Critical Appraisal of Studies Using Laboratory Animal Models

Annette M. O'Connor and Jan M. Sargeant

Abstract

In this manuscript we discuss an approach to critically appraising papers based on the results of laboratory animal experiments. The roles of external and internal validity in critically appraising the results of a paper are introduced. The risk of bias domains used by the Cochrane Handbook of Systematic Reviews of Interventions form the basis for assessing internal validity. The bias domains discussed include the selection bias, performance bias, outcome assessment bias, attrition bias, and reporting bias. Further, an approach to considering the role of chance in research findings is discussed.

Key Words: animal models; critical appraisal; preclinical research; bias; validity

Introduction

Preclinical assessment of interventions and a great deal of pathogenic mechanism research is conducted using animal models. Our understanding of the working of biological systems, such as the immune and cardiovascular systems, is often based upon the results of research on animals. Many pharmaceutical interventions in human health are first tested in animal studies for efficacy and safety. As such, animal research is considered critical to the scientific endeavor. However, how well animal studies achieve this goal is a subject of debate (van der Worp et al. 2010).

Although it frequently seems to researchers that publication is an endpoint, it is only an intermediary step in the scientific process (Sargeant and O'Connor 2013). When authors describe results, they often make the inference that the results from the study animals (study population) represent the results expected from the population the study animals came from (source population). Further, it is often inferred that the results can be applied to populations other than the source population (i.e., they can be applied to a target population). If this inference is not made, the results are not useful to end-users who wish to generalize the findings to reinforce concepts already known, generate new hypotheses, develop new research directions, or make policy based on the results.

What Is Critical Appraisal?

Critical appraisal is an essential part of the scientific process designed to assess the validity of scientific findings. The unit of assessment for critical appraisal is a single study, and the approach to assessing a single study result is present here. Critical appraisal is often conducted informally in a manner such that the rationale for a judgment is not clear. Here we seek to provide a transparent and systematic approach to critical appraisal. The approach presented here, in particular the use of risk domains, is borrowed from the Cochrane Handbook of Systematic Reviews of Interventions and translated to animal studies in preclinical medicine (Higgins et al. 2011).

Critical appraisal should be differentiated from an assessment of comprehensive reporting. Comprehensive reporting simply assesses if the study is reported in a manner that includes the important components of a study. There is substantial evidence that reporting of preclinical studies is less than comprehensive or has low reproducibility (Kilkenny et al. 2009; Prinz et al. 2011; Steward et al. 2012). Based on this evidence, there have been calls for improved reporting (Begley and Ellis 2012; Landis et al. 2012; van der Worp and Macleod 2011). Further, there are several guidelines outlining aspects of study design, analysis, and reporting that should be included in any research report; examples include the ARRIVE Guidelines (Kilkenny et al. 2010) and the "Guidance for the Description of Animal Research in Scientific Publications" (National Research Council [US] Institute for Laboratory Animal Research 2011). The target audience for these checklists or guidelines is generally authors, and the aim is to provide guidance on to how to present the study. Many, but not all, of the items in these checklists are related to enabling critical appraisal by the end-user, and for this reason sometimes these checklists are used for critical appraisal. This is inappropriate. Assessing comprehensive reporting requires an assessment of presence or absence of an item,

Annette M. O'Connor, BVSc, MVSc, DVSc, FANZCVS, is a veterinarian and Professor of Epidemiology at the Veterinary Medical Research Institute, College of Veterinary Medicine, Iowa State University, Ames, Iowa. Jan M. Sargeant, DVM, PhD, is a veterinarian and Professor of Epidemiology in the Department of Population Medicine and Director of the Centre for Public Health and Zoonoses at the Ontario Veterinary College, University of Guelph, Guelph, Ontario, Canada.

Address correspondence and reprint requests to Annette M. O'Connor, Veterinary Medical Research Institute, Bld 4, College of Veterinary Medicine, Iowa State University, Ames, IA 50010 or email oconnor@iastate.edu.

ILAR Journal, Volume 55, Number 3, doi: 10.1093/ilar/ilu038

whereas critical appraisal requires judgment about validity. For example, a study may report random allocation to treatment group and therefore have a "yes" for such an item on a checklist. However, despite random allocation to group, there may be important baseline differences in characteristics across treatments groups that suggest a high risk of bias.

Checklists for comprehensive reporting are generally made for larger areas of science and are based on a study design such as assessments of interventions and diagnostic test evaluations (Ell 2003; O'Connor et al. 2010; The Lancet 2010). These checklists can be used by authors across a variety of disciplines, because the standards for reporting across science are fairly standard. Critical appraisal, however, is more narrow and topic specific. For example, most scientists can determine if the study contains the list of outcomes assessed; however, only content experts can judge if the approach to measurement is valid and unlikely to introduce bias. This is why there are few "checklists" for critical appraisal, and those that exist (like this one) recommend tailoring the items to the topic. Although it is widely known what information is needed to assess bias, this is not synonymous with the presence of bias. The later is topic specific and requires judgment.

After critical appraisal of individual papers, the next step is to combine the results of many studies to reach a conclusion about a body of work. Formal transparent approaches to this combination are available but have infrequently been applied to laboratory animal studies and are not reviewed here (Guyatt et al. 2008, 2011; Johnson et al. 2014; Koustas et al. 2014; Lam et al. 2014; Woodruff and Sutton 2014).

Organization of the Remainder of This Paper

This remainder of this manuscript is organized as follows. First, we discuss the components of validity and introduce the terms external validity, internal validity, bias, and bias domains. We then discuss in greater detail the bias domains that affect internal validity that might affect animal experiments. For each domain of bias, we present a hypothetical example from an animal experiment publication that demonstrates an approach that would induce the bias. These examples are often extreme to make them easier to understand. Next, we provide a discussion of the rationale for why the risk domain should be assessed when critically appraising a study. We then discuss how to evaluate the role of random error when conducting a critical appraisal. Finally, we provide comments on how to reach a judgment about a paper based on the risk domains and assessment of internal validity.

