Title: Recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER)

Author List

Corresponding and first author: **Paul Whaley**, Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK | p.whaley@lancaster.ac.uk

Other authors:
Elisa Aiassa, European Food Safety Authority (EFSA), Assessment and Methodological Support unit. Via Carlo Magno 1/A, 43126 Parma, Italy | elisa.aiassa@efsa.europa.eu
Claire Beausoleil, ANSES (French Agency for Food, Environmental and Occupational Health Safety), Risk Assessment Department, Chemical Substances Assessment Unit, F-94700 Maisons-Alfort, France | claire.beausoleil@anses.fr
Anna Beronius, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden | anna.beronius@ki.se
Gary Bilotta, School of Environment and Technology, University of Brighton, Brighton, UK | g.s.bilotta@gmail.com
Alan Boobis National Heart & Lung Institute, Imperial College London, London, UK | a.boobis@imperial.ac.uk
Rob de Vries, SYRCLE, Department for Health Evidence, Radboud Institute for Health Sciences, Radboudumc, Nijmegen, The Netherlands | rob.devries@radboudumc.nl
Annika Hanberg, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden | annika.hanberg@ki.se
Sebastian Hoffmann, Evidence-based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health, Paderborn, Germany | sebastian.hoffmann@seh-cs.com
Neil Hunt, Yordas Group, Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK | n.hunt@yordasgroup.com
Carol F. Kwiatkowski, The Endocrine Disruption Exchange. P.O. Box 54, Eckert, CO, 81418, USA | kwiatkowskicf@gmail.com
Juleen Lam, University of California, San Francisco and California State University, East Bay. 28500 Carlos Bee Blvd Room 502, Hayward, CA 94542, USA | Juleen.Lam@csueb.edu
Steven Lipworth, Royal Society of Chemistry, Burlington House, Piccadilly, London, W1J 0BA, UK
Olwenn Martin, Institute for the Environment, Health and Societies, Brunel University London, Uxbridge, UK | olwenn.martin@brunel.ac.uk
Nicola Randall, Harper Adams University, Newport, Shropshire, UK | nrandall@harper-adams.ac.uk
Lorenz Rhomberg PhD ATS, Gradient, 20 University Road, Cambridge, MA 02138, USA lrhomberg@gradientcorp.com
Andrew A. Rooney, Division of the National Toxicology Program, National Institute of Environmental Health Sciences, North Carolina, USA | andrew.rooney@nih.gov
Holger J. Schünemann, McGRADE Centre and Michael G De Groote Cochrane Canada Centre, Dept. of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main Street West, Hamilton, ON, Canada | schuneh@mcmaster.ca
Daniele Wikoff, ToxStrategies, 31 College Place, Suite B118B, Asheville, NC 28801, USA | dwikoff@toxstrategies.com
Taylor Wolffe, Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK | t.wolffe@lancaster.ac.uk
Crispin Halsall, Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK | c.halsall@lancaster.ac.uk

Declaration of Interests

Due to the objective of the project being to establish, across a wide range of stakeholders, a consensus view on sound and good practice in the conduct of environmental health systematic reviews, participants in the process were selected because of their varying interests in the conduct of environmental health research. Funding was provided by Lancaster University to support travel costs of authors who would otherwise be unable to attend (PW, CH, LR, JL, AR) and Dr Jennifer McPartland (non-authoring workshop participant, see acknowledgements). With regard to the development of COSTER, the authors declare they have no apparent competing financial interests, and certify that their freedom to design, conduct, interpret, and publish the research was not compromised by any controlling sponsor. PW, as organiser of the meeting and lead author of the manuscript, declares personal fees from Elsevier Ltd (*Environment International*), the Cancer Prevention and Education Society, the Evidence Based Toxicology Collaboration, Yordas Group, and grants from Lancaster University, which are outside the submitted work but relate to the development and promotion of systematic review and other evidence-based methods in environmental health research, delivering training around these methods, and providing editorial services. Each author has declared their interests using the International Committee of Medical Journal Editors Form for Disclosure of Potential Conflicts of Interest; these are available as Supplemental Materials. The manuscript has been handled by *Environment International* according to Elsevier's conflict of interest policy.

Highlights

- COSTER is a set of recommendations for conduct of environmental health systematic reviews (EH SRs)
- COSTER's consensus process covered NGOs, academia, industry and government agencies
- COSTER is intended for use by scientists, editors, agencies, and others seeking guidance on EH SRs
- COSTER is put forward to initiate broader discussion of the EH community's expectations for SRs
- COSTER is a first step in defining robust standards for good practice in conduct of EH SRs

Abstract

Background: There are several standards that offer explicit guidance on good practice in systematic reviews (SRs) for the medical sciences; however, no similarly comprehensive set of recommendations has been published for SRs that focus on human health risks posed by exposure to environmental challenges, chemical or otherwise.

Objectives: To develop an expert, cross-sector consensus view on a key set of recommended practices for the planning and conduct of SRs in the environmental health sciences.

