Frampton *et al. Environ Evid (2017) 6:27*
DOI 10.1186/s13750-017-0102-2

Environmental Evidence

METHODOLOGY

Open Access

Eligibility screening in evidence synthesis of environmental management topics

Geoff K. Frampton[1*], Barbara Livoreil[2] and Gillian Petrokofsky[3]

Abstract

The eligibility screening step of a systematic review or systematic map (sometimes referred to as 'study selection', 'evidence selection' or 'inclusion screening') determines the scope of the evidence that may answer the review or map question. Eligibility screening involves the development, testing and application of eligibility criteria (inclusion and exclusion criteria) by an evidence synthesis review team, based on methods pre-specified in the review or map protocol. Some parts of the process require judgement, meaning that consistent and transparent reporting of the eligibility criteria and the process for applying them are essential in order to reduce the risk of introducing errors or bias. The existing Collaboration for Environmental Evidence (CEE) Guidelines for Systematic Reviews in Environmental Management (version 4.2, March 2013) give relatively limited guidance on how to conduct eligibility screening. In this paper we provide more in-depth information on good practice methods for this step of evidence synthesis, based on a critical consideration of existing guidance and current practice. Our aim is to provide recommendations to support those conducting CEE systematic reviews or systematic maps for environmental management questions; however, the methods we describe are generic and should be broadly applicable across a wide range of environmental research topics.

Keywords: Eligibility criteria, Inclusion criteria, Study selection, Systematic review, Systematic map, Quality, Bias, Reviewer agreement

Background

Systematic reviews follow a structured process in order to answer specific questions whilst minimising the risks of errors or bias [1–4]. Systematic maps also follow a structured process, to determine the extent and characteristics of a specified evidence base, but provide answers which are more descriptive [5]. Although these approaches to evidence synthesis have different objectives, they both need to identify all information relevant to the questions they are addressing, so as to reduce the risk of selective inclusion or exclusion of evidence [6]. In both approaches an extensive search for evidence should therefore have been carried out using as wide a range of information sources as possible [7]. To ensure that key evidence is not missed, searches are usually designed to have high *sensitivity*, that is, to identify as many relevant articles (or other items of evidence) as possible. However, highly sensitive searches tend to have low *specificity*, meaning poor discrimination of information that is truly irrelevant (note that sensitivity is sometimes called *recall* or *exhaustivity* and specificity is sometimes called *precision*) [7]. Consequently, sensitive searches typically return large numbers of irrelevant articles (e.g. exceeding 20,000 in an evidence synthesis on the impacts of alternative livelihood projects [8]. The challenge for those of us conducting systematic reviews or systematic maps is to separate the evidence from the large amount of irrelevant information without introducing errors or bias.

In this paper we provide a guide to good practice in planning, conducting and reporting the eligibility screening step for systematic reviews or systematic maps in environmental research. This is based on: (1) critical consideration of the existing CEE Guidelines for Systematic

*Correspondence: gkf1@soton.ac.uk
[1] Southampton Health Technology Assessments Centre (SHTAC), Faculty of Medicine, University of Southampton, Southampton, UK
Full list of author information is available at the end of the article

Reviews in Environmental Management (version 4.2, March 2013); (2) observation of current practice in environmental evidence synthesis, as indicated by the most recent systematic review and systematic map protocols published in Environmental Evidence journal (January–July 2017; n = 11 protocols); and (3) relevant guidance publications on the application of systematic reviews and systematic maps in environmental research [2, 9–12], as well as in the health and social sciences [4, 13–17]. The present paper goes beyond providing a summary of existing guidance and practice; we indicate where current practice in eligibility screening in environmental evidence synthesis could be further improved to reduce the risks of introducing errors or bias.

The eligibility screening step of a systematic review or systematic map (which may also be referred to as 'study selection', 'evidence selection' or 'inclusion screening') involves the specification of *eligibility criteria* that determine which of the primary research studies identified in searches are relevant for answering the review or map question; and the use of a systematic *screening process* for applying the eligibility criteria to the search results in such a way as to minimise the risk of introducing selection bias [18]. Both the eligibility criteria and the screening process should be specified in the evidence synthesis protocol [1]. Development of the protocol is an iterative process in which a draft version is refined based on pilot-testing and the methods are updated if necessary until they are considered adequately reproducible, efficient and objective to be applied in the full evidence synthesis. For CEE evidence syntheses the final version of the protocol should be peer reviewed and published prior to the systematic review or systematic map being conducted [1].

The following pages summarise the preparatory steps of eligibility screening, then explain the rationale and methods for determining the eligibility criteria and the screening process. We then discuss pilot-testing the eligibility criteria and screening process, before providing an overall summary of recommendations for good practice. Note that the eligibility criteria and the screening process are described in separate sections for clarity, although in practice they are intrinsically linked and would be iteratively developed and pilot-tested together when preparing an evidence synthesis protocol. We have used the existing CEE guidelines for Systematic Reviews in Environmental Management (version 4.2, March 2013; section 4.2 'Screening articles for relevance') [1] and a related report [10] as a basis for our recommendations, but in some cases our recommendations for good practice differ, as explained below.

Preparing for eligibility screening

Bibliographic searches may produce thousands or sometimes tens of thousands of references that require screening for eligibility and so it is important to ensure that search results are organised in such a way that they can be screened efficiently for their eligibility for an evidence synthesis. Key actions that will be necessary before screening can commence are to assemble the references into a library, using one or more bibliographic reference management tools; and to identify and remove any duplicate references.

Assembling references

A range of bibliographic reference management tools are available into which search results may be downloaded directly from bibliographic databases or imported manually, and these vary in their complexity and functionality. Some tools, such as Eppi Reviewer [19] and Abstrackr [20] include text mining and machine learning functionality to assist with some aspects of eligibility screening. According to recently-published evidence syntheses and protocols, the most frequently-used reference management tools in CEE evidence syntheses are Endnote and Eppi Reviewer (sometimes used in combination with Microsoft Excel), although others such as Mendeley and Abstrackr are also used. Given that reference management tools have diverse functionality and are continually being developed and upgraded, it is not possible to recommend any one tool as being 'better' than the others. An efficient reference management tool should:

- enable easy removal of duplicate articles (see "Removing duplicates" below), which can reduce substantially the number of articles;
- readily locate and import abstracts and full-text versions for articles where available;
- enable the review team to record their screening decisions for each article;
- enable articles, and any screening decisions accompanying them, to be grouped and analysed to assist the team in checking progress with eligibility screening and in identifying any disagreements between screeners.

