ISSN 1466-2523

Animal Health Research Reviews **15**(1); 3–13
doi:10.1017/S146625231400005X

Methodology

An introduction to systematic reviews in animal health, animal welfare, and food safety

A. M. O'Connor[1]* and J. M. Sargeant[2,3]

[1]*Department of Veterinary Diagnostic and Production Animal Medicine, Iowa State University College of Veterinary Medicine, Ames, IA, USA*
[2]*Centre for Public Health and Zoonoses, University of Guelph, Guelph, CA, USA*
[3]*Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph, CA, USA*

Received 9 February 2014; Accepted 25 April 2014

Abstract

In this paper, we provide an introduction to systematic reviews and discuss the process for conducting systematic reviews in animal health, animal welfare, and food safety. The research synthesis need that can be addressed by a systematic review is discussed. The use of systematic reviews to address questions about intervention effects, etiology, diagnostic tests evaluation and disease burden are discussed. The steps included in a systematic review are described.

Keywords: systematic review, evidence synthesis, meta-analysis, narrative review

Introduction

In this paper, we provide an introduction to systematic reviews and discuss the process for conducting systematic reviews in animal health, animal welfare, and food safety. Systematic reviews are a research synthesis approach used to answer a specific question for clinical decision-making or policy-making. They may or may not include meta-analysis, a formal quantitative approach to combining the results from multiple studies (Higgins and Green, 2011). Frequently, systematic reviews are misidentified as reviews with large searches, bias assessment, quantitative synthesis, or protocols. However, other methods of research synthesis can include these tools. Systematic reviews are defined by the combination of the type of question asked and the steps employed in the research synthesis process.

In veterinary science, animal health and animal welfare the use of systematic reviews methodology has grown dramatically in the past decade. The VETSRev – database of veterinary systematic reviews (http://webapps.nottingham.ac.uk/refbase/) catalogs over 370 systematic reviews. As an example of the growth in this methodology, the database contains 77 systematic reviews from 2013 and 6 from 2003 and none in 1993. The diversity of topics reviewed is very wide including control and treatment of infectious disease of livestock, companion animals, laboratory animals and wildlife, pathogens of foodborne origin, and animal welfare (Olivry and Mueller, 2003; Habacher *et al.*, 2006; Denagamage *et al.*, 2007; Waddell *et al.*, 2008; Newman *et al.*, 2010; Ferrer *et al.*, 2011; Petticrew and Davey Smith, 2012).

Research synthesis refers to the practice of combining information from multiple primary sources. There is no international designated taxonomy for research synthesis, and other terms that are used to describe a similar activity include evidence synthesis, knowledge translation, and knowledge synthesis. A key feature however is that research synthesis is a secondary use of research data and is not based on new primary data (Cooper *et al.*, 2009). Research synthesis is a fundamental part of the scientific process and decision-making. In animal health, welfare, and food safety, research synthesis methods are used to develop hypotheses for new research, place new findings in context, develop an overview understanding of a topic, answer very specific questions, or understand the range of topics that are being studied. Within the area of research synthesis, literature reviews are a common approach to combining information from multiple

*Corresponding author. E-mail: oconnor@iastate.edu

Table 1. Structured steps used to conduct systematic reviews of the literature

Step	Main purpose of steps	Brief summary
Pre	Prepare a protocol	Create a protocol that describes the approach each step of the review
1	Define the review question	Identify the type of question and the relevant acronym (PICO, PECO, PO, PIT) and refine the review question accordingly
2	Comprehensive search for studies	Identify the sources of information relevant to the review and within the resources (time or money) available. Document all decisions made. The search group should be designed around some or all of the components (PICO, PECO, PO, PIT) of the review question
3	Select relevant studies from the search results	Use several questions designed around the components of the review question (PICO, PECO, PO, PIT) to identify relevant studies captured by the search
4	Collect data from relevant studies	Extract information about sources of contextual and methodological heterogeneity and the outcomes from the relevant studies
5	Assess bias in relevant studies	Assess the risk of bias in the individual studies and the entire body of work
6	Synthesize the results	Conduct a meta-analysis if possible and assessment of sources of heterogeneity
7	Present the results	Present the results of Steps 2 to 6 using an appropriate combination of text, figures and tables
8	Interpret the results	Interpret the results and discuss limitations of the individual studies and the approach to conducting the review

PICO(S): P=Population, I=Intervention, C=Comparator, O=Outcome, and optionally S=study design.
PECO: P=Population, E=exposure, C=comparator, O=outcome.
PO: P=population, O=outcome.
PIT: P=Population, I=Index test(s), T=Target condition or disease.

sources. Cooper *et al.* (2009) quote the American Psychological Association as defining a literature review as 'The process of conducting surveys of previously published material'. This would suggest that a literature review is a narrower form of research synthesis because of the focus on published literature. However, the terminology for research synthesis and literature reviews it not entirely consistent. For example, Grant and Booth (2009) describe 14 types of published reviews, although they were not mutually exclusive: critical review, literature review, mapping review/systematic review, meta-analysis, mixed studies review/mixed methods review, overview, qualitative systematic review, rapid review, scoping review, state-of-the-art review, systematic review, systematic search and review, systematized review, and umbrella review. Given this large number of published review types it is important to clarify the types of questions that should be addressed by a systematic review (European Food Safety Authority (E.F.S.A.), 2010).

What is a systematic review?

A well-executed systematic review is a rigorous and replicable approach to identifying, evaluating, and summarizing scientific evidence relevant to a specific clinical or policy question (E.F.S.A., 2010). The explicit steps of a systematic review are described in Table 1.