Methods: A draft set of recommendations was derived from two existing standards for SRs in biomedicine and developed in a consensus process, which engaged international participation from government, industry, non-government organisations, and academia. The consensus process consisted of a workshop, follow-up webinars, email discussion and bilateral phone calls.

Results: The Conduct of Systematic Reviews in Toxicology and Environmental Health Research (COSTER) recommendations cover 70 SR practices across eight performance domains. Detailed explanations for specific recommendations are made for those identified by the authors as either being novel to SR in general, specific to the environmental health SR context, or potentially controversial to environmental health SR stakeholders.

Discussion: COSTER provides a set of recommendations that should facilitate the production of credible, high-value SRs of environmental health evidence, and advance discussion of a number of controversial aspects of conduct of EH SRs. Key recommendations include the management of conflicts of interest, handling of grey literature, and protocol registration and publication. A process for advancing from COSTER's recommendations to developing a formal standard for EH SRs is also indicated.

1. Introduction

In the fields of toxicology, epidemiology, environmental health and chemical risk assessment (henceforth abbreviated as "environmental health (EH) research"), systematic reviews (SRs) are increasingly conducted and used by academics, non-government organisations, industry and regulators to characterise health hazards and risks posed by exposure to environmental challenges (Whaley et al. 2016). One of the drivers of this growing interest is increasing recognition of the potential for systematic methods to offer a new benchmark in best practice for aggregating and summarising evidence in support of policy decisions (EFSA 2010; Rooney et al. 2014; NAS 2017, 2014; Stephens et al. 2016).

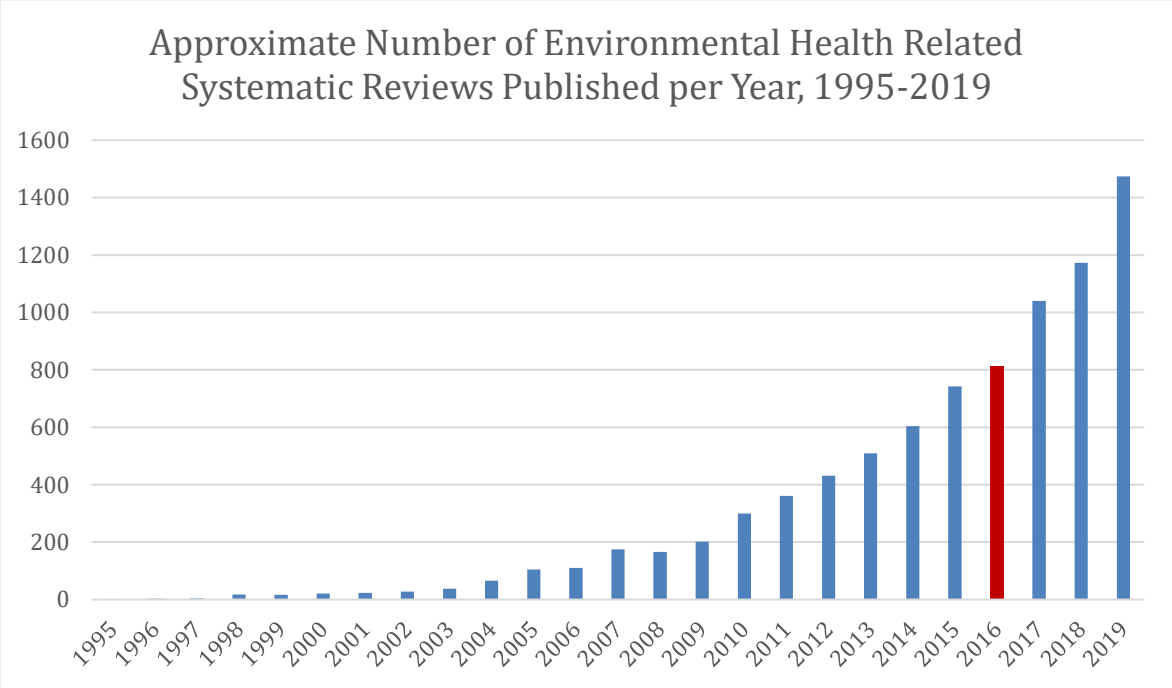

Figure 1: Chart showing annual increase in number of publications on topics related to EH research with the term "systematic review" in the title, indexed in Web of Science. The total number of publications approximately doubled between 2016 and 2020. **Search**: TITLE: ("systematic review"), **Refined by**: WEB OF SCIENCE CATEGORIES: (PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH OR TOXICOLOGY) AND [excluding] WEB OF SCIENCE CATEGORIES: (PHARMACOLOGY PHARMACY), **Timespan**: All years (1995-2019 shown). **Indexes**: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC. **Date of search**: 4 February 2020

In service of this interest, there is a burgeoning number of documents which purport to provide varying types of guidance for conducting SRs in EH research. These include, for example: a US agency handbook (NTP OHAT 2019); US and EU guidance documents (Schaefer and Myers 2017; EFSA 2015; EPA 2018); Instructions to Authors (IARC 2019b, 2019a); and general frameworks (Vandenberg et al. 2016; Woodruff and Sutton 2014).