Other features of reference management tools that review teams may find helpful to consider are: whether the software is openly accessible (e.g. Mendeley) or may require payment (e.g. Endnote, Eppi Reviewer); the number of references that can be accommodated; the number of screeners who can use the software simultaneously; and how well suited the tool is for project management tasks,

such as allocating eligibility screening tasks among the review team members and monitoring project progress.

Removing duplicates

Duplicate articles are common in search results and should be removed where possible before eligibility screening starts. Inclusion of duplicates in an evidence synthesis could lead to double-counting of data, which might introduce bias [21], as well as creating unnecessary additional screening effort. Many reference management tools enable automated identification and removal of duplicate articles (e.g. 'fuzzy matching' of references in Eppi Reviewer) and this may be particularly helpful if large numbers of duplicates are present. However, care should be taken to avoid inadvertently removing articles which are not duplicates. If an automated process is used for identifying duplicates it should not be assumed that this will always classify the articles accurately.

The eligibility criteria

Rationale for eligibility criteria

The use of pre-specified and explicit eligibility criteria ensures that the inclusion or exclusion of articles or studies from a systematic review or systematic map is done in a transparent manner, and as objectively as possible. This reduces the risk of introducing errors or bias which could occur if decisions on inclusion or exclusion are selective, subjective, or inconsistent. An objective and transparent approach also helps to ensure reproducibility of eligibility screening. Failing to consistently apply eligibility criteria, or using criteria which are not relevant to the evidence synthesis question, can lead to inconsistent conclusions from different evidence syntheses (e.g. illustrated by Englund et al. 1999 [22] for stream predation experiments and McDonagh et al. 2013 [18] for health research studies).

The eligibility criteria for a systematic review or systematic map should reflect the question being asked and therefore follow logically from the 'key elements' that describe the question structure. Many environmental questions are of the 'PICO' type, where the interest is on determining effects of an intervention within a specified population. For a PICO-type question the key elements (P, I, C, O) would specify which population(s), intervention(s), comparator(s) and outcome(s) must be reported in an article describing a primary research study in order for that article to be eligible for inclusion in the evidence synthesis (examples of PICO and other types of question structure are given by EFSA 2010 [2], James et al. 2016 [5] and Aiassa et al. 2015 [9]).

An example of eligibility criteria for an environmental intervention (i.e. PICO-type) systematic review question is shown in Box 1, for the question 'What are the environmental and socioeconomic effects of China's Conversion of Cropland to Forest Programme (CCFP) after the first 15 years of implementation?' [23]. As the example illustrates, eligibility criteria may be expressed as inclusion criteria and, if helpful, also as exclusion criteria.

Ideally, the eligibility criteria should be specified in such a way that they are easy to interpret and apply by the review team with minimal disagreement. For some systematic review or systematic map questions the eligibility criteria may be very similar to or identical to the question key elements and the question itself, whereas in other cases (e.g. as in the example in Box 1) the eligibility criteria may need to be more specific, to provide adequate information for the review team to make selection decisions.

In the example systematic review question (Box 1) it is clear that if an article describing a primary research study did not provide information on the intervention (i.e. the Conversion of Cropland to Forest Programme) then it would not be appropriate for answering the review question. As such, the article could be excluded. Similarly, an article that did not report any environmental or socioeconomic outcomes would not be relevant and could be excluded. The example question illustrates that articles can be efficiently excluded if they fail to meet one or more inclusion criteria; they are included only if they meet all the eligibility criteria.

Keeping the list of eligibility criteria short and explicit, and specifying the criteria such that an article would be excluded if it fails one or more of the criteria is a useful approach since this minimises the range of information that members of the review team would need to locate in an article and means that if an article is clearly seen not to meet one of the criteria then the remaining criteria would not have to be considered. Since a single failed eligibility criterion is sufficient for an article to be excluded from an evidence synthesis, it may be helpful to assess the eligibility criteria in order of importance (or ease of finding them within articles), so that the first 'no' response can be used as the primary reason for exclusion of the study, and the remaining criteria need not be assessed [3].

The example in Box 1 is for a relatively broad systematic review question. For a systematic map the question may be even broader since the objective of a map is to provide a descriptive output. Irrespective of how broad the question is, the process for developing eligibility criteria which we have outlined here applies both to systematic reviews and systematic maps [5].

Box 1: Example of eligibility criteria in relation to question key elements for an intervention (PICO-type) environmental systematic review question (from Rodriguez et al. [23])

Question: "What are the environmental and socioeconomic effects of China's Conversion of Cropland to Forest Programme (CCFP) after the first 15 years of implementation?"

Question key elements	Eligibility criteria
Populations (P): CCFP enrolled lands (cropland/ wasteland/ ecological trees/ economic trees) CCFP households and their individual members	Included: Both human populations and land resources, including CCFP participant households, their individual members and their CCFP enrolled lands (cropland, wasteland, ecological trees, and economic trees) Excluded: Grasslands, since they no longer form part of the CCFP and because they contribute to significantly different environmental outcomes as compared with forests
Interventions (I): CCFP (subsidies, skill-training, and enforcement with field checks)	Included: CCFP compensation subsidies, skill training for local farmers, and enforcement work with field checks, and all information on other types of subsidies that might have an impact on household livelihoods and the environment. Excluded: Natural Forest Protection Programme, as this does not overlap with the CCFP
Comparators (C): Non-enrolled sloping lands, and lands prior to CCFP implementation Non-participant households, and households prior to CCFP implementation	Included: Non-enrolled sloping lands, and lands prior to CCFP implementation; and non-participant households, and households prior to CCFP implementation. Included 'before-and-after' comparators in both human populations (i.e. the socioeconomic status of both participant and non-participant households before and after the CCFP interventions) and land resources (i.e. the environmental status of both enrolled and non-enrolled lands before and after the CCFP intervention)
Outcomes (O): Environmental outcomes (changes in water discharge, soil erosion, flood risk, local biodiversity, etc.) Socioeconomic outcomes (changes in household income structure, migration, etc.)	Included: Soil erosion and flood prevention, reconversion of forestland to cropland, land-use and forest cover change, tree survival rates, biomass and carbon storage, and biodiversity. Income, employment, food security, land access and social equality, and migration Excluded: Studies assessing potential or future outcomes of the CCFP, including model projections or other predictions of program impact, as the review only sought to assess the actual impacts of CCFP implementation (i.e. those which have already taken place)

Study design eligibility criteria were also specified by Rodriguez et al. [23]; for brevity these are not reproduced here

Study design as an eligibility criterion

The types of primary research study design (e.g. observational or experimental; controlled or uncontrolled) that can answer an evidence synthesis question will vary according to the type of question. The study design is sometimes made explicit in the key elements (e.g. 'PICO'-type questions may be stated as 'PICOD' or 'PICOS' in the scientific literature, where 'D' (design) or 'S' (study) indicates that study design is being considered) [11]. Even if study design is not explicit in the question structure it should be considered as an eligibility criterion. This is particularly important for systematic reviews since the designs of studies that are included should be compatible with the planned approach for the data synthesis step (e.g. some meta-analysis methods may specifically require controlled studies). The type of study design may also be indicative of the likely validity of the evidence, since some study designs may be more prone to bias than others. Note that in systematic reviews the full assessment of risks of bias and other threats to validity takes place at the critical appraisal step, and this should always be conducted irrespective of whether any quality-related eligibility criteria have been specified.