Clinicians, policy makers, and researchers are familiar with published reviews with the goal to provide an overview of a topic. For example, if one needs to 'come up to speed' on a particular disease by understanding the information in the scientific literature about the epidemiology, pathology, diagnosis, treatment, and control options, then a systematic review is unlikely to be useful. Published reviews that are overviews, literature reviews, or critical reviews achieve this purpose.

Another question clinicians and decision-makers frequently ask is about the specific effect of an intervention or exposure on disease occurrence, the prevalence of a disease or condition or the characteristics of a diagnostic test. For this style of question, a clinician or decision-maker has a choice of information sources. The clinician or decision-maker might consult a single research study or combine the results of multiple studies that address that question. Given variation in study results, it is preferable to incorporate the results of multiple studies that answer that question, i.e. to synthesize the research. A systematic review is an approach that can address this research synthesis need. Other research synthesis approaches that could also be used to answer the question include meta-analyses, rapid reviews, or critically appraised topics. Some steps in Table 1 can be included in these other research synthesis approaches. Therefore it is the combination of the research synthesis need and the employment of these specific steps to address it that characterize a systematic review. A systematic review is a unique research synthesis tool with two characteristics: first, it is designed to answer a specific clinical or policy question, and second, it uses an explicit stepwise methodology to answer the question (Table 1).

Questions that are suitable for systematic review

One approach to clarifying the type of question that suits a systematic review is to envision whether the specific question would be answered by a parameter with a

https://doi.org/10.1017/S146625231400005X Published online by Cambridge University Press

sampling distribution. For example, consider the following questions:

Q1. What is the difference in mortality between vaccinated or unvaccinated pigs receiving a PCV II vaccine?
Q2. What is the increase in vireamia of Virus B associated with genotype XYZ compared to genotype ABC in swine?
Q3. What is the sensitivity and specificity of the Rose Bengal test for detecting *Brucella abortus* in cattle?
Q4. What is the prevalence of *Brucella abortus* antibodies in cattle and buffalo in southeast Asia?

For these questions, it is possible to envision a primary research study that could be designed to estimate the parameter of interest (i.e. effect size, prevalence, or sensitivity, and specificity), and these questions could be addressed using systematic reviews. Furthermore, it is possible to envision multiple studies that would each obtain an estimate of the parameter. These above questions differ from the following questions:

Q5. What vaccines can reduce mortality in a swine herd?
Q6. What are the diagnostic tests that can be used to detect brucellosis infection in cattle?
Q7. What is the mechanism of pathogenesis of Salmonella in feedlot cattle?
Q8. What is the epidemiology of Salmonella in feedlot cattle?

Questions such as Q5 and Q6 generate a list. Clearly, there is a need to consult the results of multiple studies to answer this question, and therefore an extensive search may be important to ensuring that the list is comprehensive. However, these questions do not naturally relate to a parameter with a sampling distribution.

Questions Q7 and Q8 request a broad understanding about the epidemiology or pathogenesis but the state of knowledge about these topics cannot be summarized by a parameter estimate. Again, there is a need to consult the results of multiple studies to answer this research synthesis need. It would also be sensible to include an extensive search and an assessment of risk of bias in such an overview. Thus, although aspects of Table 1 can be incorporated into the approaches to answer questions Q7 and Q8, these questions are not suitable for a systematic review.

Types of review questions suitable for systematic reviews

Given that the research synthesis need contains a question that can be addressed using the systematic review methodology, the next step is to clarify the type of systematic review question. Systematic review questions can further be classified as questions about interventions, etiology, disease burden (prevalence/incidence) and detection. The steps of the systematic review are the same for each of these question types. However, within each the approach to searching for data, the study designs to include, data to extract, sources of bias, data analysis, and method of presentation differ. Due to the focus of most health agencies on interventions, the systematic review methodology is most developed for questions about interventions. Knowledge of sources of biases and methods of meta-analysis for review questions about causation, disease burden, and diagnostic test evaluation are evolving and improving (Begg, 1987; Begg and Mazumdar, 1994; Deeks *et al.*, 2005; Glasziou *et al.*, 2008; Higgins *et al.*, 2013).

What is unique about systematic reviews in animal health, animal welfare, and food safety?

Much of what has already been written about the approach to systematic reviews in human health, education, public health, criminology, and sociology directly apply to animal health, animal welfare, and food safety. For example, we might consider that because animals are patients with owners this is unique; however, similar situations are encountered in human health frequently (decisions made for infants or incapacitated patients by caregivers). Similarly, the clustered populations in livestock production may be considered unique; however, clustered populations are common in education, criminology, and public health. Therefore methodologies for conducting systematic reviews that can be translated to animal health, animal welfare, and food safety are available.

Two aspects that we consider unique to animal health, animal welfare, and food safety are the use of study designs that deliberately induce disease in the species of interest (challenge studies) and the use of euthanasia in animal populations. How to weight the value of information in challenge studies in the assessment of intervention and etiology review questions is unclear. Little has been written on the topic, and ideas are still evolving; however, currently it appears that the value of challenge studies is specific to the review topic. For some outcomes, challenge studies have little relevance to the real world applications on an intervention. For example, single disease models of bovine respiratory disease may bear little resemblance to disease that occurs in feedlots. However, findings from challenge models for single pathogens such as tetanus may be very relevant. Furthermore, it is unclear if differences in estimates from challenge studies and field studies should be defined as a methodological bias. The impact of euthanasia in systematic reviews of animal health, animal welfare, and food safety is unique because it alters end points in the primary study. Methodologists need to consider the impact of different endpoints for the same outcome when incorporating such studies into systematic reviews.

https://doi.org/10.1017/S146625231400005X Published online by Cambridge University Press

Steps of a systematic review

In the following sections, we briefly outline the steps in a systematic review, listed in Table 1.