The challenge for the reader is in how SR guidance documents vary in their levels of comprehensiveness and detail, domains of applicability, the extent to which they have been tested and validated, and what they define (either implicitly or explicitly) as being essential SR methodology. For example, the US National Toxicology Program Office of Health Assessment and Translation (NTP OHAT) handbook is for SRs conducted in support of hazard assessment within a US regulatory framework (Rooney et al. 2014; NTP OHAT 2019), whereas the Navigation Guide Framework (Woodruff and Sutton 2014) is intended for a more general research context. While the Navigation Guide and NTP OHAT approaches are largely similar (with steps including development of a protocol, comprehensive search strategies, employment of a Cochrane-derived risk of bias approach to appraising study quality, and use of a GRADE-based approach to assessing confidence in a body of evidence) there are some differences between the two. Other approaches have larger differences. For example, the SYRINA framework (Vandenberg et al. 2016) lays out a wide range of options for SR teams to choose from, and a draft SR-based risk assessment methodology for the US Toxic Substances Control Act (EPA 2018) scores study quality rather than implementing Cochrane guidance on risk of bias assessment (Singla et al. 2019). Others differ in their use of protocols, their approach to critical appraisal of included studies, and their methods for assessing certainty in the evidence. Furthermore, some EH SR guidance documents are intended to apply to the entire environmental health risk assessment process, while others focus on a particular stage of it. Many SR guidance documents have also been developed for specific purposes and are not necessarily intended to represent a broader community view of general good practice. Overall, these documents do not provide a collectively consistent, general overview of good practice in the planning and conduct of EH SRs.

The development and promulgation of clear, expert guidance on good practice is considered by institutions including the US Institute of Medicine to be an important contributor to ensuring the quality of biomedical SRs (Eden et al. 2011). The potential value of developing such guidance specific to EH SRs was recognised in a 2014 expert workshop on applying SR methods to chemical risk assessment. Among other strategic proposals, the workshop recommended "development of a recognised 'gold standard' for SRs in toxicology and risk assessment […] to address the growing number of purported SRs of unclear validity which are increasingly prevalent in the environmental health literature" (Whaley et al. 2016).

A broad cross-section of relevant stakeholders was therefore convened, with the objective of developing a comprehensive set of recommendations for the planning and conduct of SRs in EH research. These recommendations are based on standard practices and processes for conduct of

SRs in other fields, and put forward to initiate broader discussion as to what the EH community's collective expectations for SR methods ought to be.

2. Methods

A workshop was held on 2 December 2016, attended by 31 participants from academic, policy, regulatory, non-government and industry backgrounds (see Supplemental Information 01). Participants were prioritised for invitation to the workshop from an initial longlist of 62 drawn up by PW and CH, based on a mixture of having a publishing history demonstrating at least some experience in systematic review or the principles thereof, professional reputation, economic sector, and word-of-mouth recommendation. An overall balance of expertise in SR, weight-of-evidence methods, chemical risk assessment, toxicology, epidemiology, environmental health research and chemicals policy was sought across the final group of participants, along with balanced representation from each stakeholder group including a target of at least two NGO participants. Lancaster University provided £5000 to facilitate balanced participation, covering travel costs for participants who would not otherwise be able to attend the workshop.

The recommendations for good practice were developed using a consensus methodology. "Consensus" was defined following the terminology of the International Organization for Standardization (ISO) as "general agreement, characterized by the absence of sustained opposition to substantial issues by any important part of the concerned interests and by a process that involves seeking to take into account the views of all parties concerned and to reconcile any conflicting arguments" (ISO/IEC 2004).

The consensus process was seeded by two discussion documents drafted by PW (see Supplements 02 and 03). A draft set of recommendations (Supplement 03), initially given the working title of "ECOSYS-CRA" before being renamed "COSTER", was created by combining version 2.3 of the Cochrane *MECIR* standards (Chandler et al. 2013) with the US Institute of Medicine *What Works in Health Care: Standards for Systematic Reviews* (Eden et al. 2011), henceforth referred to as "MECIR" and "IOM" respectively. The MECIR and IOM standards were taken to already represent a high degree of consensus and expectation of effectiveness of sound-practice requirements relating to general SR methods in biomedicine, thereby providing a solid basis for interpretation into a set of recommendations for EH SRs.

The draft recommendations were discussed element-by-element at the workshop by two break-out groups working in parallel, chaired by PW and JL. Feedback was solicited on four areas. (a) Which of the proposed elements would constitute "sound and good practice" for EH

SRs, and should therefore be included in a final set of recommendations? (b) Should any of the included elements be reformulated for the EH SR context, and if so, how? (c) Were there any additional elements that should be included for the EH SR context and, if so, how should they should be formulated? (d) Were there questions for clarification and follow-up? Further detail on the assumptions, methodological decisions, and structure of the consensus process behind COSTER is provided in Supplement 02.