The screening process

Rationale and overview of the screening process

The process of eligibility screening aims to ensure that the eligibility criteria are applied consistently and impartially so as to reduce the risk of introducing errors or bias in an evidence synthesis. Articles identified in searches are typically structured in having a title, an abstract (or summary), and/or a 'full text' version such as an academic journal paper, agency report, or internet pages. Eligibility screening can be applied at these different levels of reading to impose a number of filters of increasing rigor and thus screening is normally a stepwise process. The exact approach is a matter of preference, although CEE guidelines recommend that at least two filters are applied: (a) a first reading of titles and abstracts to efficiently remove articles which are clearly irrelevant; and (b) assessment of the full-text version of the article [1].

Depending on the nature of the evidence synthesis question and the number of articles requiring screening, titles and abstracts may be screened separately or together. If only an insignificant number of articles can be excluded on title alone (e.g. as found in a systematic review of the environmental impacts of poverty rights regimes by Ojanen et al. 2017 [24]), then combining the title and abstract screening in a single step may be more efficient. In cases where insufficient information is available in the title or abstract to enable an eligibility decision to be made, or if the abstract is missing, then the full-text version should be obtained and examined. An overview of the eligibility screening process is shown in Fig. 1.

As shown in Fig. 1, the screening process starts out with individual articles but final eligibility decisions are made at the level of studies, taking into account any linked articles that refer to the same study (see "Identifying linked articles" below). The evidence selection decision process is conservative at each step so that only articles which do not meet the inclusion criteria are excluded [1]; in any cases of doubt, articles proceed to the next step for further scrutiny. If after full-text screening the eligibility of a study remains unclear, further information should be sought, if feasible (e.g. by contacting the authors), to enable the study to be included or excluded. Any studies whose eligibility still remains unclear after this process should be listed in an appendix to the systematic review or systematic map report. In systematic reviews, an option could be to include studies of unclear relevance in a sensitivity analysis. The approach for handling unclear studies should be considered during protocol development and specified in the systematic review or systematic map protocol.

A single set of eligibility criteria can be used to screen titles, abstracts and full-text articles (e.g. Rodriguez et al. [23] used the eligibility criteria shown in Box 1 for screening titles and abstracts and then applied the same criteria to full-text articles). However, if the information reported in titles and abstracts is limited it may be efficient to use a smaller subset of the eligibility criteria to screen the titles and/or abstracts, and apply the more detailed full set of eligibility criteria for the screening of full-text articles. Whichever approach is used, the eligibility criteria applied at each step should be clearly stated in the protocol.

Identifying linked articles

If the same data are included more than once in an evidence synthesis this can introduce bias [21, 25, 26]. Therefore, the unit of analysis of interest in a systematic review or map is usually individual primary research studies (e.g. observational studies, surveys, or experiments), rather than individual articles.

Investigators often report the same study in more than one article (e.g. the same study could be reported in different formats such as conference abstracts, reports or journal papers, or in several different journal papers [27], and we refer to these as 'linked articles'. Although there is often a single article for each study, it should never be assumed that this is the case [3]. Linked articles may range from being duplicates (i.e. they fully overlap and do not contribute any new information) to having very little overlap. Articles which are true duplicates should be removed to avoid double-counting of data. The remaining linked articles which refer to a study should be grouped together and screened for eligibility as a single unit so that all available data pertinent to the study can be considered when making eligibility decisions.

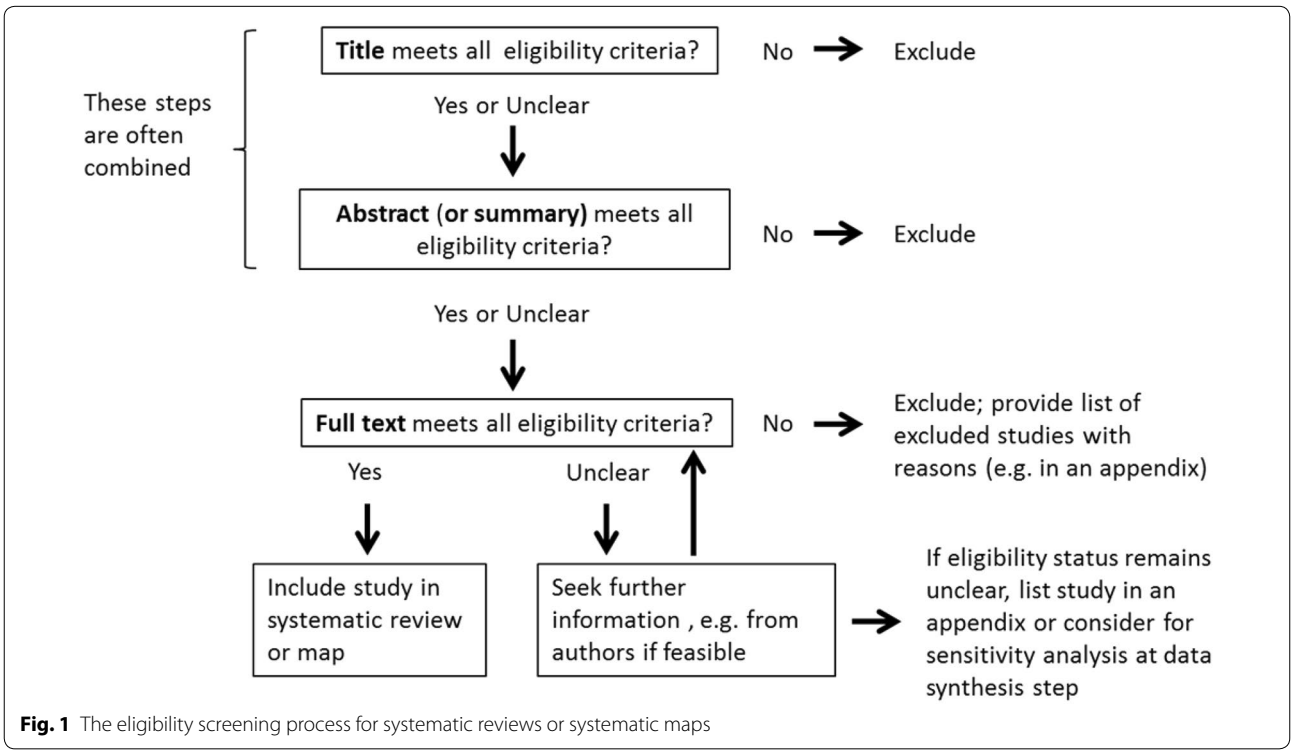

Fig. 1 The eligibility screening process for systematic reviews or systematic maps