Pre-step: assemble a review team and develop a systematic review protocol

In contrast to many overview reviews, systematic reviews are conducted by a review team. The review team should be sufficient to cover the content expertize needed for the review and to complete the workload in a timely manner. Often the review team requires expertize on the intervention, the outcome(s), and the population. One or more people may fill this role. Expertize in research synthesis methods, critical appraisal, meta-analysis, and information retrieval is also required. It is not unusual to have several staff that conduct some steps of the review, especially the search, the screening, and the data extraction, under the supervision of the review team leader.

Once the review team is in place the approach to each step of the review can be designed and documented in the protocol. The development of an explicit protocol is probably the most unique and critical aspect of a systematic review. Frequently overview reviews do not start with an explicit question, but rather a theme or objective. This means that the focus of an overview can evolve over time dependent upon findings identified during the conduct of the review. For an overview review this may be acceptable. However, a systematic review is designed to explicitly answer a question in the manner reminiscent of the way primary research studies are designed to test a specific hypothesis. Furthermore, as the systematic reviews are often commissioned and conducted to be an aid in decision-making for a clinical or policy question, it is critical that the review addresses that specific question requested at the start of the project. Designing a protocol ensures that the question is answered as it was designed.

After development, the protocol should be peer reviewed prior to the conduct of the review. Peer review should be conducted by the content experts and systematic review experts. The aim of the review is not to change the review question but rather identify the issues in addressing the question that the review team has not considered. In human health and environmental health, some journals offer peer review of systematic review protocols (for an example, see: http://www.biomedcentral.com/authors/protocols). Once the review protocol has been finalized, it should be registered and made publically available. The PROSPERO database (http://www.crd.york.ac.uk/PROSPERO/) will register systematic review protocols provided they are within the PROSPERO scope. Systematic review protocols that relate the use of animals as models for human disease can be registered at the CAMARADES website (http://www.camarades.info). Currently a registry for review protocols that relate to other use of animals and the wider range of food safety issues is not available.

The process for developing the protocol and registration should be included in the final report of the review (Liberati *et al.*, 2009; Moher *et al.*, 2009).

Step 1: Define the review question

Systematic reviews should begin by defining explicitly the components of the review question. For reviews about interventions the format to create the review question is summarized by the acronym PICO(S), which stands for P=Population, I=Intervention, C=Comparator, O=Outcome, and optionally S=study design. For reviews about etiology the format to create the review question is summarized by the acronym PECO(S): P=Population, E-Exposure, C=Comparator, O=Outcome, and optionally S=study design (E.F.S.A., 2010). For reviews about disease burden (prevalence or incidence) the format uses the acronym PO: P=Population and O=Outcome (E.F.S.A., 2010). For reviews about diagnostic test accuracy (DTA) the format uses the acronym PIT: P=Population, I=Index test(s), and T=Target condition or disease (Deeks, 2001).

By defining each of these components, it will be clear to the end user what studies were relevant to the review question. Relevant studies address the component of the question, which are explicitly defined by eligibility criteria. For animal populations, the population is frequently defined by a combination of species, production system, age, and/or reproductive status. The intervention refers to a therapeutic or preventive intervention applied by an investigator, clinician, or policy group. The exposure refers to a factor that may either increase risk or protect against the outcome. In reality, there may be situations where there is little difference between intervention and exposure. For example, in human health exercise may be an intervention or a protective factor. Therefore sometimes the distinction is not clear. Often, PICO questions are limited to deliberate exposure and questions of intervention effect, whereas PECO questions relate to disease etiology.

The comparator can be either an active or non-active comparator or unexposed category. An active comparator would likely be the current recommended standard of care or a common standard of care. A non-active comparator may be a placebo or a non-treated group. Frequently, when the comparator is non-active or unexposed this is not stated explicitly in the review question as it is implied.

The outcome of interest or target condition must also be clarified for any review question (PICO, PECO, PO, and PIT). Phrases such as 'effect on production' or 'impact on welfare' are too vague for systematic reviews and must

https://doi.org/10.1017/S146625231400005X Published online by Cambridge University Press

be refined. It is preferable to describe outcomes as study subject metrics that can be quantified. For instance, the effect on production can be measured by the effect of the exposure on average daily gain. The impact on welfare may be measured by the effect on time resting. When multiple outcomes are of interest, these should all be specified. Some reviews collect data on multiple important outcomes including adverse events, such reviews are really several systematic reviews conducted simultaneously to obtain a more complete picture of the effect of an intervention. For DTA reviews, if there is a reference standard, this is used to determine whether or not the target condition is present. A unique aspect of the DTA reviews is the index test. This is the test(s) that is being evaluated.

Some review teams limit the scope of the review by the study design. For reviews about the effect of an intervention, it is common to limit the review to randomized controlled trials with naturally occurring disease (Higgins and Green, 2011). In animal health, welfare, and food safety, controlled trials that occur using induced models of disease may also be used to assess interventions and the review team must decide about the relevance of results from such studies. If considered relevant, the results from such studies will be included in the review. For reviews about etiology, data from observational studies, and controlled trials with natural or induced disease may be relevant. Study designs for DTA are poorly understood and the review team should include, or consult, individuals with expertize on the available designs and sources of bias that can occur in DTA studies. A longer discussion about how study design can be considered in each review question is available elsewhere (O'Connor and Sargeant, 2014).

The review question should be reported in the report of the review using the relevant acronym (Liberati *et al.*, 2009; Moher *et al.*, 2009).