GB and CH took notes of the discussion in each group. Comments were collated into a redrafted document and, in response to a request by workshop participants, cross-checked by PW against the Campbell Collaboration *MEC2IR* standard (Campbell Collaboration 2014). This was to check for any further possible elements that might be included as recommendations in COSTER. The COSTER recommendations were then discussed in a series of six one-hour webinars held between January and June 2017, chaired by PW and attended on average by six participants (EA, ABe, RdV, KG, AH, NH, SH, CK, JL, OM, LR, AR, HS, KS, DW, CH, TW participated in at least one). The webinars were followed by email exchanges and bilateral phone calls between PW and various authors to finalise wording and agree that consensus had been reached.

The consensus process was closed by PW on 24 January 2018; participating authors confirmed agreement with the consensus by signing off as co-authors of this manuscript. Non-authoring contributors are listed in the Acknowledgements.

The manuscript went through three rounds of journal peer-review, during which the framing and implications of COSTER as a consensus process and resulting set of good-practice recommendations were revised and clarified. The most significant change was the reframing of COSTER from a “code of practice” to a set of recommendations. While the process followed in COSTER was intended to emulate formal standardisation processes, the peer-reviewers suggested the authors were potentially over-reaching in describing what they had achieved, and that the formal language of standardisation was an impediment to communication of the core messages of the manuscript. The authors therefore removed reference to formal standards, instead presenting COSTER as a set of recommendations for good practice. The COSTER recommendations themselves, as they were the result of the consensus process, were not changed in peer-review. For transparency, previous versions of the manuscript are archived on Zenodo.org (Whaley et al. 2019).

3. Results

COSTER presents 70 recommendations for good practice in the conduct of EH SRs, distributed across 8 steps of the SR process. If followed, the recommendations should result in a EH SR having the following three characteristics which are considered, in the opinion of the authors, as critical for the scientific quality of EH SR projects:

1. *Utility*: addressing an important research question and advancing community understanding of an environmental health issue via a methodology of synthesising existing research;

2. *Transparency*: encouraging comprehensive consideration of the assumptions and methods employed in a SR such that, if they are adequately reported, a reader is able to appraise the validity of the SR's findings and assess their relevance to a given decision-making context;

3. *Credibility*: minimising the risk that a SR's findings are biased either by limitations in the evidence base itself or in the processes used to locate and synthesise that evidence.

The eight COSTER domains cover the following methodological elements of the SR process: planning the SR; searching for evidence; selecting evidence for review; extracting data; critically appraising each individual included study; synthesising the evidence; interpreting the evidence and summarising what it means for the review question; and drawing conclusions (see Figure 2). The recommendations within each domain are listed in Table 1. An explanation of key recommendations is provided in Table 2, and guidance on how to use COSTER is presented in the Discussion section of this manuscript.

In total, 20 of the 31 workshop participants, plus TW, signed off as a manuscript author. Eight participants did not participate in the consensus process beyond the workshop; they were not asked why, but when reasons were given they related to restrictions imposed by the governance policies of employing organisations in relation to employees' endorsement of guidance documents, or a lack of personal capacity to contribute to a lengthy process of discussion and manuscript development. Only one participant who was involved in the development of the manuscript itself ultimately felt they could not sign off as an author, citing differences between COSTER and the official policies of the organisation with which they were affiliated, and the potential for confusion that might cause if their authorship was misinterpreted as organisational endorsement. None of the participants opposed publication of COSTER.

Figure 2. Conceptual structure of COSTER with objectives for each stage of the SR process

COSTER v1.0.0: Recommendations for the planning and conduct of environmental health systematic reviews
1. Planning the Review and Preparing the Protocol
1.1 Securing capacity, competencies and tools
1.1.1 Ensure the review team has sufficient combined competence to conduct the systematic review, including relevant expertise in: information science (for e.g. search strategies); evidence appraisal; statistical methods; domain or subject expertise; systematic review methods.
1.1.2 Identify information management practices for each stage of the review, including reference and knowledge management tools, systematic review software, and statistics packages.
1.1.3 Exclude people or organisations with apparent conflicts of interest relating to the findings of the review from analysis and decision-making roles in the review process.
1.1.4 Disclose the roles and all potential conflicts of interest of all people and organisations involved in planning and conducting the review, including all providers of financial and in-kind support.
1.2 Setting the research question to inform the scope of the review ("problem formulation")
1.2.1 Demonstrate the need for a new review in the context of the scientific value of the question, the importance to stakeholders of the question being asked, and the findings of any pre-existing primary research and/or evidence syntheses.
1.2.2 Articulate the scientific rationale for each question via development of a theoretical framework which connects e.g. the exposure to the outcomes of interest (or otherwise as appropriate given the objectives of the review).
1.2.3 For each research question to be answered by the review, prospectively define a statement of the research objective in terms of one or more of the following components, selected as appropriate: ● Population (objects of investigation, i.e. the entities to which exposures or interventions happen) ● Exposure or Intervention (the administered change in conditions of the objects of investigation, to include timing, duration and dose) ● Comparator (the group to which the intervention or exposure groups are being compared) ● Outcome (the change being measured in the intervention or exposure group) ● Study design (specific design features of relevant research) ● Target condition (the object of a test method for diagnosis or detection)