It may be difficult to determine whether articles are linked, as related articles do not always cite each other [28, 29] or share common authors [30]. Some 'detective' work (e.g. checking whether the same data appear in more than one article, or contacting authors) may therefore be needed by the review team. Although it would be ideal to identify linked articles that refer to the same study early on the screening process, it may only become clear at the full-text screening stage that articles are linked. Once the links between articles and studies have been identified, a clear record will need to be kept of all articles which relate to each study. This may be done using a separate document or spreadsheet, or using grouping or cross-referencing functions available in bibliographic reference management tools.

Number and expertise of screeners

Eligibility decisions involve judgement and it is possible that errors or bias could be introduced during eligibility screening if the process is not conducted carefully.

Possible problems that could arise at the eligibility screening step are:

- Some articles might be misclassified due to the way members of the review team interpret the information given in them in relation to the eligibility criteria;
- One or more articles might be missed altogether, due to human error;
- Review team members may (knowingly or not) introduce bias into the selection process, since human beings are susceptible to *implicit bias* and experts in a particular topic often have pre-formed opinions about the relevance and validity of articles [3, 31].

Appropriate allocation of the review team to the eligibility screening task, in terms of the number and expertise of those involved, is important to ensure efficiency [10] and can help to minimise the risk of errors or bias. If any members of the review team are authors of articles identified in the searches then the allocation of screening tasks should ensure that members of the review team do not influence decisions regarding the eligibility of their own articles.

Number of screeners

It has been estimated that when eligibility screening is done by one person, on average 8% of eligible studies would be missed, whereas no studies would be missed when eligibility screening is done by two people working independently [32]. The same authors also suggested that use of two reviewers to screen eligibility increased the number of randomised studies identified by 9%. To ensure reliability of the eligibility screening process, articles providing guidance on conducting systematic reviews in environmental research [2, 11, 12] and health research [14, 15, 18] recommend that eligibility screening should be performed where possible by at least two people. The screeners need not necessarily be the same two people for all articles or for all screening steps. Options could be for one person to screen the articles and the second person to then check the first screener's decisions; or both screeners may independently perform the selection process and then compare their decisions. Independent screening is preferable since it avoids the possibility that the second screener could be influenced by the first screener's decision.

The current CEE Guidelines for Systematic Reviews in Environmental Management (version 4.2, March 2013) [1] do not provide recommendations for the number of people who should conduct eligibility screening, although the Guidelines implicitly suggest that a single screener may be acceptable provided that an assessment of screener reliability is conducted. According to the latest CEE evidence synthesis protocols published in Environmental Evidence journal (January–July 2017), screening by a single person, subject to a check of screener reliability using a subset of articles, is the currently practised approach in most cases.

A potential problem with eligibility screening being conducted by a single screener is that any errors in the classification of articles by the screener, or any articles missed from classification, may go undetected, if checking by a second screener is not done on an adequate number of articles. This is why the use of a minimum of two screeners is now considered mandatory in some health research systematic reviews [16, 17]. Reliability checking can be done (e.g. using screener agreement statistics) but this has limitations which should be taken into consideration, as we explain below (see "Assessing screener agreement").

Eligibility screening can be a time-consuming process, typically taking an hour or more for a screener to assess 200 titles or 20 abstracts [10]. If the evidence base is extensive such that large numbers (e.g. tens of thousands) of articles would need to be screened, it might not always be feasible for two or more screeners to work on all screening steps. Consideration may then need to be given as to whether the systematic review or systematic map question, or the eligibility criteria, should be refined (e.g. narrowing the scope) to make the evidence synthesis manageable within the available resources. Discussion with relevant stakeholders, e.g. research commissioners, may be helpful in resolving any difficulties if the level of rigor expected of eligibility screening will be difficult to achieve within the available resources. Employing a single screener at one or more steps of the eligibility screening process, subject to checking screener

reliability, is a pragmatic approach which may be justifiable on a case-by-case basis depending on the nature of the topic and how critical it is to minimise the risk of selection bias [33], but should not be considered as being reflective of best practice (see "Assessing screener agreement" below).

It may be tempting to consider employing a single screener for titles, since the information available in a title is usually relatively limited and titles can often indicate that an article is irrelevant without the need to expend detailed effort in screening [10]. However, selection bias could arise at title screening (just as it could at abstract or full-text screening) if a screener is not impartial, and this could be especially important for evidence syntheses on contentious topics. Furthermore, in our experience it is not uncommon for a small proportion (~1%) of articles to be completely missed from screening by a single reviewer, due to human error (e.g. screener fatigue when assessing thousands of articles). For these reasons, good practice would be to employ a minimum of two screeners at the title screening as well as abstract and full-text screening steps.

For systematic maps the need to minimise selection bias may seem less critical than for systematic reviews, since the output and conclusions of systematic maps are often descriptive. Nevertheless, an underlying expectation of systematic maps is that the searching and eligibility screening steps should be conducted with the same rigor as for systematic reviews [5]. It is therefore good practice in all types of evidence synthesis that at least two people conduct eligibility screening of each article. We recommend that deviations from this should only be made as exceptions, where clear justification can be provided, and is agreed among all relevant stakeholders. This is important for maintaining the integrity of systematic evidence synthesis as a 'gold standard' or 'benchmark' approach for minimising the risk of introducing errors or bias, and to avoid creating confusion as to whether the methods employed in specific evidence syntheses truly constitute those of a systematic review or systematic map, rather than, for example, a traditional literature review or rapid evidence assessment [10].

If a pragmatic decision is made by the review team to proceed with a systematic review or systematic map involving a large number of articles to screen and to use only one screener for some of the articles then, for consistency with good practice as defined above, the following information should be provided in the protocol and final evidence synthesis report:

- a clear justification for using one screener to screen all and a second to screen only a sample, stating which steps of the screening process this will be applied to;
- evidence of the reliability of the approach (i.e. the reliability of the screener's decisions should be tested and reported; see "Assessing screener agreement" below);
- acknowledgement that the use of one screener to screen all and a second to screen only a sample at one or more steps of eligibility screening is a limitation (this should be stated in the conclusions section, critical reflection or limitations section, and, if possible, also in the abstract).