Step 2: Conduct an extensive search for studies

The aim of conducting an extensive search is to ensure as many relevant results as possible are included in the review. The rationale is to reduce the bias associated with the accessibility of studies based on their outcome, sometimes called retrieval bias. Retrieval bias is a subtype of publication bias, in that studies with more favorable or interesting outcomes are published in higher profile locations that are easier to access (Scherer *et al.*, 1994; Krzyzanowska *et al.*, 2004). In animal health, evidence of bias toward publication of positive findings is not currently available. In one study that assessed this question for trials that reported assessment of vaccines for swine and bovine diseases, so very few conference abstracts were subsequently published that the power to detect such a bias was limited (Brace *et al.*, 2010). For food safety outcomes, there is evidence of bias: abstracts reporting at least one positive outcome were more likely to be published (OR=2.6: 1.1, 6.2) and were published faster (HR=2.3: 1.1, 4.7). Time to publication decreased with the number of positive outcomes reported (HR=1.1: 1.0, 1.3) (Snedeker *et al.*, 2010a, b).

The search should be designed based on some or all of the concepts included in the review question (i.e. the PICO(S), PECO(S), PO(S) or PIT(S) components). It is not always necessary to include all components in the search. For example, frequently the outcomes are not included in the search for reviews of interventions, as these may not be explicitly reported in the abstracts of all studies. The aim is to design a search that will capture as many relevant studies as possible (high sensitivity) with as few irrelevant studies (high precision) as possible (Higgins and Green, 2011). This inevitably involves a trade-off to achieve as few false negatives as possible (relevant papers missed) even if this results in a large number of false positives (irrelevant papers included).

The search should be extensive. Therefore consideration should be given to the range of electronic citations databases to be used. In animal health and welfare reviews, the inclusion of CAB Abstracts is likely to be important, as data suggest that it provides the most comprehensive coverable of animal health topics (Grindlay *et al.*, 2012). In food safety, a large list of relevant databases exists and a librarian familiar with indexing in those databases should be consulted. Often particular conferences are of interest, and it should be verified if these are indexed, and if not handsearching of the relevant years of conferences proceedings should be included. Other important sources of non-peer reviewed literature, which may be unique to a topic, should be considered. Identification of these unique topic-specific information sources is an important role of content experts in the review protocol development process. More details about performing the search are available (Higgins and Green, 2011; Grindlay *et al.*, 2012). Approaches to reporting the search for systematic reviews are described in the PRISMA statement (Liberati *et al.*, 2009; Moher *et al.*, 2009).

Step 3: Selecting relevant studies from the results of the search

Once the citations have been retrieved it is necessary to evaluate them and identify those relevant to the review. This step is called screening because the aim is to screen out non-relevant studies. This step could also be called eligibility screening, as non-relevant studies are not eligible. To conduct the screening, a series of short questions are applied to each citation. The questions are designed during the protocol development and pre-tested to ensure agreement by the review team that they effectively exclude non-relevant studies and identify relevant studies. Initially, relevance screening is

https://doi.org/10.1017/S146625231400005X Published online by Cambridge University Press

performed using the titles and abstracts for citations identified by the search. If the information required to answer the questions is not provided in the title or abstract, it may be necessary to obtain the full paper to exclude or include a manuscript. An example set of questions that might be used for a review of a *Bordatella* vaccine to reduced upper respiratory disease in dogs might be:

- Does the title or abstract describe a primary research study of dogs? (population)
- Does the title or abstract describe an assessment of registered Bordatella vaccines? (intervention)
- Does the title or abstract include a negative control group? (comparator)
- Does the title or abstract describe upper respiratory disease and clinical signs as an outcome? (outcome)

Step 4: Collecting data from relevant studies

Once the screening is complete, the next step in a systematic review is the extraction of data from the relevant studies. In this step, the results of the relevant studies are extracted, as are the potential sources of contextual heterogeneity (characteristics of the population, intervention, or comparator that could impact the study results).

One of the key features of systematic reviews is the emphasis on extraction and reporting of the magnitude of outcomes and precision around estimates. Unlike overviews, the inference from either hypothesis testing (significant or not significant) or author's interpretation from the original research publication is generally not reported. The type of outcome depends upon the nature of the data and the question. For PICO and PECO questions, the aim is to obtain an estimate of the intervention effect size. For disease outcomes, effect sizes are often expressed as the risk ratio, rate ratio, prevalence ratio, risk odds ratio, prevalence odds ratio, or exposure odds ratio. Production outcomes such as milk yield and average daily gain are often continuous, in which case the mean difference or standardized mean difference is of interest. Some studies report group-level information (e.g. the proportion experiencing the outcome for the intervention group and for the control group) and the end user is required to calculate the effect size, others report the effect size directly. The measures of variation for either the group-level measure or summary statistic must also be extracted or calculated. Knowledge of approaches to analysis is helpful when extracting data.

To place the results in context, end users need to be aware of possible sources of heterogeneity in intervention effect among the relevant studies (Deeks *et al.*, 2011; Khan *et al.*, 2012). Although the review question has limited the eligibly of the studies, often there are still sources of heterogeneity within the relevant studies. The content experts on the review team should identify these sources for data extraction. Differences in the populations studied may contribute to differences in observed intervention effects. If the population was not very refined in the review question there may be sources of heterogeneity such as age, sex, or breed that need to be extracted. The comparator is often not a source of variation in animal health, animal welfare, and food safety trials, as many studies use non-active controls. However, if an active control is used, as with the intervention, it may also vary. In particular, when the comparator is the prevailing standard of care this can vary over time or between countries. If the variation in the comparator is substantial, this may preclude a simple pairwise meta-analysis and require advanced methods of analysis.

