-analysis in context. London: BMJ Publishing Group.

Festing MFW, Overend P, Gaines Das R, Cortina Borja M, Berdoy M. 2002. The Design of Animal Experiments. London: Laboratory Animals Ltd.

Hartung T. 2010. Evidence-based toxicology - the toolbox of validation for the 21st century? ALTEX 27:253–263.

Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. 2013. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. PLoS Med 10:e1001489.

Higgins JP, Green S. 2008. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: John Wiley & Sons Ltd.

Hirst TC, Vesterinen HM, Sena ES, Egan KJ, Macleod MR, Whittle IR. 2013. Systematic review and meta-analysis of temozolomide in animal models of glioma: was clinical efficacy predicted? Br J Cancer 108:64–71.

Hooijmans CR, de Vries RB, Rovers MM, Gooszen HG, Ritskes-Hoitinga M. 2012. The effects of probiotic supplementation on experimental acute pancreatitis: a systematic review and meta-analysis. PLoS One 7:e48811.

Hooijmans CR, Ritskes-Hoitinga M. 2013. Progress in using systematic reviews of animal studies to improve translational research. PLoS Med 10:e1001482.

Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. 2014. SYRCLE's risk of bias tool for animal studies. BMC Medical Research Methodology 14:43.

Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. 2010. Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. Lab Anim 44:170–175.

Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R. 2014. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 383:166–175.

Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG. 2009. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS One 4:e7824.

Kimmelman J, Anderson JA. 2012. Should preclinical studies be registered? Nat Biotechnol 30:488–489.

Korevaar DA, Hooft L, ter Riet G. 2011. Systematic reviews and meta-analyses of preclinical studies: publication bias in laboratory animal experiments. Lab Anim 45:225–230.

Krauth D, Woodruff TJ, Bero L. 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. Environ Health Perspect 121:985–992.

Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. 1992. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med 327:248–254.

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

Leenaars M, Ritskes-Hoitinga M, Griffin G, Ormandy E. 2012b. Background to the Montréal Declaration on the Synthesis of Evidence to Advance the 3Rs Principles in Science, as Adopted by the 8th World Congress on Alternatives and Animal Use in the Life Sciences, Montréal, Canada, on August 25, 2011. Altex Proceedings:35–38.

Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. 2008. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke 39:2824–2829.

McGonigle P, Ruggeri B. 2014. Animal models of human disease: challenges in enabling translation. Biochem Pharmacol 87:162–171.

Mignini LE, Khan KS. 2006. Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. BMC Med Res Methodol 6:10.

Mullen PD, Ramirez G. 2006. The promise and pitfalls of systematic reviews. Annu Rev Public Health 27:81–102.

O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH, Howells DW. 2006. 1,026 experimental treatments in acute stroke. Ann Neurol 59:467–477.

Percie du Sert N, Chapman K, Sena ES. 2012. 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. Poster 12th annual meeting Safety Pharmacology Society (Oct 2012, Phoenix AZ).

Percie du Sert N, Rudd JA, Apfel CC, Andrews PL. 2011. Cisplatin-induced emesis: systematic review and meta-analysis of the ferret model and the effects of 5-HT(3) receptor antagonists. Cancer Chemother Pharmacol 67:667–686.

Peters JL, Sutton AJ, Jones DR, Rushton L, Abrams KR. 2006. A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. J Environ Sci Health B 41:1245–1258.

Ritskes-Hoitinga M, Leenaars M, Avey M, Rovers M, Scholten R. 2014. Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. Cochrane Database Syst Rev 3:ED000078.

Roosen A, Woodhouse CR, Wood DN, Stief CG, McDougal WS, Gerharz EW. 2012. Animal models in urinary diversion. BJU Int 109:6–23.

Russell WMS, Burch RL. 1959. The Principles of Humane Experimental Technique. London: Methuen.

Selderslaghs IW, Blust R, Witters HE. 2012. Feasibility study of the zebrafish assay as an alternative method to screen for developmental toxicity and embryotoxicity using a training set of 27 compounds. Reprod Toxicol 33:142–154.