1.3 Defining eligibility criteria
1.3.1 Define and justify unambiguous and appropriate eligibility criteria for each component of the objective statement.
1.3.2 Define the points at which screening for eligibility will take place (e.g. pre-screening based on title/abstract, full text screening, or both)
1.3.3 For interventions, exposures and comparators: define as relevant to review objectives the eligible types of interventions and/or exposures, methods for measuring exposures, the timing of the interventions/exposures, and the interventions/exposures against which these are to be compared.
1.3.4 For outcomes: define as relevant to review objectives the primary and secondary outcomes of interest (including defining which are apical and which are intermediate), what will be acceptable outcome measures (e.g. diagnostic criteria, scales) and the timing of the outcome measurement.
1.3.5 For study designs: define eligible study designs per design features rather than design labels.
1.3.6 Include all relevant, publicly-available evidence, except for research for which there is insufficient methodological information to allow appraisal of internal validity.
1.3.7 Include evidence which is relevant to review objectives irrespective of whether its results are in a usable form.
1.3.8 Include relevant evidence irrespective of language.
1.3.9 Exclude evidence which is not publicly available.
1.4 Planning the review methods at protocol stage
1.4.1 Design sufficiently sensitive search criteria, so that studies which meet the eligibility criteria of the review are not inadvertently excluded.
1.4.2 Design “characteristics of included studies” table.
1.4.3 Define the risk of bias assessment methods to be used for evaluating the internal validity of the included research. If observational studies are included, this should cover identification of plausible confounders.
1.4.4 Design the methods for synthesising the included studies, to cover: qualitative and quantitative methods (with full consideration given to synthesis methods to be used when meta-analysis is not possible); assessment of heterogeneity; choice of effect measure (e.g. RR, OR etc.); methods for meta-analysis and other quantitative synthesis; pre-defined, appropriate effect modifiers for sub-group analyses.
1.4.5 Define the methods for determining how, given strengths and limitations of the overall body of evidence, confidence in the results of the synthesis of the evidence for each outcome is to be captured and expressed. (For reviews which include multiple streams of evidence, this may need to be defined for each stream.)
1.4.6 For reviews which include multiple streams of evidence (e.g. animal and human studies), define the methods for integrating the individual streams into an overall result. This should include a description of the relative relevance of populations (e.g. species, age, comorbidities etc.), exposures (e.g. timing, dose), and outcomes (direct or surrogate, acute or chronic model of disease, etc.), as appropriate, per which inferences about predicted effects in target populations can be made from observed effects in study populations.
1.4.7 Pilot-test all components of the review process in which reviewer performance could affect review outcomes. This includes the design and usability of the data extraction form/s, and the conduct of the risk of bias assessment.

1.5 Publishing the protocol
1.5.1 Create a permanent public record of intent to conduct the review (e.g. by registering the protocol in an appropriate registry) prior to conducting the literature search.
1.5.2 As appropriate for review planning and question formulation, secure peer-review and public feedback on a draft version of the protocol, incorporating comments into the final version of the protocol.
1.5.3 Publish the final version of the protocol in a public archive, prior to screening studies for inclusion in the review.
1.5.4 Clearly indicate in the protocol and review report any changes in methods made after testing or conduct of any steps of the review.
2. Searching for Evidence
2.1 Search all the key scientific databases for the topic, including national, regional and subject-specific databases.
2.2 Define reproducible strategies for identifying and searching sources of grey literature (databases, websites etc.).
2.3 Structure search strategies for each database, electronic and other source, using appropriate controlled vocabulary, free-text terms and logical operators in a manner which prioritises sensitivity.
2.4 Search within the reference lists of included studies and other reviews relevant to the topic (“hand-searching”) and consider searching in the reference lists of documents which have cited included studies.
2.5 Search by contacting relevant individuals and organisations.
2.6 Document the search methods and results in sufficient detail to render them transparent and reproducible.
2.7 Re-run all searches and screen the results for potentially eligible studies within 12 months prior to publication of the review (screening at least at the level of title plus abstract). In deciding whether to incorporate new studies in the review, the importance of a possible change in results should be weighed against any delay in publication. Potentially eligible studies which have not been incorporated should be listed as “awaiting classification”.
3. Screening Evidence for Inclusion
3.1 Screening of each piece of evidence for inclusion to be conducted by at least two people working independently, with an appropriate process (e.g. third-party arbitration) for identifying and settling disputes.
3.2 Document decisions in enough detail to allow presentation of the results of the screening process in a PRISMA flow chart.
3.3 Studies which are excluded after assessment of full text should be listed in a table of excluded studies along with the reason for their exclusion (one reason is sufficient).
3.4 Do not exclude multiple reports of the same research (e.g. multiple publications, conference abstracts etc.); instead collate the methodological information from each of the reports as part of the data extraction process for each unit of evidence.
4. Extracting Relevant Data from Included Study Reports
4.1 Collect characteristics of the included studies in sufficient detail to populate the planned "characteristics of included studies" table.