Step 5: Assess the risk of bias in relevant studies

Assessment of risk of bias aims not just to convey to the end user the presence or absence of design features that are associated with bias. Instead the rationale behind transparently reporting the potential for bias is to alert the end user of potential concerns and uncertainties about the evidence summary, which should be considered in the decision-making process. At the protocol development stage, the systematic review team should have considered the potential sources of systematic bias that could occur in studies relevant to the specific review question. The sources of bias may be subject-specific, and certainly are specific to the type of review question, i.e. intervention, exposure, diagnostic test assessment, and prevalence estimate.

For PICO(S) and PECO(S) questions, checklists from reporting guidelines such as the CONSORT statement (Moher *et al.*, 2010; Schulz *et al.*, 2010) or the REFLECT statement (O'Connor *et al.*, 2010; Sargeant *et al.*, 2010) should not be employed as a risk of bias tool. These reporting guidelines provide a yes or no answer about the presence of absence of design features. However, there is not always a high risk of bias when a design feature is absent. Based on the presence of these features and the topic, the review team must decide what is the risk of bias. The Cochrane Handbook of Systematic Reviews for Interventions has an entire section devoted to assessing the risk of bias for randomized controlled trials (Higgins *et al.*, 2011). The domains of bias in this tool include selection, performance, detection, and attrition bias. For intervention studies in animal health using results from clinical trials, this tool can be used with a minimal amount of modification. The domains of bias used in the Cochrane risk of bias tool correspond to bias terms more commonly used in veterinary medicine: confounding, selection, and information bias. One topic where additional consideration may be required for veterinary medicine includes accounting for the impact of non-independence of populations in the analysis. Non-independent populations are quite common in veterinary settings: examples include

https://doi.org/10.1017/S146625231400005X Published online by Cambridge University Press

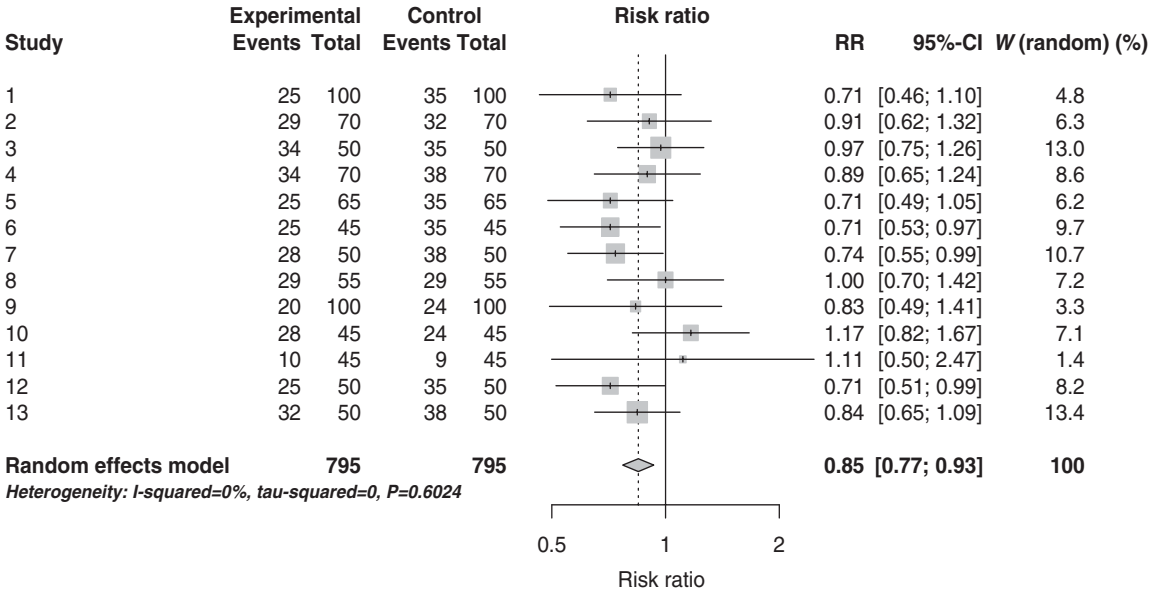

Fig. 1. Pairwise meta-analysis using hypothetical data. The forest plots include data from 13 studies, used the risk ratio from each study as the summary statistics and calculates an overall risk ratio in a random effects models.

shelter medicine, racing performance, and production medicine.

For other types of review questions (such as diagnostic test assessments, exposure assessment, and prevalence estimates) the sources of bias differ. There are also several published tools for the assessment of bias for diagnostic test assessments, exposure assessment, and prevalence estimates. The QUADAS statement is designed as a quality assessment tool of diagnostic test assessments. As it is a risk of bias tool, the QUADAS tool requires tailoring to the specific review question and identifies four domains for bias in diagnostic tests. For diagnostic test evaluation, the STARD statement is a reporting guideline that can be used to learn more about sources of bias in diagnostic tests (Bossuyt *et al.*, 2003; Christopher, 2007). Recently, proposed guidelines for quality assessment of prevalence studies have been published (Shamliyan *et al.*, 2011; Giannakopoulos *et al.*, 2012). New tools are also available for the assessment of risk of bias in non-randomized studies which may be useful for questions of etiology (Higgins *et al.*, 2013; Valentine and Thompson, 2013).

Quantitative scores of quality or bias should not be used (Whiting *et al.*, 2005; Higgins *et al.*, 2011). Such scoring systems are clearly arbitrary and do not convey the importance of bias to the topic.

The role of study design, rather than bias within the design, as a potential source of heterogeneity is poorly understood in the assessment of interventions. The easiest solution is to limit systematic reviews to a single design at the start of the review (e.g. only randomized controlled trials may be deemed eligible). In this situation, methodological heterogeneity only relates to biases that occur due to execution as described above. The availability of challenge studies, where researchers are able to induce the disease experimentally and assess interventions, makes this issue particularly relevant to veterinary science. In the authors' opinion, it is currently unclear if heterogeneity that might be expected due to different study designs should be classified as biases. However, if different designs are included in the review, this information must be reported to the end user.