Ultimately, it is the review team's responsibility to ensure that, where possible, methods are used which minimise risks of introducing errors and bias, and that any limitations are justified and transparently reported.

Expertise of screeners

There is no firm 'rule' about how many of the screeners should be topic experts. Given the complexity of environmental topics it is important that the team has adequate expertise in evidence synthesis and the question topic to ensure that important factors relating to the evidence synthesis question are not missed [10]. However, topic experts may lack impartiality as they are likely to be very familiar with the literature relevant to the evidence synthesis question which may risk selective screening decisions being made [31]. A pragmatic approach to reduce the risks of any conflicts of interest within a review team could be to include screeners with different backgrounds and expertise, to ensure diversity of stakeholder perspectives.

Assessing screener agreement

An assessment of agreement between screeners during pilot-testing can help to ensure that the eligibility screening process is reproducible and reliable. If necessary, the eligibility criteria and/or screening process may be modified and re-tested to improve the agreement between screeners. Agreement can be assessed by: recording the observed proportions of articles where pairs of screeners agree or disagree on their eligibility decisions; calculating a reviewer agreement statistic; and/or descriptively tabulating and discussing any disagreements.

A widely used statistic for assessing screener agreement is Cohen's kappa [34], which takes into account the level of agreement between screeners that would occur by chance. But interpretation of kappa scores is subjective since there is no consensus as to which scores indicate 'adequate' agreement, and the concept of 'adequate' agreement is itself subjective. CEE's Guidelines for Systematic Reviews in Environmental Management (version 4.2) [1] suggested a minimum Kappa value of 0.5 should be achieved, which was interpreted as indicating

'substantial agreement'. However, when interpreting screener agreement it should be borne in mind that potentially important discrepancies between screeners can occur even if screener agreement statistics indicate high overall rates of agreement (Box 2).

To assess screener agreement, a sample (as large as possible) of the articles identified in searches should be screened by at least two people and their agreement determined. The size of the sample should be justified by the review team and the articles comprising the subset should be selected randomly to avoid bias towards certain authors, topics, years or other factors.

Use of a kappa statistic to guide pilot-testing of eligibility screening where two or more people will screen each article is a pragmatic approach to optimise efficiency of the process, in which case the limitations of the agreement statistic and its somewhat arbitrary interpretation are not critical. However, recently-published evidence syntheses and protocols indicate that the kappa statistic is increasingly being used for a different purpose: to demonstrate high reviewer agreement in support of employing only one screener to assess the majority of articles. The potential insensitivity of overall screener agreement measures to specific discrepancies in screener agreement (Box 2) suggests that a kappa statistic might not be adequate as a justification that a single screener has sufficient reliability in their screening decisions to protect against the risk of introducing errors or selection bias.

According to the most recently-published protocols, CEE evidence syntheses often assess screener agreement based on a subset of 10% of articles or 100 abstracts (whichever is the larger), although some have used 5% of articles or an unspecified 'small proportion' of articles. These subsets seem rather small, and it could be questioned how a review team would be confident in minimising the risk of selection bias if as many as 90% of articles are not checked. Therefore, we recommend that as large a subgroup of articles as possible is screened by at least two reviewers—the ideal would be 100%.

As there is no consensus on what 'adequate' rates of agreement are (unless reaching 100%), the review team should justify the level of agreement reached and explain in the evidence synthesis report whether relying on a single screener may have led to any relevant studies being excluded. If so, an explanation should be given as to how this would affect interpretation of the evidence synthesis conclusions. Presentation of a decision matrix showing the combinations of screener agreements (e.g. as in Box 2) may be helpful to support any discussion and interpretation of screener reliability.

Box 2: Example of screener agreement interpretation

Screener agreement is illustrated, for two screeners making three possible eligibility decisions (include, exclude or unclear) on 8000 articles. Data are hypothetical but are reflective of a typical evidence synthesis scenario in which the majority of articles identified in searches are excluded during screening. The overall observed agreement between screeners for these data is 99.4% and Cohen's kappa is 0.62.

Screener 1	Screener 2			
	Include	**Exclude**	**Unclear**	**Total**
Include	35	15	3	*53*
Exclude	18	7911	4	*7933*
Unclear	7	5	2	*14*
Total	*60*	*7931*	*9*	*8000*

The data illustrate that, despite good overall agreement as indicated by the observed agreement and the kappa score, discrepancies exist in the include/exclude decisions made by screener 1 and screener 2 which could be critical for a systematic review or systematic map (where the aim should be not to miss any relevant articles). In this example, screener 1 excluded 18 of the 60 articles which screener 2 included (30%), whilst screener 2 excluded 15 of the 53 articles which screener 1 included (28%). At these rates of agreement, employing either screener alone could result in different sets of articles being selected for inclusion.

Resolving disagreements

A process for resolving any disagreements between screeners should be agreed by the review team and to ensure consistency this should be pre-specified in the protocol. An approach which appears to be commonly used [35], and which works efficiently in our experience, is that the screeners meet to discuss their disagreements to reach a consensus; if consensus is not reached a third opinion could then be sought, from another member of the review team or the project advisory group. The exact approach is a matter of preference; for example, abstracts over which there is disagreement could be discussed by the screeners before proceeding to the full-text screening step (to avoid obtaining full-text articles unnecessarily), or the articles could be directly passed to the full-text screening step (to enable decisions to be based on all available information). Records of all screening decisions should be kept to ensure that, if necessary, the review team can justify their study selection. Screening decisions can often be recorded conveniently in user-definable fields in reference management tools. Pilot-testing

the screening process, described below, can be helpful to identify whether some screeners differ systematically from others in the eligibility decisions they make.

Pilot testing the eligibility criteria and process

Pilot testing is important for validating reproducibility and reliability of the method. Pilot testing can:

- Check that the eligibility criteria correctly classify studies;
- Provide an indication of how long the screening process takes, thereby assisting with planning the full evidence synthesis;
- Enable agreement between screeners to be checked; if agreement is poor this should lead to a revision of the eligibility criteria or the instructions for applying them;
- Provide training for the review team in how to interpret and apply the eligibility criteria, to ensure consistency of understanding and application;
- Identify unanticipated issues and enable these to be dealt with before the methods are finalised.

The eligibility screening process should be tested on a sample of articles. There is no firm 'rule' about how many articles should be tested, but the review team will need to satisfy themselves that the eligibility criteria will correctly identify articles that can answer the evidence synthesis question without needing any further amendments. Higgins and Green [3] suggested using around 10–12 articles, including ones which are thought by one screener to be definitely eligible, definitely ineligible, and doubtful, and can be screened by one or more further members of the review team to assess consistency. Pilot testing should be performed for each separate step of the screening process that will be conducted, i.e. the title, abstract (or title plus abstract) and full-text screening steps.