Throughout the paper, we provide guidance on where in the manuscript a critical appraiser might expect to find the information needed to assess each bias domain. For this component, we make reference to "The ARRIVE Guidelines for Reporting Animal Research" (Kilkenny et al. 2010). Although we make reference to the ARRIVE Guidelines checklist items, end-users should be aware that not all authors report in a manner consistent with the ARRIVE Guidelines. There is good empirical evidence that reporting of laboratory animal studies is not comprehensive (Banwell et al. 2009; Crossley et al. 2008; Sena et al. 2007a, 2007b, 2010; van der Worp et al. 2007; Wheble et al. 2008). End-users should also be aware that even when authors do use the ARRIVE Guidelines (or similar lists), the authors are not obliged to follow the checklist item order. For example, there is no strict requirement that the number of study animals be reported in the first part of the results. Based on historical, journal, discipline, or author's preference, the methods and materials section often contains this information. The reporting of the information is what is important, not its order in the manuscript.

The Components of Critical Appraisal

If the results of studies are to be inferred to represent those from a source and target population, it is important to evaluate the external validity and internal validity of the study. External validity refers to the extent to which the results can be inferred to a target population. Internal validity refers to how the study results represent the source population. Critical appraisal of a scientific report can evaluate both concepts.

External Validity

The concept of external validity is a difficult one in laboratory animal studies, because the target population is unclear and varies by end-user. By the very nature of using animals as models for basic biological functions, there is the implication that the results are in some way generalizable to other populations, such as humans. However, often with laboratory animal studies, the next step in the scientific process is to develop a new hypothesis to be tested in other animals, in which case the target population of the laboratory animal study would be laboratory animal populations. Further, there is some empirical evidence in the literature suggesting that results from animal models do not generalize to "other" external populations of interest (Perel et al. 2007; Pound et al. 2004; van der Worp et al. 2010). In this paper we assume that the critical appraiser has already made the decision that the study results, if internally valid, will apply to a target population. We do not provide guidelines for assessing external validity other than to mention that information about external validity often relates to the population studied.

Internal Validity

The most common aspect considered in critical appraisal is internal validity. Internal validity relates to the question, "Are the results of the study population representative of the source population?" Threats to internal validity occur due to bias. For example, consider an experiment that looks at differences in behavior in mice subjected to 2 treatments (A and B). In the experiment, the outcome is a binary variable

("yes" or "no") that classifies each animal as having a characteristic of interest or not. There are 10 mice per treatment group. The data demonstrate that in treatment A, 8 of the 10 mice score a "yes," while in treatment B, only 4 of the 10 mice score a "yes." The comparison of the treatments can be summarized using the ratio of the proportions with "yes" outcomes in treatment A to treatment B (i.e., 8/10 ÷ 4/10 = 2). From these results, we would infer that treatment A doubles the risk of a "yes" response. It would also be reasonable to summarize the comparison of the treatments by calculating the difference in risk of the "yes" response (i.e., 80% – 40% = 40%). If a "yes" response is favorable (e.g., a "yes" outcome indicates better or improved cognitive function), treatment A may be considered a candidate therapy to be moved forward to the next stage of testing. If a "yes" response is not favorable (e.g., a "yes" outcome indicates lower or diminished cognitive function), treatment A may be considered a risk factor or potential cause of the behavior, and further research may evaluate the causal mechanisms. The question the critical appraiser must ask is, "Did the observed ratio or difference arise because of the treatments, systematic bias, or imprecision?" Alternatively, internal validity can be framed as the question, "What is the potential for factors other than the variable of interest to have influenced the results of the study?"

For the remainder of this paper, we will discuss a systematic approach to critically appraising the internal validity of study findings. We will discuss sources of systematic error (i.e., bias) as well as the impact of imprecision or random error on the results.

Bias

Bias refers to a systematic deviation from the true state of nature (i.e., the association observed in the study population differs from the true association in the source population). In comparative research, we aim to measure the effect of an intervention among the study groups (e.g., how average liver weight differs across 3 treatments). However, the observed difference in the study population may differ from the true result in the source population for 2 reasons: bias or imprecision. Imprecision relates to random error and is discussed at the end of the manuscript. Bias is caused by systematic errors. Systematic errors are a function of how the study was conducted and, unlike random error, cannot be resolved by increasing the sample size. For example, if researchers aim to compare the weights of 2 study groups, they may obtain an imprecise estimate of the difference in weights, because only 5 animals per group were studied. Simply increasing the sample size can decrease the extent of imprecision. However, if the researchers use a scale that is systematically incorrect (e.g., one that always underestimates weight), increasing the sample size will not decrease this systematic error.

The nomenclature of systematic errors is only moderately consistent across disciplines and may therefore create some confusion. In epidemiology, the categories of systematic error are usually referred to as selection bias, confounding bias, and information bias (Dohoo et al. 2010). However, in the clinical trial literature, the terminology used for categorizing systematic errors is selection bias, performance bias, detection bias, attrition bias, and reporting bias (Higgins et al. 2011). In trials, selection bias may create a confounding bias. In this manuscript, we will use the latter terms for consistency with the risk-of-bias tool developed by the Cochrane Collaboration (Higgins et al. 2011). The rationale for using the risk-of-bias domains recommended by the Cochrane Collaboration is that the biases associated with animal research are consistent with those associated with human health and animal experiments more closely represent controlled trials that observational studies. Further, there are communication advantages to using of a common terminology (Higgins et al. 2011). For example, failure to blind leads to the same concerns about outcome assessment bias in human-based research and animal research.

Risk-of-Bias Domains

Selection Bias

Mice in the study were allocated to treatment group by sex; treatment A consisted of female mice and treatment B consisted of male mice:

Example 1

Selection bias occurs when there are systematic differences in baseline characteristics between the treatment groups being compared (Higgins et al. 2011). In laboratory animal studies, selection bias occurs when nonrandom factors influence the allocation of animals to treatment groups (Starks et al. 2009). In example 1, it is possible that the sex of the group may account for the observed difference in the group outcomes, rather than the treatment. Although extreme, this is an example of selection bias.

When critically appraising a paper for evidence of selection bias, end-users should ask, "Are the groups comparable such that an observed difference is likely attributable to the treatment rather than a confounder?" A confounder is a factor that is related to the outcome in the source population independent of the treatments and related to the treatments in the study population. A common potential confounder in animal studies is sex. If the sex of animals is unevenly distributed across the treatment groups and sex has an effect on the outcome, then the observed difference in treatments may be simply a sex effect.