Step 6: Synthesize the results

Having extracted the data about the results, study characteristics, and potential sources of contextual or methodological (risk of bias) heterogeneity, the next step is to synthesize the results. If the data are amenable to quantitative analysis, a meta-analysis can be conducted, although a meta-analysis is not always a component of a systematic review. Similarly, many meta-analyses do not employ the systematic review methodology. Meta-analysis aims to combine, for each outcome, the observed result from each of the relevant studies into one estimate.

The essential elements of a meta-analysis entail planning, conducting, and interpreting the meta-analysis. Planning requires deciding which comparisons to make, what summary effect measure to use, which model to use, and which sources of heterogeneity to assess. For example, when comparing two interventions possible comparisons include differences in mortality, morbidity, and weight gain. Conducting a meta-analysis involves assessing heterogeneity and calculating the summary effect measure if appropriate. Often, forest plots are used to display the results (Lewis and Clarke, 2001). An example is provided in Fig. 1. Each row of Fig. 1 illustrates the results of a comparison from a primary study, with the

https://doi.org/10.1017/S146625231400005X Published online by Cambridge University Press

Table 2. Evidence profile for hypothetical review of an intervention to prevent disease.

No. of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Quality of the evidence[1]
Outcome: risk of disease							
15	12 randomized and 3 non-randomized challenge studies	No serious risk of bias	No serious risk of bias	Serious[2]	No serious imprecision	Publication bias[3]	⊕⊕ LOW

[1]GRADE Working Group grades of evidence. High quality: Further research is very unlikely to change our confidence in the estimate of effect. Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate. Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. Very low quality: We are very uncertain about the estimate.
[2]Studies all used animals younger than would be expected to develop the disease and all are challenge models.
[3]The funnel plot suggested evidence that larger studies showed effects closer to the null.

solid square representing the parameter estimate and the horizontal line representing the confidence interval on that estimate. If a meta-analysis is performed, the summary estimate is included as a diamond, with the center representing the summary parameter estimate and the points of the diamond representing the confidence intervals on that estimate.

Interpreting the meta-analysis requires consideration of how to interpret findings about heterogeneity and the summary effect measure. The summary effect estimate may reflect no effect, a strong protective effect, a weak protective effect, a strong deleterious effect, or a weak deleterious effect. Based on the same level of confidence (for example a 95% confidence interval (CI)), these may be informative (a narrow 95% CI) or non-informative (a wide 95% CI). It is beyond the scope of this summary to describe meta-analysis in detail and numerous other sources are available (Borenstein *et al.*, 2009; Cooper *et al.*, 2009). Guidelines for presenting the results of meta-analysis are described in the PRISMA statement (Liberati *et al.*, 2009; Moher *et al.*, 2009).

Frequently in animal health, animal welfare, and food safety, it is not possible to combine the results in a meta-analysis. This often occurs due to a low number of relevant studies or because of poor reporting or differences in metrics used to measure the outcomes. In this situation, meta-analysis is not possible and the presentation of the results and discussion takes the place of this step.

Step 7: Presenting the results

The presentation of the results of the review should include the following components:

- The results of the search and study selection;
- Summary information about the characteristics of the studies identified as relevant to the review (including those that could not be included in a meta-analysis);
- Risk of bias assessment for the individual studies;
- Outcomes reported by the relevant studies;
- Meta-analysis results including subgroup analysis and/or meta-regression;
- Risk of bias across the studies.

There is an entire document devoted to describing how to present the results of a systematic review and these should be considered a minimum. As already mentioned, the PRISMA statement, its accompanying explanation, and its elaboration document should be strictly adhered to when reporting the review (Liberati *et al.*, 2009; Moher *et al.*, 2009).

One of the most difficult aspects of presenting a review is presenting a summary of the overall assessment of the body of work. Approaches to presenting this information have evolved since the publication of PRISMA. One approach used by the Cochrane Collaboration is to

https://doi.org/10.1017/S146625231400005X Published online by Cambridge University Press

Table 3. Summary of findings table for hypothetical review of an intervention to prevent disease

Outcome	Illustrative comparative risks (95% CI)		Relative effect (95% CI)	Absolute effect (95% CI)	No. of Participants (studies)	Quality of the evidence[1]
	Assumed risk	Corresponding risk				
	No intervention	Intervention				
Risk of disease	Study population 695 cases per 1000	515 cases per 1000 (452–577)	RR 0.74 (0.65–0.83)	181 fewer cases per 1000 (from 118 fewer to 243 fewer)	2547 (15 studies)	⊕⊕ low due to indirectness and publication bias
	Low risk population (10%) 100 per 1000	74 cases per 100 (65–82)		26 fewer cases per 1000 (from 17 fewer to 35 fewer)		

[1]GRADE Working Group grades of evidence. High quality: Further research is very unlikely to change our confidence in the estimate of effect. Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate. Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. Very low quality: We are very uncertain about the estimate.

provide an evidence profile and summary of findings table (Higgins and Green, 2011). These two tables together aim to summarize the findings with an overall assessment of bias (i.e. across all studies, as opposed to the risk of bias within studies that was previously described). A summary of findings table is based on a meta-analysis and therefore may not be possible; however, even when meta-analysis cannot be conducted an evidence profile can be created and is informative for end-users. The evidence profile provides a structured means of summarizing the risk of bias for each outcome included in the systematic review. The issues considered in the evidence profile include inconsistency (heterogeneity in study effect sizes) (Guyatt *et al.*, 2011e), risk of bias in the studies (Guyatt *et al.*, 2011g), indirectness (i.e. relevance of the study populations to the target population) (Guyatt *et al.*, 2011d), imprecision (related to the number of studies and number of subjects) (Guyatt *et al.*, 2011b), and other considerations (usually publication bias) (Guyatt *et al.*, 2011f). Table 2 provides an example of an evidence profile and Table 3 provides a summary of findings from a hypothetical review.