4.2 Extraction of study characteristics and outcome data to be conducted by at least two people working independently with an appropriate process (e.g. third-party arbitration) for identifying and settling disputes.
4.3 Assessment of risk of bias to be conducted separately from data extraction. Ideally, and where appropriate, risk of bias assessment should be conducted between extraction of study characteristics and extraction of outcome data (study results).
4.4 Correct for errors and omissions in data reported in included studies by: (1) collecting the most detailed numeric data possible; (2) examining relevant retraction statements and errata for information; (3) obtaining where possible relevant unpublished data which is missing from reports and studies.
4.5 Check accuracy of the numeric data in the meta-analysis utilising an appropriate process (e.g. third-party control).
5. Appraising the Internal Validity of Included Studies
5.1 Appraise internal validity of each included study via the risk of bias assessment methodology specified in the protocol.
5.2 Assess risk of bias per outcome or outcome-exposure pair (as appropriate) rather than per study.
5.3 Risk of bias assessment is to be conducted by at least two people working independently, with an appropriate process (e.g. third-party arbitration) for identifying and settling disputes.
5.4 Apply the risk of bias assessment tool thoroughly and consistently to each included study, recording each risk of bias judgement made by each reviewer, and any disagreements and how they were resolved.
5.5 If there is empirical evidence which supports a judgement, comment but do not guess on likely direction and (if possible) magnitude of effect of bias.
5.6 Provide appropriate explanation for judgement of risk of bias, making reference to decision processes described in the protocol, and using supporting quotes from study reports or noting if information was not available.
6. Synthesising the Evidence / Deriving Summary Results
6.1 Undertake (or display) meta-analyses only when studies are sufficiently comparable as to render the combined result meaningful.
6.2 Transform all scales (where appropriate) into common measures of outcome, explaining how each scale has been reinterpreted in the review.
6.3 Use appropriate methods to assess the presence and extent of between-study variation (statistical heterogeneity) when undertaking a meta-analysis.
6.4 If important statistical heterogeneity is observed, explain how this is accommodated in developing appropriate summary results for the review (e.g. by not pooling at all, by conducting subgroup analyses etc.)
6.5 Assess the potential for publication bias in the data (i.e. systematic differences between the evidence which was accessible to the review, and the evidence which was not).
6.6 Assess potential impact of risk of bias in the synthesis, based on the results of the appraisal of risk of bias in the included studies (e.g. sub-group analysis excluding studies at high risk of bias; appropriate qualitative or quantitative approaches).

6.7 Test the robustness of the results using sensitivity analyses (such as the impact of notable assumptions, imputed data, borderline decisions and studies at high risk of bias).
6.8 If subgroup analyses are conducted, follow the subgroup analysis plan specified in the protocol, avoiding over-interpretation of any particular findings; sensible post-hoc analyses may also be carried out.
7. Interpreting Results
7.1 Interpret the internal validity of the overall body of evidence by considering results of the appraisal of internal validity (risk of bias) of each included study. The review should describe the potential for biased summary results due to limitations in study design and conduct (e.g. extent of randomisation, blinding, confounding etc.) and the implications of these limitations for drawing conclusions based on the overall body of evidence.
7.2 Interpret the consistency of the overall body of evidence, accounting for explainable and unexplainable variation between studies. If a meta-analysis has been conducted, consider statistical heterogeneity. Where appropriate, conduct sub-group and sensitivity analyses.
7.3 Interpret any subgroup analyses without selective reporting of results or placing undue emphasis on specific findings.
7.4 Interpret the precision of the results of any syntheses, taking care to interpret statistically non-significant results as findings of uncertainty rather than no effect, unless the confidence intervals are sufficiently narrow to rule out an important magnitude of effect.
7.5 Interpret the magnitude of the observed effect.
7.6 Interpret the dose-response relationship in the observed results.
7.7 Interpret the potential effects of reporting and publication biases (e.g. unreported outcome data, unpublished studies etc.) on the observed results.
7.8 Interpret the external validity of the overall body of evidence. Any inferences or predictions about effects in target populations which are made based on effects observed in the populations in the included studies should accord with the considerations defined in the protocol about the relative relevance of populations (e.g. species, age, comorbidities etc.), exposures (e.g. timing, dose), and outcomes (direct or surrogate, acute or chronic model of disease, etc.), as appropriate. Deviations from these considerations must be explained and justified.
7.9 Include the “summary of findings” table.
7.10 Summarise the quality of the overall body of evidence into an appropriate overall statement of confidence in the results of the synthesis.
8: Drawing Conclusions
8.1 Draw out implications based only on findings from the synthesis of studies included in the review.
8.2 Describe implications for research based on Population-Exposure-Comparator-Outcome or other appropriate formula consistent with that specified in the research objective.
8.3 Avoid describing policy implications in terms of specific actions authors feel that decision-makers should take. If authors feel it is necessary to describe policy implications, articulate them in terms of hypothetical scenarios rather than making specific policy recommendations.