Sena E, Wheble P, Sandercock P, Macleod M. 2007. Systematic review and meta-analysis of the efficacy of tirilazad in experimental stroke. Stroke 38:388–394.

Sena ES, Briscoe CL, Howells DW, Donnan GA, Sandercock PA, Macleod MR. 2010a. Factors affecting the apparent efficacy and safety of tissue plasminogen activator in thrombotic occlusion models of stroke: systematic review and meta-analysis. J Cereb Blood Flow Metab 30:1905–1913.

Sena ES, Currie GL, McCann SK, Macleod MR, Howells DW. 2014. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. J Cereb Blood Flow Metab 34: 737–742.

Sena ES, van der Worp HB, Bath PM, Howells DW, Macleod MR. 2010b. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. PLoS Biol 8:e1000344.

Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, Hing C, Kwok CS, Pang C, Harvey I. 2010. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 14(8):iii, ix-xi, 1–193.

Stephens ML, Andersen M, Becker RA, Betts K, Boekelheide K, Carney E, Chapin R, Devlin D, Fitzpatrick S, Fowle JR 3rd, Harlow P, Hartung T, Hoffmann S, Holsapple M, Jacobs A, Judson R, Naidenko O, Pastoor T, Patlewicz G, Rowan A, Scherer R, Shaikh R, Simon T, Wolf D, Zurlo J. 2013. Evidence-based toxicology for the 21st century: opportunities and challenges. ALTEX 30:74–103.

ter Riet G, Korevaar DA, Leenaars M, Sterk PJ, Van Noorden CJ, Bouter LM, Lutter R, Elferink RP, Hooft L. 2012. Publication bias in laboratory animal research: a survey on magnitude, drivers, consequences and potential solutions. PLoS One 7:e43404.

van der Graaf AM, Wiegman MJ, Plosch T, Zeeman GG, van Buiten A, Henning RH, Buikema H, Faas MM. 2013. Endothelium-dependent relaxation and angiotensin II sensitivity in experimental preeclampsia. PLoS One 8:e79884.

van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. 2010a. Can animal models of disease reliably inform human studies? PLoS Med 7:e1000245.

van der Worp HB, Macleod MR, Kollmar R, European Stroke Research Network for H. 2010b. Therapeutic hypothermia for acute ischemic stroke: ready to start large randomized trials? J Cereb Blood Flow Metab 30:1079–1093.

van der Worp HB, Sena ES, Donnan GA, Howells DW, Macleod MR. 2007. Hypothermia in animal models of acute ischaemic stroke: a systematic review and meta-analysis. Brain 130:3063–3074.

van Drongelen J, Hooijmans CR, Lotgering FK, Smits P, Spaanderman ME. 2012. Adaptive changes of mesenteric arteries in pregnancy: a meta-analysis. Am J Physiol Heart Circ Physiol 303:H639–657.

van Luijk J, Bakker B, Rovers MM, Ritskes-Hoitinga M, de Vries RB, Leenaars M. 2014. Systematic reviews of animal studies; missing link in translational research? PLoS One 9:e89981.

Vesterinen HM, Sena ES, Egan KJ, Hirst TC, Churolov L, Currie GL, Antonic A, Howells DW, Macleod MR. 2014. Meta-analysis of data from animal studies: a practical guide. J Neurosci Methods 221:92–102.

Vesterinen HM, Sena ES, ffrench-Constant C, Williams A, Chandran S, Macleod MR. 2010. Improving the translational hit of experimental treatments in multiple sclerosis. Mult Scler 16:1044–1055.

Wever KE, Menting TP, Rovers M, van der Vliet JA, Rongen GA, Masereeuw R, Ritskes-Hoitinga M, Hooijmans CR, Warle M. 2012. Ischemic preconditioning in the animal kidney, a systematic review and meta-analysis. PLoS One 7:e32296.