If relevant articles are found to have been excluded, irrelevant articles are included, or a large number of 'unclear' judgements are being made by the review team, then the eligibility criteria should be revised and re-tested until an acceptable discrimination between relevant and irrelevant articles is achieved. The finally-agreed eligibility criteria should then be specified in the evidence synthesis protocol.

Recording and documenting eligibility screening

Reporting the eligibility criteria, screening process and screening results

The final evidence synthesis report may refer to the protocol for a description of most of the methods followed [1]. A concise summary of the eligibility criteria and screening process should also be given in the final report, including in the abstract or summary. An explanation must be provided in the final report if the methods employed differed from those specified in the protocol. It is particularly important to consider whether any changes to the protocol could have introduced errors or bias.

We recommend that, where possible, the final evidence synthesis report should, as a minimum, provide information as specified in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta Analyses) checklists for systematic reviews [36] and for abstracts [37]. The PRISMA checklists are endorsed by over 200 academic journals and by several organisations dedicated to improving standards in evidence synthesis, as well as the Council of Science Editors. Although most of the journals endorsing the PRISMA checklists are currently in the health area (reflecting that this is where the checklists were originally developed), the criteria in the checklists are in principle generic and applicable across disciplines. The PRISMA checklists ensure that key aspects of systematic reviews or systematic maps are reported consistently. In summary, the items in the PRISMA checklists [36, 37] that relate to eligibility screening are:

- Abstract: specify the criteria for inclusion.
- Methods: eligibility: specify the eligibility criteria, giving the rationale.
- Methods: study selection: state the process for selecting studies.
- Results: study selection: give the numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.

It is good practice to include a flow diagram in the evidence synthesis report to show how many unique articles were identified (i.e. after removing any duplicates), and to indicate how many of these were excluded at the title, abstract and full-text screening steps [2, 3, 11, 14, 18, 36]. The flow diagram should also clarify the relationship between articles and studies so that it is clear how many articles and unique studies were included in the systematic review or systematic map; and should give reasons why any studies were excluded at the full-text selection step. A template for a flow diagram based on PRISMA principles [36] is shown in Fig. 2.

The flow diagram template (Fig. 2) may be adjusted to display how the eligibility screening was conducted. For example, the diagram may be expanded to accommodate further panels if titles and abstracts are screened separately. In addition to the flow diagram, a list of the studies which were excluded at the full-text screening step should be provided, indicating the reasons for

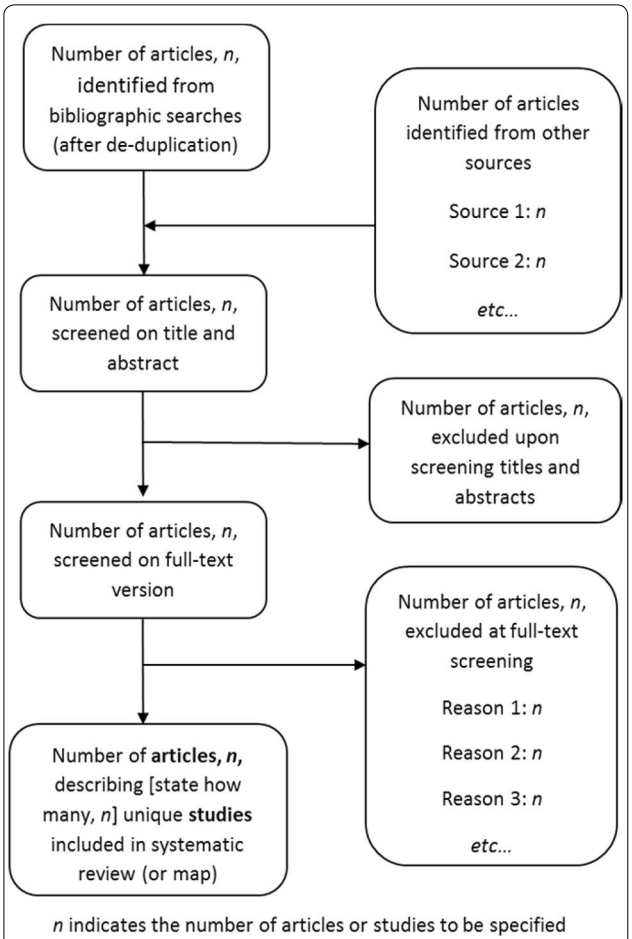

Fig. 2 Flow diagram for reporting the results of eligibility screening

exclusion (e.g. as an appendix to the evidence synthesis report). Whilst the template in Fig. 2 indicates the minimum information on the results of eligibility screening that should be reported, some authors advocate specifying further information. For example, the flow diagram could include an indication of how many of the included studies contributed to any meta-analyses (e.g. [38]), or an indication of how many studies informed quantitative and qualitative analyses for the primary outcome of interest [3, 36].

Any definitions and instructions on interpretation of the eligibility criteria used by the review team should be reported at least in the protocol. Details of the screening process which should be documented in the protocol and also stated concisely in the evidence synthesis report are: the number of screeners involved at each eligibility screening step; whether screening decisions were independent; the expertise of the screeners; the pilot-testing process; any assessments of screener agreement, with justification for the methods chosen; the process employed for resolving any screener disagreements; and how any missing or unclear information was handled.

Any limitations in the eligibility screening process should be mentioned in the Discussion (or Critical Reflection) section of the final evidence synthesis report so that readers can consider them when interpreting the overall findings of the evidence synthesis [36]. If there are any serious limitations in the eligibility screening criteria or the screening process which could affect the overall conclusions of the systematic review or systematic map these should, where possible, also be mentioned in the abstract or summary.

Keeping an archive of screening decisions

It is important that a record is kept of all eligibility screening decisions so that judgements made during conduct of the systematic review or systematic map are transparent and, if necessary, defensible (e.g. if any readers query why a particular study was not included). A record of the screening decisions should be saved (e.g. in a reference management tool or relational database) that can easily be interrogated to display articles which were included, excluded, or deemed unclear at each selection step. The tool or database containing the full set of screening decisions should be archived in such a way that it can be made available if requested by any readers of the systematic review or systematic map report.

Improving current reporting practice

Evidence synthesis protocols published in Environmental Evidence journal during January–July 2017 (n = 11), which we assume reflect current practice, suggest that there are some improvements that could be made in the reporting of eligibility screening in CEE evidence syntheses:

- most of the publications did not specify the number of screeners for one or more of the screening steps;
- most of the publications did not state which of the screening steps their reported screener agreement assessment would be applied to; or whether screener agreement would be assessed for all steps;
- none of the publications provided a justification for the size of the sample of articles to be used in assessing screener agreements; and nearly half (45%) of the publications did not report whether the sample would be selected randomly.