Table 1: The full list of COSTER recommendations for the planning and conduct of environmental health systematic reviews. The recommendations should be read alongside the explanatory notes in Table 2.

4. Discussion

4.1 How to use COSTER

4.1.1 Target audience of COSTER

COSTER is intended to be usable by any entity or practitioner responsible for or interested in conducting an EH SR project, and who needs a benchmark against which different possible approaches can be evaluated. Such entities include: independent scientists; journal editors receiving SR submissions; research teams wishing to conduct a SR; research commissioners seeking confidence that a contractor will conduct a successful SR project; quality assurance units in research-associated organisations seeking to implement consistent, good-quality SR practices; and regulatory authorities and scientific agencies seeking to demonstrate compliance with an agreed set of practices for conduct of research.

4.1.2 Managing the number of recommendations in COSTER

SRs are complex, multi-disciplinary projects that typically take 12-36 months to conduct (Borah et al. 2017; Haddaway and Westgate 2019). While 70 may seem like a large number of recommendations for a research team to follow, COSTER is comparable in size to IOM, which consists of 82 performance elements across 4 domains, and MECIR 1.07, which consists of 75 performance elements across 10 domains. COSTER is intended to be used in parallel to the development, conduct, and reporting of a systematic review in an iterative manner, which mirrors many of the considerations that should naturally arise for research teams undertaking each of these steps. Therefore, following COSTER's recommendations is unlikely to constitute an additional burden for a well-designed and well-conducted SR. In other scenarios, COSTER should help identify oversights and limitations in methods that might threaten the integrity of a SR project.

4.1.3 How should adherence with COSTER be described?

When research teams report the use of COSTER in planning and conducting a SR, they are encouraged to avoid broad summary statements such as "COSTER was followed" or "we adhered to the recommendations of COSTER". Although prevalent in the literature, such self-reported statements are usually only partly true and may therefore mislead the reader about the exact methods used (Page and Moher 2017). Instead, authors should report that COSTER was used to inform the planning and conduct of a SR, and transparently describe whether and how they were able to respond to each recommendation. The recommendations are numbered to facilitate this process. Where researchers elect to depart from COSTER, it is helpful if the reasons for doing so are explained.

4.2 Comparing COSTER to other SR standards and guidelines

COSTER is the first explicit effort by EH research practitioners and stakeholders to validate commonly-used biomedical SR standards for their particular cultural and research context. Table 2 highlights key explanatory points for COSTER according to themes that are either unique to the context of EH research, address aspects of systematic review conduct for which it has historically been difficult to achieve consensus on recommended practice, are potentially controversial given current SR practices in the field of EH, or provide a novel contribution to progressing SR practices in general. Where COSTER closely follows the conventions of IOM and MECIR, we refer the reader to Eden et al. (2011) and Higgins et al. (2019) for detailed explanation as to why the recommendations are considered good practice in SR.

Table 2: Explanation and elucidation of key recommendations of COSTER

Recommendations	1.1.1 through 1.5.4
Theme	Project planning
Contribution of COSTER	Emphasis on importance of standard practices in biomedical SRs for environmental health research
Explanation: COSTER recommends conducting EH SRs according to pre-published protocols. Following a pre-published protocol can reduce the risk that changes in methods mid-project will bias the results of a SR, by enabling comparison of the completed review with what was planned in the protocol (Centre for Reviews and Dissemination 2020). Protocol publication also provides an opportunity for external peer-review of proposed methods and subsequent early identification of errors which, if left unresolved, could undermine the validity of a resource-intensive project (Munafò et al. 2017). Although not yet common practice, some EH SRs are being conducted according to pre-published protocols – see e.g. Mandrioli et al. (2018), Matta et al. (2019), and Hansen et al. (2019). COSTER follows MECIR and IOM in providing comprehensive recommendations for the planning and protocol phase of a SR.	
Recommendations	1.1.3, 1.1.4
Theme	Disclosure and management of interests
Contribution of COSTER	Distinction between potential and apparent conflicts of interest relating to team selection in SRs
Explanation: COSTER recommends defining a conflict of interest (COI) as "a situation in which financial or other personal considerations would be considered by a reasonable person to have the potential to compromise or bias professional judgment and objectivity", and classifying COIs in two categories. These are: "apparent" conflicts of interest, defined as situations "in which a reasonable person would think that the professional's judgment is	

likely to be compromised"; and "potential" conflicts of interest, which are situations "that may develop into an apparent conflict of interest". This follows the Columbia University framework for "Responsible Conduct of Research" (Columbia University 2004).

The authors believe this approach offers a way to operationalise the description and handling of risks that COIs pose to the integrity of a SR project. Firstly, all interests are declared. Then, the classification of "potential" is applied to any interest for which the degree of conflict is unlikely to present a risk to the integrity of the project, while the classification of "apparent" is applied to any interest for which the degree of conflict may present excess risk to the integrity of the project. Persons with apparent conflicts of interests are excluded from involvement in decision-making processes.