Ideally, we would like to measure the effect of an intervention by observing the effects of treatment A and treatment B at exactly the same time on exactly the same animals; this is referred to as the counterfactual. In this situation, the observed difference in the outcome would be attributable to only the difference in treatment. Of course, the counterfactual cannot be observed. In lieu of being able to observe the counterfactual, researchers try to create intervention groups that are exchangeable (Greenland and Robins 1986; Lindley and Novak 1981; Pearl 2009). For example, imagine a hypothetical exchange of the 2 treatment groups; if the treated group becomes untreated, and vice versa, and the effect of the treatment remains the

same, then the groups are exchangeable. Exchangeability will arise if there is balance across the groups for all the factors that may affect the outcome, so that we achieve close to the counterfactual comparison. Any difference in the outcome observed between the treatment groups can be attributed to the difference in treatment. Therefore, when critically appraising a study, end-users should look for the use of design tools that would create exchangeable groups.

Randomization is the best known of the design tools to balance the distribution of confounders across the intervention groups. Randomization of study subjects to intervention groups means that the differences observed in the makeup of groups are due to chance ("random error"). Randomization is the only design tool that can address known and unknown confounders. As such, it is an important tool and should be used in all studies. Numerous studies have demonstrated that the results of randomized and nonrandomized studies differ, suggesting randomization reduces bias (Burns and O'Connor 2008; Jerndal et al. 2010; Pedder et al. 2014; Sargeant et al. 2009; Schulz et al. 1995). Thus, randomization serves as an indicator of a reduced risk of bias. However, randomization does not guarantee the balancing of confounders and, therefore, does not guarantee the absence of selection bias as a cause of observed group differences. The reporting of randomization should not be treated as synonymous with the absence of selection bias.

The ability of randomization to balance both known and unknown confounders decreases as the study becomes smaller. This is an important point. Because of the importance of sample size in the efficacy of randomization, even in randomized studies, end-users need to consider the size of the groups to determine the potential for randomization to balance known and unknown confounders. By and large, laboratory animal studies are very small, so this is relevant to critical appraisal of laboratory animal studies. For example, consider an experiment with 20 animals—10 males and 10 females. If sex is associated with the outcome in the source population, the researcher would want to ensure that there was an even number of males and females in each group to balance the sex effect. If the researchers relied solely on simple random allocation, it is possible that the sex of the animals would not be evenly distributed across the groups. This concept can be illustrated with the idea of a coin toss—if a coin is tossed 20 times, 10 heads would not be guaranteed despite the use of a random method. As shown in Figure 1, the most common result is to have 10 heads; however, other results are possible. In Figure 2, it can be seen that the potential for uneven groups increases with decreasing sample size. For example, when 6 coin tosses are made, it is quite common for only 1 head to occur (around 10% of the time), rather than the expected 3 heads from 6 tosses. As the sample size increases, the probability of extreme unbalance become rarer, and the 50% heads result is observed with greater frequency. The same concept can be applied to small experiments, which laboratory animal studies often are.

Given the potential for study size to affect the efficacy of simple randomization to control selection bias, critical

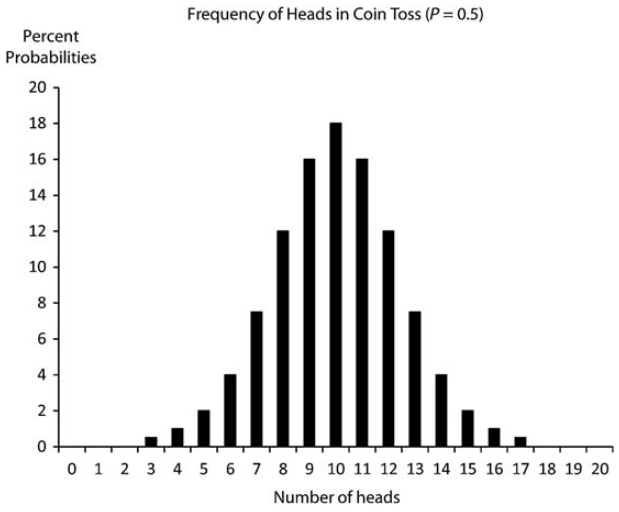

Figure 1 Sampling distribution of number of head from repeated tosses of 20 fair coins.

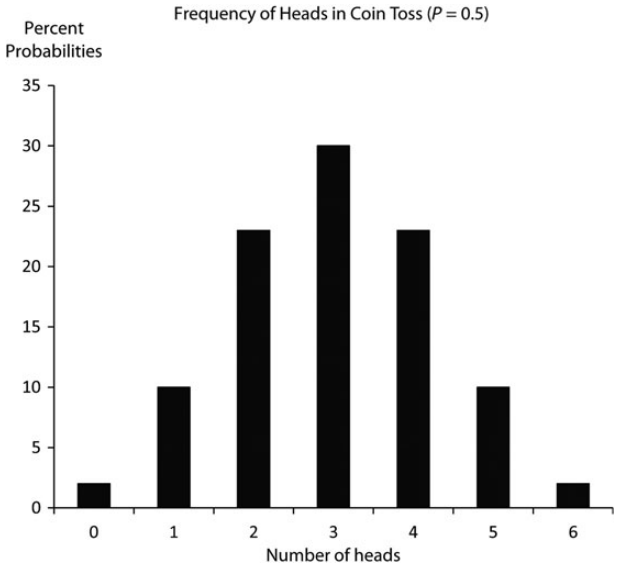

Figure 2 Sampling distribution of number of head from repeated tosses of 6 fair coins.

appraisers should look for the use of more complex approaches to random allocation. The use of such tools may increase the confidence that the risk of selection bias is low. It is possible to employ restricted randomization to deliberately force the even distribution of known confounders and increase the potential for exchangeability. Block and stratified randomization can be used, although the researcher may not use these exact terms, which are more common in clinical trials. In stratified randomization, the researchers could stratify animals by sex, and randomly allocate within each sex, ensuring that sex was evenly distributed across treatment groups. If a continuous variable is a potential confounder, the researchers may use block randomization. For example, 20 piglets might be organized by weight from heaviest to lightest, creating

5 blocks of 4 animals each. Then all possible allocation sequences of 2 treatments (labeled A and B) for a group of 4 can be created (i.e., AABB, ABBA, ABAB, BABA, BBAA, BAAB) (Altman and Bland 1999). The researcher may randomly allocate a sequence to each block. Such an approach aims to balance the distribution of weight across the treatment groups and reduce selection bias and confounding.