More information about creating the summary of findings tables and evidence profiles provided in a series of publications from the GRADE working group (Guyatt *et al.*, 2008, 2011a, b, c, d, e, f, g; Balshem *et al.*, 2011).

Step 8: Interpret the results and discussion

As with any research project, conclusions about the results of the review and a discussion of potential biases of the review should be discussed to enhance the end user's understanding of the issues related to the review and the review authors' interpretation of the meaning of the results. The PRIMSA guidelines suggest that the discussion and interpretation include a summary of the evidence, a discussion of the limitations of the review, and the overall conclusions (Moher *et al.*, 2009). The discussion of the evidence should include a consideration of the magnitude of the summary effect and the precision of that estimate, as well as a discussion on the potential impact of bias. The summary effect estimate may suggest no effect, or may reflect a strong protective effect, a weak protective effect, a strong deleterious effect, or a weak deleterious effect. Based on the width of the confidence interval and quality of body of work, the certainty about this effect will differ. The discussion of the impact should include the expected direction of bias rather than mere mention that such biases could occur.

The discussion of the limitations of the review should include two components; the issues identified in the studies themselves and also the approach to the review. Despite attempts to be transparent and comprehensive, all reviews have limited resources. For instance, the review may have been restricted to only studies published in English or perhaps assumptions were made in the

https://doi.org/10.1017/S146625231400005X Published online by Cambridge University Press

approach to extracting measures of variation for continuous outcomes; these types of issues should be included in the discussion on the review limitations.

Conclusion

Systematic reviews are one tool available for summarizing evidence for decision-making in animal health, animal welfare, and food safety. The tool is appropriate when a specific question is asked about an intervention, an exposure, a diagnostic test or disease burden (prevalence/incidence). Critical aspects of systematic reviews are an *a priori* question, a protocol, a comprehensive search, an assessment of bias within studies in the review informing the review, extraction of the magnitude of effect rather than the statistical inference, and comprehensive reporting of the results.

The methodological approach to systematic reviews for interventions is well developed and generally transfers well to veterinary settings. Unique issues that arise in systematic reviews in animal health, animal welfare, and food safety are how to consider the evidentiary value of challenge studies in systematic reviews and the impact of euthanasia.

References

Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S and Guyatt GH (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology* **64**: 401–406.

Begg CB (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine* **6**: 411–423.

Begg CB and Mazumdar M (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50**: 1088–1101.

Borenstein M, Hedges LV, Higgins JPT and Rothstein HR (2009). *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D and De Vet HC (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clinical Chemistry* **49**: 1–6.

Brace S, Taylor D and O'connor AM (2010). The quality of reporting and publication status of vaccines trials presented at veterinary conferences from 1988 to 2003. *Vaccine* **28**: 5306–5314.

Christopher MM (2007). Improving the quality of reporting of studies of diagnostic accuracy: let's STARD now. *Veterinary Clinical Pathology* **36**: 6.

Cooper HM, Hedges LV and Valentine JC (2009). *The Handbook of Research Synthesis and Meta-Analysis*. New York: Russell Sage Foundation.

Deeks JJ (2001). Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *British Medical Journal* **323**: 157–162.

Deeks JJ, Macaskill P and Irwig L (2005). The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology* **58**: 882–893.

Deeks JJ, Higgins JPT and Altman DG (2011). *Chapter 9: Analysing data and undertaking meta-analyses*. In: Higgins JPT and Green S (eds) *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Denagamage TN, O'connor AM, Sargeant JM, Rajic A and Mckean JD (2007). Efficacy of vaccination to reduce Salmonella prevalence in live and slaughtered swine: a systematic review of literature from 1979 to 2007. *Foodborne Pathogens and Disease* **4**: 539–549.

European Food Safety Authority (E.F.S.A.) (2010). Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA Journal* **8**: 1–90.

Ferrer M, Bildstein K, Penteriani V, Casado E and De Lucas M (2011). Why birds with deferred sexual maturity are sedentary on islands: a systematic review. *PLoS ONE* **6**: e22056.

Giannakopoulos NN, Rammelsberg P, Eberhard L and Schmitter M (2012). A new instrument for assessing the quality of studies on prevalence. *Clinical Oral Investigations* **16**: 781–788.

Glasziou P, Irwig L and Deeks JJ (2008). When should a new test become the current reference standard? *Annals of Internal Medicine* **149**: 816–822.

Grant MJ and Booth A (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal* **26**: 91–108.

Grindlay DJ, Brennan ML and Dean RS (2012). Searching the veterinary literature: a comparison of the coverage of veterinary journals by nine bibliographic databases. *Journal of Veterinary Medical Education* **39**: 404–412.

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P and Schunemann HJ (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal* **336**: 924–926.

Guyatt GH, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, Debeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P and Schunemann HJ (2011a). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology* **64**: 383–394.

Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux P, Montori VM, Freyschuss B, Vist G, Jaeschke R, Williams Jr JW, Murad MH, Sinclair D, Falck-Ytter Y, Meerpohl J, Whittington C, Thorlund K, Andrews J and Schunemann HJ (2011b). GRADE guidelines 6. Rating the quality of evidence-imprecision. *Journal of Clinical Epidemiology* **64**: 1283–93.

Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, Alderson P, Glasziou P, Falck-Ytter Y and Schunemann HJ (2011c). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology* **64**: 395–400.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G, Akl EA, Post PN, Norris S, Meerpohl J, Shukla VK, Nasser M and Schunemann HJ (2011d). GRADE guidelines: 8. Rating the quality of evidence-indirectness. *Journal of Clinical Epidemiology* **64**: 1303–10.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Glasziou P, Jaeschke R, Akl EA, Norris S, Vist G, Dahm P, Shukla VK, Higgins J, Falck-Ytter Y and Schunemann HJ (2011e). GRADE

https://doi.org/10.1017/S146625231400005X Published online by Cambridge University Press

guidelines: 7. Rating the quality of evidence-inconsistency. *Journal of Clinical Epidemiology* **64**: 1294–302.

Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams Jr JW, Meerpohl J, Norris SL, Akl EA and Schunemann HJ (2011f). GRADE guidelines: 5. Rating the quality of evidence-publication bias. *Journal of Clinical Epidemiology* **64**: 1277–82.

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams Jr JW, Atkins D, Meerpohl J and Schunemann HJ (2011g). GRADE guidelines: 4. Rating the quality of evidence–study limitations (risk of bias). *Journal of Clinical Epidemiology* **64**: 407–415.

Habacher G, Pittler MH and Ernst E (2006). Effectiveness of acupuncture in veterinary medicine: systematic review. *Journal of Veterinary Internal Medicine* **20**: 480–488.

Higgins JPT, Altman DG and Sterne JAC (2011). Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT and Green S (eds) *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Higgins JPT and Green S (eds) (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated March 2011]*. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Higgins JPT, Ramsay C, Reeves BC, Deeks JJ, Shea B, Valentine JC, Tugwell P and Wells G (2013). Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 12–25.

Khan KS, Ball E, Fox CE and Meads C (2012). Systematic reviews to evaluate causation: an overview of methods and application. *Evidence-based Medicine* **17**: 137–41.

Krzyzanowska MK, Pintilie M, Brezden-Masley C, Dent R and Tannock IF (2004). Quality of abstracts describing randomized trials in the proceedings of American Society of Clinical Oncology meetings: guidelines for improved reporting. *Journal of Clinical Oncology* **22**: 1993–1999.

Lewis S and Clarke M (2001). Forest plots: trying to see the wood and the trees. *British Medical Journal* **322**: 1479–1480.

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J and Moher D (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of Clinical Epidemiology* **62**: e1–e34.

Moher D, Liberati A, Tetzlaff J and Altman DG (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology* **62**: 1006–1012.

Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M and Altman DG (2010). CONSORT 2010 Explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal* **340**: c869.

Newman J, Westgarth C, Pinchbeck G, Dawson S, Morgan K and Christley R (2010). Systematic review of human-directed dog aggression. *Veterinary Record* **166**: 407.

O'Connor AM, Sargeant JM, Gardner IA, Dickson JS, Torrence ME, Dewey CE, Dohoo IR, Evans RB, Gray JT, Greiner M, Keefe G, Lefebvre SL, Morley PS, Ramirez A, Sischo W, Smith DR, Snedeker K, Sofos J, Ward MP, Wills R and Consensus Meeting Participants. (2010). The REFLECT Statement: methods and processes of creating reporting guidelines for randomized controlled trials for livestock and food safety. *Journal of Food Protection* **73**: 132–139.

O'Connor AM and Sargeant JM (2014). Meta-analyses including data from observational studies. *Preventive Veterinary Medicine* **113**: 313–22.

Olivry T and Mueller RS (2003). Evidence-based veterinary dermatology: a systematic review of the pharmacotherapy of canine atopic dermatitis. *Veterinary Dermatology* **14**: 121–146.

Petticrew M and Davey Smith G (2012). The monkey puzzle: a systematic review of studies of stress, social hierarchies, and heart disease in monkeys. *PLoS ONE* **7**: e27939.

Sargeant JM, O'Connor AM, Gardner IA, Dickson JS, Torrence ME, Dohoo IR, Lefebvre SL, Morley PS, Ramirez A and Snedeker K (2010). The REFLECT Statement: reporting guidelines for randomized controlled trials in livestock and food safety: explanation and elaboration. *Journal of Food Protection* **73**: 579–603.

Scherer RW, Dickersin K and Langenberg P (1994). Full publication of results initially presented in abstracts. A meta-analysis. *Journal of the American Medical Association* **272**: 158–162.

Schulz KF, Altman DG and Moher D (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Obstetrics Gynecology* **115**: 1063–1070.

Shamliyan TA, Kane RL, Ansari MT, Raman G, Berkman ND, Grant M, Janes G, Maglione M, Moher D, Nasser M, Robinson KA, Segal JB and Tsouros S (2011). Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists. *Journal of Clinical Epidemiology* **64**: 637–657.

Snedeker KG, Campbell M, Totton SC, Guthrie A and Sargeant JM (2010a). Comparison of outcomes and other variables between conference abstracts and subsequent peer-reviewed papers involving pre-harvest or abattoir-level interventions against foodborne pathogens. *Preventive Veterinary Medicine* **97**: 67–76.

Snedeker KG, Totton SC and Sargeant JM (2010b). Analysis of trends in the full publication of papers from conference abstracts involving pre-harvest or abattoir-level interventions against foodborne pathogens. *Preventive Veterinary Medicine* **95**: 1–9.

Valentine JC and Thompson SG (2013). Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 26–35.

Waddell LA, Rajic A, Sargeant J, Harris J, Amezcua R, Downey L, Read S and Mcewen SA (2008). The zoonotic potential of Mycobacterium avium spp. paratuberculosis: a systematic review. *Canadian Journal of Public Health* **99**: 145–155.

Whiting P, Harbord R and Kleijnen J (2005). No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Medical Research Methodology* **5**: 19.