As these aspects of reporting relate to important components of the evidence synthesis methods, review teams should ensure they are fully reported in evidence synthesis protocols and final reports.

Summary and recommendations

The eligibility screening step of a systematic review or systematic map is a well-structured process that determines which evidence will be available for answering a systematic review or systematic map question. Adherence to good practice in eligibility screening reduces the risk of introducing errors or bias into the evidence synthesis. Some parts of the process require judgement, meaning that consistent and transparent reporting of the eligibility criteria and the process for applying them are needed to ensure a clear understanding of how eligibility decisions were made by the review team.

Eligibility screening can be divided into planning, application and reporting phases, although there may be some overlap of these during pilot-testing and protocol development, as iterative improvements are made to the eligibility criteria and screening process. To optimise the efficiency of eligibility screening and minimise the risk of introducing errors or bias, the following approaches for planning, conducting and reporting the eligibility screening step are recommended as good practice.

Planning eligibility screening

- Consider how and whether stakeholders may be involved in the eligibility screening process and how the expertise of the review team will influence decisions; ensure that screeners do not influence eligibility decisions for any articles on which they appear as authors.
- Draft a set of eligibility criteria that reflect the structure of the evidence synthesis question; consider using a standard template specifying the eligibility criteria, with instructions, to ensure that screeners are consistent in their interpretation and application of the criteria.
- Decide how many screeners will conduct eligibility screening at each step, and whether this will be the same for title, abstract and full-text steps (good practice is that at least two screeners conduct each screening step).
- Decide how to assess screener agreement and justify the approach, using as large as possible a randomly-selected sample of references, taking into consideration the potential limitations of screener agreement statistics discussed above.
- Decide how to resolve any screener disagreements and how to handle any ambiguous or missing information.
- If automation of any processes will be employed (e.g. automated de-duplication in reference management software, or text-mining to assist eligibility screening, ensure that the limitations of these approaches are considered and adequate checks for reliability are conducted; it should not be assumed that automated processes will be reliable at identifying or classifying information.
- Pilot-test the eligibility criteria using the specified screening process and if necessary revise and re-test the criteria and/or process to improve efficiency and accuracy;
- Report the final eligibility criteria and screening process in the protocol.

Applying eligibility screening

- Identify and remove duplicates from the search results; if appropriate, follow up any ambiguous or missing information with study authors.
- Apply the protocol-specified eligibility criteria and screening process to titles, abstracts and/or full-text articles, whilst checking for links between articles and studies; resolve any disagreements between screeners.
- Retain a copy of the screening decisions (e.g. in a reference management tool or relational database).

Reporting eligibility screening

Information on eligibility screening should be reported in the protocol and the final evidence synthesis report, as follows.

In the protocol

- State the eligibility criteria and any accompanying instructions that will be provided to screeners on how to apply them.
- Specify the number of screeners intended to conduct screening at each step, with justification.
- Report how any screener agreement assessments will be conducted at each step of the process, with a justification for the size of samples of articles and whether they were selected randomly.
- State the intended processes for handling screener disagreements and any missing or unclear information.

In the final evidence synthesis report

- Include a statement of whether there were any deviations from the protocol in the eligibility criteria or the screening process, with explanations.
- Specify the eligibility criteria, and any relevant instructions used by screeners to interpret them (e.g. in a template, if used).
- State the number of people who conducted screening at each of the title, abstract and full-text screening steps, and whether they worked independently.
- Provide results of any screener reliability assessments that were conducted at each of the title, abstract and full-text screening steps.

- Provide the eligibility screening results, presented in a flow diagram, preferably following PRISMA standards.
- Provide a list (e.g. in an appendix) of all articles excluded at the full-text screening step, giving the reason(s) that each article was excluded.
- Provide a list of any articles which had unclear eligibility status after completion of full-text screening, with explanation why they could not be classified.
- Provide a statement of any limitations of the eligibility criteria or the screening process, including the implications of any deviations from the protocol, and how these would influence the overall conclusions of the evidence synthesis.

Review teams should follow good practice in eligibility screening, to maintain the integrity of systematic review and systematic mapping as robust benchmark approaches for minimising the introduction of errors and bias in evidence synthesis. Quicker and cheaper ways to conduct evidence synthesis, such as 'quick scoping reviews' and 'rapid evidence assessments' [10] are increasingly in demand, driven by the needs both of review teams and end-users of evidence syntheses [33]. The detailed recommendations we provide in the current paper serve to define good practice if the aim is to minimise risks of introducing errors and bias in evidence synthesis. If the requirement is to conduct a more rapid synthesis whilst being as rigorous as possible then our recommendations may serve to assist pragmatic, transparent, discussions as to which part(s) of the eligibility screening methods could be adjusted to expedite more rapid syntheses.

Recently published evidence synthesis protocols suggest that there may be opportunities to improve current practice in eligibility screening and reporting, and we have considered this in the recommendations presented above. In particular, we suggest review teams should consider carefully: whether employing fewer than two screeners would adequately protect against the risk of introducing errors or bias; and the possible limitations of reviewer agreement statistics when applied in eligibility screening.

Finally, we acknowledge that evidence synthesis is a dynamic area in which new methods are emerging. In particular, automated approaches such as text mining and machine learning appear to offer considerable promise for making eligibility screening quicker and less onerous [20]. If using automated approaches, review teams will need to demonstrate that the eligibility screening process is reliable and does not put systematic evidence syntheses at increased risk of errors or bias. A basic principle underpinning our recommendations is that studies should not be selectively excluded from evidence synthesis, and this would apply irrespective of whether eligibility screening is conducted by human beings and/or machine processes.

Authors' contributions
GF drafted the manuscript. GF, BL and GP read, commented on, and revised the manuscript. All authors read and approved the final mansucript.

Author details
[1] Southampton Health Technology Assessments Centre (SHTAC), Faculty of Medicine, University of Southampton, Southampton, UK. [2] Fondation pour la Recherche sur la Biodiversité, Paris, France. [3] Department of Zoology, Biodiversity Institute, University of Oxford, Oxford, UK.

Acknowledgements
We thank the Editor and three anonymous reviewers for their constructive comments on the submitted manuscript.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Consent for publication
Not applicable.

Ethics approval
Not applicable.

Funding
CEE kindly provided travel support and Oxford Martin School, University of Oxford, kindly provided a room, to support a 2-day workshop in which all authors discussed and revised the manuscript.