Other designs tools also exist to distribute confounders across intervention groups. The most commonly used design tool in laboratory animal studies is restriction of the study population to exclude a factor that is known to or would likely affect the outcome. In our example, restriction would correspond to only conducting the experiment in male mice or only conducting the study in female mice. Restriction is such a fundamental approach to controlling for confounding in laboratory animal studies that many researchers may not recognize it as a design tool. Laboratory animal studies routinely are restricted to a source population of the same age and genetic lines, and by doing so, important known (or suspected) confounders are removed. It is also important to recognize that the use of restriction as a design tool means that variation is often lower in laboratory animal studies compared with human randomized controlled trials that are often designed to include a diverse group of people. Because of this lower variation, restriction may in part explain why large sample sizes are often not needed in laboratory animal studies to observe differences. Unfortunately, restriction also has the impact of reducing external validity. Of note, the National Institutes of Health recently have been discouraging restriction of studies to one sex of animal (Clayton and Collins 2014).

Finally, even in studies that employ all of the tools recommended to increase the exchangeability of groups and reduce selection bias, groups may still be uneven. This should be evaluated by looking at the baseline characteristics of the groups, frequently reported in the initial portion of the results section.

Where Might We Expect Information About Selection Bias to Be in the Report?

When assessing the potential for selection bias, the critical appraiser needs to look at the eligibility criteria for the study subjects, the approach to allocation of study subjects to treatment groups, the sample size, and the baseline information for study subjects in each of the treatment groups. In the ARRIVE Guidelines, the information that would enable the end-user to judge the balance of known factors between the groups is described in Items 6, 8, 10, and 11. The information requested in these items will enable the end-user to know the study population, housing, and demographics and the baseline data of animals in each treatment group.

Performance Bias

> *Mice in the infected group were housed in the newer B3 isolation facility in individual cages. However, as there was no biological need and to reduce costs, the untreated control mice were housed in two group cages in the older Bl2 facility:*
>
> *Example 2*

Performance bias occurs when there are systematic differences between groups in the care that is provided, or in exposure to factors other than the treatment (Higgins et al. 2011). When critically appraising papers of laboratory animal studies, end-users should ask, "Was the approach to husbandry the same for all treatment groups and was caregiving done without knowledge of the treatment group?" If the answer is no, then the potential for performance bias may be high.

In example 2, it is clear that the management of the animals is different between the treatment groups and this could influence the performance of the animals. In such a scenario, the potential is high for differences in treatment group outcomes to be affected by the different approaches to housing. Obviously, there are more subtle ways in which animals may be managed that could still contribute to differences in the performance of the groups. In animal studies, performance bias may arise in many ways. First, as in the example, the animals may be managed differently by design. Location in facilities, group sizes, and diet are all factors that should be the same across treatment groups. Secondly, the animals may be managed differently by nonblinded caregivers. Blinding of the caregiver is important in all animal studies. Because there is a daily interaction between caregivers and animals, it is important that caregivers are not aware of the treatment groups, so they do not inadvertently manage one group differently. Blinding of caregivers to the treatment group should not be confused with blinding of outcome assessment, which is related to the potential for detection bias (see below). Blinding of the caregiver to treatment group can be difficult for studies that use only 2 cages (i.e., one for each treatment). Such studies theoretically have a high potential for performance bias, if a caregiver feels they "know" the treatment allocation. This is an argument for having more than one cage per treatment.

Where Might We Expect Information About Performance Bias to Be in the Report?

To assess performance bias, the critical appraiser needs to know about the management of the animals, the outcomes of interest, and the approach to blinding of caregivers as to which treatment groups the animals (or animal cages) belong. In the ARRIVE Guidelines, the information that would enable the end-user to judge performance bias is described in Items 6, 9, and 13.

Detection Bias

> *The cages of the mice in the group treated with the new fantastic drug were labeled with the term "fantastic drug" and the cages of the mice that did not received the treatment were labeled "old drug" Each day the staff feeding the mice were asked to described how active the*

mice where on a scale of 1 to 5, with 1 being not active and 5 being very active:

Example 3

Detection bias can occur when there are systematic differences between treatment groups in the way in which the outcome is assessed (Higgins et al. 2011). When critically appraising a paper, end-users should ask, "Was the approach to assessing the outcomes the same in both groups and done without knowledge of the group?" If the answer is no, the potential for detection bias may be high.

In example 3, it is clear that the people measuring the outcome are also aware of the treatment groups of the mice. The caregivers may have the expectation that the mice receiving the new treatment will perform better and inadvertently overestimate the activity levels of these mice. The observed differences in the outcomes between groups could be due to differential measurement of the outcome between the 2 groups rather than treatment. This approach clearly has the potential to introduce detection bias.

The most common tool used to prevent detection bias is blinding of the individuals assessing the outcome as to which animals (or cages) belong to which treatment groups. Empirical evidence shows that failure to blind the outcome assessor in laboratory animal studies is associated with more favorable outcomes (Jerndal et al. 2010; Minnerup et al. 2010). Ideally, all studies should blind outcome assessment; however, the risk of bias due to failure to blind may be high or low depending upon the outcome. For example, when the outcome is death, the risk of bias due to lack of blinding may be very low, as the outcome is extremely objective. Other outcomes that may seem objective and quantifiable, such as cells per field, may be prone to bias if the reader is allowed to pick a "suitable slide field." Risk of bias should be judged for every outcome reported and the judgment reached may differ among different outcomes within the same experiment. Assessment of the risk of detection bias for a particular outcome often requires content expertise.

When blinding is used, it is important not only that the outcome assessor not know what specific treatment an animal has received, but also what treatment group an animal is in (e.g., group labeled "A" or group labeled "B"), even if specific treatments are not named. This is important, because knowing the outcome for one animal of a group may influence the interpretation of the outcome for the next animal in the same group. Even if a bias does not exist towards a particular outcome, knowing the treatment group may inadvertently reduce group-level variation. For example, imagine the scenario where the first animal is known to have received treatment A and has severe lung consolidation. The pathologist may be more attuned to lung congestion in the next animal in group A. Similarly, the pathologist may be more inclined to call the pathology severe, if the last animal in group A was known to have severe pathology. This bias would make groups more consistent, reducing heterogeneity, and may decrease the *p* value as the true extent of variation is understated. Thus, bias can be introduced, even without knowing exactly what treatment A is.

Where Might We Expect Information About Performance Bias to Be in the Report?

To assess detection bias, the critical appraiser needs to know about the management of the animals, including any approaches to identification, the measurement of the outcomes of interest, and the approach to blinding of the outcome assessor. In the ARRIVE guidelines, the information that would enable the end-user to determine the likelihood that the outcome was differentially assessed is described in Items 6, 9, and 13.

Attrition Bias

"Twenty-four animals were allocated to 2 treatment groups (12 in each group). The average weight (±SD) of the mice in treatment A was 2.5 g ± 0.3 (n = 7) and the average weight in treatment B was 2.4 g ± 0.3 (n = 12)."

Example 4

Attrition bias refers to systematic differences in withdrawals from treatment groups (Higgins et al. 2011). When critically appraising a paper, end-users should ask, "Was the loss of animals from the groups minimal and unrelated to the treatment groups?" If the answer is no, the potential for bias due to attrition may be high. The impact of attrition is that the outcome is not observed for all animals. In the example above, it appears that some animals have not completed the study. If the outcome of those missing animals was systematically different across the groups, then bias can occur. For example, imagine a study that measured weight gain over a 50-day period as an outcome, as in example 4. The study also had weight loss as a withdrawal criterion (e.g., if an animal loses a certain amount of weight they must be withdrawn from the study and euthanized). Further, consider the impact of these criteria, if the effect of treatment A is to create a bimodal distribution of diseases (i.e., animals within the group are either severely affected or show no clinical signs), and if the effect of treatment B is to create a wide spectrum of disease, from severe to mild to no clinical signs. In such a scenario, severely affected animals in treatment A are more likely to be removed from the study, leaving only healthy animals. The unobserved animals would have had very low weight gain and, if included, would have reduced the treatment A average weight gain. However, because only the healthy animals are actually observed to the end of the study, the average weight gain observed in treatment A is higher than the true unobserved average, and the observed difference in group averages is due to attrition bias rather than a true treatment effect.

Attrition bias is a major issue in human studies where attrition may be voluntary (participants choose to leave the study) or nonvoluntary (e.g., participants die prior to the end of the

study). In human clinical trials, participants can leave the study if they choose. When they leave, the reason for leaving often is not known. If the reason for leaving is associated with the treatment, a bias can occur. This almost certainly cannot happen in laboratory animal studies. Therefore the potential for voluntary attrition bias in laboratory animal studies is almost nonexistent; however, nonvoluntary attrition can occur if animals die or are withdrawn because they met *a priori* criteria for removal prior to observing the outcome. Because nonvoluntary attrition can inadvertently introduce bias, authors are encouraged to report information about attrition for each study group. Authors should also consider including outcomes that will be unaffected by attrition; in the hypothetical example about weight loss, comparing survival time in each group would not be affected by attrition bias. However, often outcomes cannot be designed to address attrition bias and other approaches must be employed. Some design or statistical approaches to minimizing the impact of attrition can be used, such as using the last observed outcome or imputing the missing data. If these approaches are reported in a manuscript, the assessment of their validity often requires a statistician be consulted.

Where Might We Expect Information About Performance Bias to Be in the Report?

The ARRIVE Guidelines propose that authors report the number of animals enrolled in each group, the number of animals completing the study, and the number included in the analysis. It is also possible that authors will report their approach to dealing with missing data including any imputation approaches in the statistical analysis section of the manuscript (ARRIVE checklist items 13, 14, and 15).

Reporting Bias

> *"Outcomes measured were weight of the liver, kidney, brain, spleen and lungs. ...*
>
> *Results: The average weight of the kidneys in treatment A was 1.5 g (SD = 0.5, n = 12) and the average weight in treatment B was 0.9 g (SD = 0.4, n = 12) (mean difference = 0.6, 95% CI = 0.2 to 0.98, p value = 0.003867). The conclusion reached was ..."*
>
> *Example 5*

In example 5, we see that although the authors reported assessing 5 outcomes, the results of only 1 are reported. This is an example of incomplete reporting. When critically appraising a paper, end-users should ask, "Were the results of all outcome variables assessed reported completely?" If the answer is no, the potential for reporting bias may be high. Often incomplete reporting can only be theoretically assessed. If the authors had omitted reference to the lungs in the methods and materials, the reader would have no evidence about the inclusivity of the outcomes reported. Reporting bias refers to the absence of important results from the study (Higgins et al. 2011). The concept of reporting bias is sometimes difficult to conceptualize, when critically appraising a single study. It is likely more accurate to say that studies fail to report some results and that the actual bias occurs in subsequent reviews that use studies that fail to report all results. The failure to report something makes it impossible to ask the question, "Are there other explanations for the results observed?" Reporting bias is relevant when the aim of the critical appraisal exercise is to combine and summarize a larger body of work. If there is evidence of reporting bias and some results are missing, the results of the summary will be biased because of the absence of some studies. At the individual paper level, reporting bias may hide multiplicity issues and, as such, affect the ability of the critical appraiser to know the role of chance in the findings (see the discussion of multiplicity below).

Where Might We Expect Information About Reporting Bias to Be in the Report?

As shown in example 5, evidence of reporting bias is often found by comparing the outcomes reported (ARRIVE checklist item 16) and the statistical methods proposed (ARRIVE checklist item 13) and the variables assessed (ARRIVE checklist item 12). It is also possible to compare the outcomes studied in the report with those proposed in the original study proposal or protocol, if available. Finally, some journals now suggest the inclusion of a statement of comprehensive reporting (Altman and Moher 2013), and the inclusion of such a statement should suggest that the risk of reporting bias is low. An example of such a statement modified for animal studies might be, "The authors affirm that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the review as planned have been documented and explained. The authors have indicated where results from these study animals are reported in other publications, and have included citations for these publications where relevant."