Publisher's Note

Received: 4 January 2017 Accepted: 30 August 2017
Published online: 20 September 2017

References

1. CEE (Collaboration for Environmental Evidence). Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2. 2013. http://www.environmentalevidence.org/wp-content/uploads/2014/06/Review-guidelines-version-4.2-final.pdf.
2. EFSA (European Food Safety Authority). Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA J. 2010;8(6):1637.
3. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Chichester: Wiley-Blackwell; 2011.
4. Petticrew M, Roberts H. Systematic reviews in the social sciences. A practical guide. Oxford: Blackwell; 2006.
5. James KL, Randall NP, Haddaway NR. A methodology for systematic mapping in environmental sciences. Environ Evid. 2016;5:7.
6. Bayliss HR, Beyer FR. Information retrieval for ecological syntheses. Res Synth Methods. 2015;6(2):136–48.
7. Livoreil B, Glanville G, Haddaway NR, Bayliss H, Bethel A, Flammerie de la Chapelle F, Robalino S, Savilaakso S, Zhou W, Petrokofsky G, Frampton G. Systematic searching for environmental evidence using multiple tools and sources. Environ Evid. 2017;6:23.

8. Roe D, Booker F, Day M, Zhou W, Allebone-Webb S, Hill NAO, Kumpel N, Petrokofsky G, Redford K, Russell D, Shepherd G, Wright J, Sunderland TCH. Are alternative livelihood projects effective at reducing local threats to specified elements of biodiversity and/or improving or maintaining the conservation status of those elements? Environ Evid. 2015;4:22.
9. Aiassa E, Higgins JPT, Frampton GK, Greiner M, Alfonso A, Amzal B, Deeks J, Dorne JL, Glanville J, Lovei GL, Nienstedt K, O'Connor AM, Pullin A, Rajic A, Verloo D. Applicability and feasibility of systematic review for performing evidence-based risk assessment in food and feed safety. Crit Rev Food Sci Nutr. 2015;55:1026–34.
10. DEFRA (Department for Environment Food and Rural Affairs). Emerging tools and techniques to deliver timely and cost effective evidence reviews. London: DEFRA; 2015.
11. Rooney AR, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. Systematic review and evidence integration for literature-based environmental health science assessments. Environ Health Perspect. 2014;122(7):711–8.
12. Sargent JM, O'Connor AM. Conducting systematic reviews of intervention questions II: relevance screening, data extraction, assessing risk of bias, presenting the results and interpreting the findings. Zoonoses Public Health. 2014;61(suppl 1):39–51.
13. Campbell Collaboration. Systematic reviews: policies and guidelines version 1.2. Oslo: The Steering Group of the Campbell Collaboration; 2014.
14. CRD (Centre for Reviews and Dissemination). Systematic reviews. CRD's guidance for undertaking reviews in health care. York: University of York; 2009.
15. Higgins JPT, Deeks JJ. Selecting studies and collecting data. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Chichester: Wiley-Blackwell; 2011. p. 151–86.
16. Higgins JPT, Lasserson T, Chandler J, Tovey D, Churchill R. Methodological expectations of Cochrane intervention reviews (MECIR). London: Cochrane; 2016.
17. Institute of Medicine. Finding what works in health care. Standards for systematic reviews. Washington DC: Institute of Medicine of the National Academies; 2011.
18. McDonagh M, Peterson K, Raina P, Chang S, Shekelle P. Avoiding bias in selecting studies—methods guide for comparative effectiveness reviews. Rockville: Agency for Healthcare Research and Quality (AHRQ); 2013.
19. Social Science Research Unit. Eppi Reviewer 4. London: Social Science Research Unit, Institute of Education, University of London; 2016. https://eppi.ioe.ac.uk/cms/Default.aspx?alias=eppi.ioe.ac.uk/cms/er4. Accessed July 2017.
20. Rathbone J, Hoffman T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev. 2015;4:80.
21. Tramer MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. BMJ. 1997;315:635–40.
22. Englund G, Sarnelle O, Cooper SD. The importance of data-selection criteria: meta-analyses of stream predation experiments. Ecology. 1999;80(4):1132–41.
23. Rodriguez LG, Hogarth NJ, Zhou W, Xie C, Zhang K, Putzel L. China's conversion of cropland to forest program: a systematic review of the environmental and socioeconomic effects. Environ Evid. 2016;5:21.
24. Ojanen M, Zhou W, Miller DC, Nieto SH, Mshale B, Petrokofsky G. What are the environmental impacts of property rights regimes in forests, fisheries and rangelands? Environ Evid. 2017;6:12.
25. Choi WS, Song SW, Ock SM, Kim CM, Lee J, Chang WJ, Kim SH. Duplicate publication of articles used in meta-analysis in Korea. Springer Plus. 2014;3:182.
26. von Elm E, Tramer MR, Jüni P, Egger M. Does duplicate publication of trials introduce bias in systematic reviews? A systematic review [abstract]. In: 11th Cochrane Colloquium: 2003 Oct 26–31; Barcelona, Spain.
27. von Elm E, Poglia G, Walder B, Tramer MR. Different patterns of duplicate publication: an analysis of articles used in systematic reviews. JAMA. 2004;291:974–80.
28. Bailey BJ. Duplicate publication in the field of otolaryngology-head and neck surgery. Arch Otolaryngol. 2002;126:211–6.
29. Barden J, Edwards JE, McQuay HJ, Moore RA. Oral valdecoxib and injected parecoxib for acute postoperative pain: a quantitative systematic review. BMC Anesthesiol. 2003;3:1.
30. Gøtzsche P. Multiple publication of reports of drug trials. Eur J Clin Pharmacol. 1989;36:429–32.
31. Gøtzsche PC, Ioannidis JPA. Content area experts as authors: helpful or harmful for systematic reviews and meta-analyses? BMJ. 2012;345:e7031.
32. Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Stat Med. 2002;21:1635–40.
33. Langer L, Erasmus Y, Tannous N, Stewart R. How stakeholder engagement has led us to reconsider definitions of rigour in systematic reviews. Environ Evid. 2017;6:20.
34. Altman DG. Measuring agreement. In: Altman DG, editor. Practical statistics for medical research. London: Chapman and Hall; 1991.
35. Petersen K, bin Ali N. Identifying strategies for study selection in systematic reviews and maps. Proceedings of the 5th International Symposium on Empirical Software Engineering and Measurement (ESEM), Banff, AB, Canada, September 2011.
36. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Goetzsche PC, Ioannidis JPA, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.
37. Beller EM, Glasziou PP, Altman DG, Hopewell S, Bastian H, Chalmers I, et al. PRISMA for abstracts: reporting systematic reviews in journal and conference abstracts. PLoS Med. 2013;10(4):e1001419. doi:10.1371/journal.pmed.1001419.
38. Sagoo GS, Little J, Higgins JPT. Systematic reviews of genetic association studies. PLoS Med. 2009;6(3):e1000028.