Random Error

> *"Mortality in treatment A was 90% (9 of 10) and 10% in treatment B (1 of 10) (Fisher's exact p-value = 0.001093)."*
>
> *Example 6*

> *"Mortality in treatment A was 90% (9 of 10) and 10% in treatment B (1 of 10) (Fisher's exact p-value < 0.05)."*
>
> *Example 7*

Apart from systematic sources of error, random error may be another explanation for observed differences between group outcomes. When critically appraising a paper, end-users should ask, "Is there a low probability that chance played a role in the observed difference?" The end-user should use the *p* value provided by the original study to assess this

question and should report the p value threshold used by the end-user as the basis for considering an outcome rare enough to reject the null hypothesis. A small p value suggests a low probability of chance, a large p value suggests a high probability of chance. It is known from sampling distribution theory that if we were to obtain multiple random samples from a subset of individuals from the same population and measure a population parameter (e.g., mean weight), we would not obtain identical means for each sample, even though each sample came from the same source population. "Random error" means that the difference observed is simply a function of the random sampling of 2 groups of individuals from the same population. The p value of a test statistic under the null hypothesis is the approach used to express the probability that the observed difference in the outcomes between groups is a function of random error. For the data provided above, after performing a Fisher's exact test to assess treatment group differences, we obtain a p value of 0.00109. A chi-square test would not be appropriate for this dataset, because the sample size is too small. This p value suggests that 0.109% of the time with repeated samplings, we would expect to obtain the same (or greater) observed difference in proportions in mortality between 2 groups of 10 randomly drawn individuals from the same population. In example 6, we would conclude that the chances are small, because the p value suggests that the observed difference of 90% between groups would be observed only rarely.

Failure of authors to report the exact p value prevents a critical appraiser from evaluating the probability that random error is the reason for the observed treatment effect. Often authors simply report that the finding was "statistically significant," meaning that the probability that the observed difference arose from random error is ≤5% (example 7). The ability to assess the role of random error would be substantially improved if one could differentiate between a p value of 0.05 or 0.0001. For example, a p value of 0.05 implies that something occurs 1 in 20 occasions. In reality, this is not a very rare outcome, when we consider that it is quite common to roll 2 sixes with a pair of dice, an event with a probability of 1 in 36. The magnitude of the p value should not be used to interpret the size of the treatment effect (i.e., a small p value does to equate to a large difference, just a rare difference). This is a common misinterpretation. Indeed, in laboratory animal studies, most treatment effects are quite large, as the studies are small and only have the statistical power to detect large differences. One issue to be aware of is that within a test such as an ANOVA it is possible to adjust for the p value of pairwise comparison using statistical methods. Examples of approaches include calculation of the least significant difference, a Bonferonni adjustment, a Tukey adjustment , Tukey-Kramer adjustment, and numerous possible approaches (Ramsey and Schafer 2013). However, such approaches only adjust the p value for pairwise comparisons conducted within the ANOVA test and do not account for the conduct of multiple ANOVA tests.

A 95% confidence interval conveys to the end-user how precisely the estimate of the treatment effect is known. However, it can also be used to provide indirect information about random error. If the 95% confidence interval includes the null value (1 for ratios and 0 for differences), then the p value is >0.05. For example, if a study reports a risk ratio of 2 and the 95% confidence interval for the ratio is from 0.386 to 2.44, the p value must be >0.05. However, sometimes it is more difficult to determine the exact p value from a confidence interval, because the degrees of freedom are not reported, and in those situations, the exact role of chance might only be inferred rather than exactly calculated.

Multiple Testing

> *"Statistical Methods. The study had 3 treatment groups, and we measured wet weights, dry weights, and surface area for the liver, spleen, kidney, heart, lungs, stomach, and brain. We compared the average outcomes across all possible pairwise comparisons of the 3 groups (i.e., group 1 vs. group 2, group 2 vs. group 3, and group 1 vs. 3) using t-tests. We used a p value of 0.05 to assess statistical significance.*
>
> *Results. The wet weights, dry weights, and surface area of liver, spleen, kidney, heart, lungs, stomach, and brain were not statistically different across any of the pairwise treatment group comparisons with the exception of a statistically significant difference in the surface area of the spleen between Group 1 and Group 3 ($p < 0.05$)."*
>
> *Example 8*

Another factor to evaluate related to random error is multiplicity. It is common, by convention, to accept that a comparison is "statistically significant" if the p value is < 0.05. The 0.05 corresponds to a type I error (i.e., the probability that the observed difference or larger would occur in a population where the means were the same is 5% or less). However, when multiple comparisons are conducted within the same experiment, although the probability of a type I error for each comparison is 5%, the probability that at least one comparison was significant due to random error is considerably higher.

Authors should consider when multiple outcomes are assessed that differences may have arisen by chance. As well, across the entire study, the more tests conducted, the more opportunity there is for random error. In human clinical trials, studies are often designed to assess a limited number of outcomes, reducing the potential for this multiplicity of random errors. However, it is unclear if limiting the number of outcomes in laboratory animal studies is desirable. Laboratory animal studies are meant to provide preliminary proof of concept or data on treatment safety and thus play a critical role in discovery and hypothesis generation (Henderson et al. 2013). They also often end in euthanasia of the animals. Thus, a large number of outcomes may be justified. For discovery studies, we believe that it is preferable that authors report all outcomes assessed with group-level and summary-level data (including measures of precision) and authors consider not conducting inferential statistics on all outcomes (e.g., consider testing only those outcomes for which the study

Table 1 Sample form that might be used to document the approach to critical appraisal of a laboratory animal study designed to compare outcomes among groups

Area to assess	Question to ask	Possible responses	Design tools associated with control or reduction of risk	ARRIVE guideline items
External validity	"If the study was conducted in a manner that suggests little internal bias, will it be useful for the 'next step' because the population is relevant to 'the next step?'"	No—don't assess Yes—continue to assess internal validity	Inclusion criteria for relevant populations, housing, and intervention used	7, 8, and 9
Internal validity (using risk-of-bias domains from Higgins et al. 2011)				
Selection bias	"Are the groups comparable such that an observed difference is likely attributable to the treatment rather than a confounder?"	Yes—low ROB Unclear—unclear ROB No—high ROB	Blinded allocation to group, restriction, randomization, restricted randomization	6, 8, 10, and 11
Performance bias	"Was the approach to husbandry the same for all treatment groups and was caregiving done without knowledge of the treatment group?"	Yes—low ROB Unclear—unclear ROB No—high ROB	Blinding of caregivers, use of multiple cages per treatment	6, 9, and 13
Detection bias	"Was the approach to assessing the outcomes the same in both groups and done without knowledge of the group?"	Yes—low ROB Unclear—unclear ROB No—high ROB	Blinding of outcome assessors, use of repeatable and objective outcome measures	6, 9, and 13
Attrition bias	"Was the loss of animals from the groups minimal and unrelated to the treatment groups?"	Yes—low ROB Unclear—unclear ROB No—high ROB	Minimization of loss to follow-up and complete reporting of loss to follow-up for each treatment group	13 to 15
Reporting bias	"Were the results of all outcome variables assessed reported completely?"	Yes—low ROB Unclear—unclear ROB No—high ROB	Comprehensive reporting and a well-designed study protocol	12 to 17
Random error				
Test-level error	"Is there a low probability that chance played a role in the observed difference?"	Yes—low risk of random error in the test Unclear—unclear risk of random error in the test No—high risk of random error in the test	The exact p-value and the 95% confidence interval	10, 13, and 16

Continued

Table 1 Continued

Area to assess	Question to ask	Possible responses	Design tools associated with control or reduction of risk	ARRIVE guideline items
Study-level error	"Did the authors limit the number of hypothesis tests conducted to those the study was designed (powered) to assess?"	Yes—low risk of random error across the tests Unclear—unclear risk of random error across the tests No—high risk of random error across the tests	The power of the study, the number of tests conducted	10, 13, and 16
Conclusion	Based on the overall assessment of the above items what is your assessment of internal validity?	High internal validity Low internal validity Unclear internal validity		Items above

Abbreviation: ROB, risk of bias.

was specifically designed—i.e., the basis for the power calculation). For the critical appraiser, if reporting suggests a large number of outcomes were tested, the role of random chance in the study increases. This is the area where reporting bias affects the critical appraiser. If the authors do not report all outcomes assessed, the person appraising the study cannot accurately gauge the role of random error due to multiplicity in the study results. For example, if a researcher tests and publishes 20 outcomes and only one is statistically significant, a mindful reader could recognize the likelihood of false statistical significance and adjust for this. However, if instead the researcher publishes the 1 significant outcome with only 4 nonsignificant outcomes, 20% of outcomes appear significant, and the potential for the reader to recognize the random nature of the significant findings is eliminated.

How to Reach a Conclusion about a Paper

We propose that for individual papers, critical appraisers first assess external validity (i.e., if the study is judged to be internally valid, then the result would be useful to apply to the next step in the scientific process). Only then assess internal validity. To assess internal validity, for each of the risk domains, we would propose following the approach proposed by the Cochrane Collaboration (Higgins et al. 2011). For each domain, assess whether the risk of bias is high, low, or unclear. The unclear response usually arises if there is a lack of comprehensive reporting. If reporting is comprehensive, a judgment of high or low risk should be made and the rationale indicated. After assessing each bias domain and the potential for random error, evaluate the findings together. The final judgment can be recorded as valid or invalid. We present a template table (Table 1) that might be used for a critical appraisal exercise. We also present an example of a completed table (Table 2) to illustrate how the rationale can be included to enhance transparency.

Tempting as it may be to create a score or a cut-off, this approach should not be used (Higgins et al. 2011). Scores and cut-points have been discredited in favor of making a judgment about the validity of study findings (likely valid or likely invalid) and using risk-of-bias assessments to support that judgment. This avoids subjective selection of weightings for different bias domains and also allows the critical appraiser to judge the risk of bias for each domain in the specific context of the study or outcome. It is not unreasonable for the reviewer to decide that on the basis of a single flaw that the results are likely not internally valid.

We believe that recording the reasoning behind an assessment will be helpful for defending decisions about validity. Decisions about validity of results have been made for a very long time and knowledge of risk-of-bias domains has long been known. The 2 novel aspects to critical appraisal incorporated here are the formal partitioning of the risk-of-bias domains and the transparent documentation of the conclusions reached about these domains.

Table 2 Completed critical appraisal form for a hypothetical study using examples in the text

Area to assess	Question to ask	Possible responses	Rationale
External validity	"If the study was conducted in a manner that suggests little internal bias, will it be useful for the 'next step' because the population is relevant to 'the next step?'"	Yes	Population, housing, and intervention are relevant to next step in the scientific process.
Internal validity			
Selection bias	"Are the groups comparable such that an observed difference is likely attributable to the treatment rather than a confounder?"	No—high ROB	Example 1: Animals are assigned by sex and sex may be a confounder, if related to the outcome.
Performance bias	"Was the approach to husbandry the same for all treatment groups and was caregiving done without knowledge of the treatment group?"	No—high ROB	Example 2: The treatment groups are housed very differently and, therefore, the housing rather than the treatment could be the cause of the observed differences.
Detection bias	"Was the approach to assessing the outcomes the same in both groups and done without knowledge of the group?"	No—high ROB	Example 3: The outcome assessors are clearly aware of the treatment assignment and the approach to measurement is highly subjective, so these factors could be the cause of the observed differences between groups rather than the treatment.
Attrition bias	"Was the loss of animals from the groups minimal and unrelated to the treatment groups?"	No—high ROB	Example 4: Some animals are missing from one group and no explanation is provided.
Reporting bias	"Were the results of all outcome variables assessed reported completely?"	Unclear—unclear ROB	Example 5: Several measured outcome are not reported; this suggests incomplete reporting and may indicate a non-significant finding.
Random error			
Test level	"Is there a low probability that chance played a role in the observed difference?"	Yes—low risk of random error	Example 6: The p-value is very small suggesting the observed difference is rare under the null hypothesis.
Study level	"Did the authors limit the number of hypothesis tests conducted to those the study was designed (powered) to assess?"	No—high risk of random error	Example 8: The authors appear to have conducted $9 \times 7 = 63$ hypothesis tests without adjustment for multiplicity, and only found one significant outcome. It is unclear if this is a primary (important) outcome.
Conclusion		Low internal validity	

Abbreviation: ROB, risk of bias.

Conclusion

The critical appraisal of studies is a fundamental aspect of the scientific process. Here we propose applying the "risk-of-bias" domains used by the Cochrane Collaboration to the critical appraisal of laboratory animal studies. Critical appraisal requires content expertise and making judgments cannot be avoided. However, transparency in the criteria assessed and conclusions reached can make the process of critical appraisal easier to conduct, more deliberate, and perhaps more reproducible.

References

Altman DG, Bland JM. 1999. How to randomise. BMJ 319:703–704.

Altman DG, Moher D. 2013. Declaration of transparency for each research article. BMJ 347:f4796.

Banwell V, Sena ES, Macleod MR. 2009. Systematic review and stratified meta-analysis of the efficacy of interleukin-1 receptor antagonist in animal models of stroke. J Stroke Cerebrovasc Dis 18:269–276.

Begley CG, Ellis LM. 2012. Drug development: Raise standards for preclinical cancer research. Nature 483:531–533.

Burns MJ, O'Connor AM. 2008. Assessment of methodological quality and sources of variation in the magnitude of vaccine efficacy: A systematic review of studies from 1960 to 2005 reporting immunization with Moraxella bovis vaccines in young cattle. Vaccine 26:144–152.

Clayton JA, Collins FS. 2014. Policy: NIH to balance sex in cell and animal studies. Nature 509:282–283.

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.

Dohoo IR, Martin W, Stryhn H. 2010. Veterinary Epidemiologic Research. 2nd edition. VER Inc, Prince Edward Island, Canada.

Ell PJ. 2003. STARD and CONSORT: Time for reflection. Eur J Nucl Med Mol Imaging 30:803–804.

Greenland S, Robins JM. 1986. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15:413–419.

Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schünemann HJ, Grade Working Group. 2008. Going from evidence to recommendations. Br Med J 336:1049–1051.

Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. 2011. GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 64:380–382.

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 J, Altman D, Sterne J. 2011. Chapter 8: Assessing risk of bias in included studies. In: Higgins J, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions. Available online (http://www.cochrane-handbook.org), accessed June 25, 2014.

Jerndal M, Forsberg K, Sena ES, Macleod MR, O'Collins VE, Linden T, Nilsson M, Howells DW. 2010. A systematic review and meta-analysis of erythropoietin in experimental stroke. J Cereb Blood Flow Metab 30:961–968.

Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. 2014. The Navigation Guide—evidence-based medicine meets environmental health: Systematic review of human evidence for PFOA effects on fetal growth. Environ Health Perspect 122:1028–1039.

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. PLoS Biol 8:e1000412.

Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, 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.

Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. 2014. The Navigation Guide—evidence-based medicine meets environmental health: Systematic review of nonhuman evidence for PFOA effects on fetal growth. Environ Health Perspect 122:1015–1027.

Lam J, Koustas E, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. 2014. The Navigation Guide—evidence-based medicine meets environmental health: Integration of animal and human evidence for PFOA effects on fetal growth. Environ Health Perspect 122:1040–1051.

Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, et al. 2012. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490:187–191.

Lindley DV, Novak MR. 1981. The role of exchangeability in inference. Ann Statist 9:45–58.

Minnerup J, Wersching H, Diederich K. 2010. Methodological quality of preclinical stroke studies is not required for publication in high-impact journals. J Cereb Blood Flow Metab 30:1619–1624.

National Research Council (US) Institute for Laboratory Animal Research. 2011. Guidance for the Description of Animal Research in Scientific Publications. Available online (http://www.ncbi.nlm.nih.gov/books/NBK84205/), accessed June 25, 2014.

O'Connor AM, Sargeant JM, Gardner IA, Dickson JS, Torrence ME, Consensus Meeting P, Dewey CE, Dohoo IR, Evans RB, Gray JT, et al. 2010. The REFLECT statement: Methods and processes of creating reporting guidelines for randomized controlled trials for livestock and food safety by modifying the CONSORT statement. Zoonoses Public Health 57:95–104.

Pearl J. 2009. Causality. Cambridge: Cambridge University Press. p. 285.

Pedder H, Vesterinen HM, Macleod MR, Wardlaw JM. 2014. Systematic review and meta-analysis of interventions tested in animal models of lacunar stroke. Stroke 45:563–570.

Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, Macleod M, Mignini LE, Jayaram P, Khan KS. 2007. Comparison of treatment effects between animal experiments and clinical trials: Systematic review. BMJ 334:197.

Pound P, Ebrahim S, Sandercock P, Bracken MB, Roberts I. 2004. Where is the evidence that animal research benefits humans? BMJ 328:514–517.

Prinz F, Schlange T, Asadullah K. 2011. Believe it or not: How much can we rely on published data on potential drug targets? Nat Rev Drug Discov 10:712.

Ramsey FL, Schafer DW. 2013. The statistical sleuth: A course in methods of data analysis. Boston: Brooks/Cole, Cengage Learning.

Sargeant JM, O'Connor AM. 2013. Issues of reporting in observational studies in veterinary medicine. Prev Vet Med 113:323–330.

Sargeant JM, Saint-Onge J, Valcour J, Thompson A, Elgie R, Snedeker K, Marcynuk P. 2009. Quality of reporting in clinical trials of preharvest food safety interventions and associations with treatment effect. Foodborne Pathog Dis 6:989–999.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273:408–412.

Sena E, van der Worp HB, Howells D, Macleod M. 2007a. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci 30:433–439.

Sena E, Wheble P, Sandercock P, Macleod M. 2007b. 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. 2010. 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.

Starks H, Diehr P, Curtis JR. 2009. The challenge of selection bias and confounding in palliative care research. J Palliat Med 12:181–187.
Steward O, Popovich PG, Dietrich WD, Kleitman N. 2012. Replication and reproducibility in spinal cord injury research. Exp Neurol. 233:597–605.
The Lancet. 2010. CONSORT 2010. The Lancet 375:1136.
van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. 2010. Can animal models of disease reliably inform human studies? PLoS Med 7:e1000245.
van der Worp HB, Macleod MR. 2011. Preclinical studies of human disease: Time to take methodological quality seriously. J Mol Cell Cardiol 51:449–450.
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.
Wheble PC, Sena ES, Macleod MR. 2008. A systematic review and meta-analysis of the efficacy of piracetam and piracetam-like compounds in experimental stroke. Cerebrovasc Dis 25:5–11.
Woodruff TJ, Sutton P. 2014. The Navigation Guide systematic review methodology: A rigorous and transparent method for translating environmental health science into better health outcomes. Environ Health Perspect 122:1007–1014.