COSTER allows for interests to be financial and non-financial. Similar to IOM, COSTER recognises that any potential COI can, in the right circumstances, become an apparent COI, and that all potential COIs should therefore be declared, evaluated and managed. COSTER distinguishes itself from the IOM approach to COIs by emphasising that individuals with apparent conflicts of interest need only be excluded from analysis and decision-making roles in the review process. This leaves open the possibility of their involvement in advisory capacity as individuals with specialist knowledge on which review teams can draw, while insulating the integrity of the review process from their apparent COIs by prohibiting their involvement in decision-making. This allows EH SRs to utilise the full range of expertise in a field in which many practitioners will likely have apparent COIs.

The authors emphasise that the intent of these recommendations is not to limit participation in EH SRs by excluding people with affiliation to broad sectors (e.g. academic grant holders, industry, or NGOs), but rather to make such associations transparent. In lieu of declaration of interest forms built specifically for environmental health research, SR authors could consider using forms such as those published by the International Committee of Medical Journal Editors (International Committee of Medical Journal Editors 2013).

Recommendations	1.2.2, 1.4.6, 7.8
Theme	Interpreting external validity of the evidence, and integrating multiple evidence streams
Contribution of COSTER	Adaptation of biomedical SR standards to specific context of EH research

Explanation: Operationalising the interpretation of indirect, non-human and *in vitro* evidence in the course of predicting health risks in target human populations is a fundamental challenge in adapting SR methods to environmental health. For healthcare interventions, IOM specifies the use of an "analytical framework which clearly lays out the chain of logic that links the health intervention to the outcomes of interest". COSTER applies

this concept in its recommendations for the assessment of the external validity of evidence, to account for the importance in EH research of consistent, unbiased interpretation of an evidence base which is often indirect.

EH researchers are increasingly interested in how the analysis of indirect mechanistic evidence can be organised via predictive biological networks (Villeneuve et al. 2014b, 2014a) or Key Characteristics frameworks (Smith et al. 2016; Arzuaga et al. 2019; Luderer et al. 2019) to help anticipate whether an environmental challenge will cause an adverse health outcome. In anticipation of the development of systematic approaches to developing and assessing the plausibility of such networks or framework analyses, in recommendation 1.2.2 COSTER asks that protocols include the basic elements of a theoretical framework for interpreting the external validity of included studies. The framework should describe why and to what extent the review team will consider different populations (e.g. species, developmental stage), exposures (e.g. timing, dose, similarity of substance / read-across) and outcomes (e.g. apical, intermediate) to be comparable to the target populations, exposures and outcomes of interest. Recommendation 7.8 asks that interpretation of the results of synthesis are made in accordance with this pre-specified framework.

Recommendations	1.2.3, 1.3.3, 1.3.4, 1.3.5, 1.3.9
Theme	Formulation of research objectives
Contribution of COSTER	Formal clarification of use of PECO-style statements in formulating SR objectives in EH research

Explanation: COSTER recommends formulating SR objectives in a structured format using context-appropriate elements of the PECOTS (Population-Exposure/Intervention-Comparator-Outcome-Target Condition-Study Design) mnemonic. SRs that investigate health effects of exposures and interventions (such as amelioration of the effects of exposures) are both expressly allowed for in COSTER.

COSTER also makes granular recommendations about the specific aspects of the PECOTS elements that should be considered in establishing the objectives of an EH SR. Because elements such as timing of exposure are a potentially critical issue in reliably identifying health risks of environmental exposures, COSTER recommends these be considered and defined as necessary. Specific guidance on good practice in the formulation of PECO statements by Morgan et al. (2018) has been published since the COSTER recommendations were finalised, to which prospective authors may wish to refer.

Recommendations	1.3.6 to1.3.9, 3.4
Theme	Including informally published or previously unpublished literature, regardless of usability in the planned analysis

Contribution of COSTER	Provides unambiguous rationale for exclusion of study reports due to insufficient information content

Explanation: COSTER recommends that grey literature (i.e. studies that have not been published in peer-reviewed journals) should be included in systematic reviews. This is because the relevance of evidence is determined by the SR objectives, not by the publication status of that evidence, the language the evidence is in, nor its compatibility with the analyses planned by the reviewers.

The inclusion of grey literature can act as a safeguard against the influence of publication bias; however, researchers should never assume that the grey literature which can be located is representative of the grey literature overall. The authors of COSTER also acknowledge that inclusion of grey literature can be daunting and for some SR authors may be controversial (Adams et al. 2016; Paez 2017). Therefore, COSTER provides an explicit rationale for where researchers can draw the line on including grey literature in a SR, as follows.

Firstly, in keeping with the SR principle of transparency, COSTER recommends that only publicly available information about a study be eligible for inclusion (recommendation 1.3.9). The authors note that a SR that brings into the public domain previously inaccessible information can be the mechanism by which such data becomes publicly accessible and therefore eligible for inclusion. This has happened with SRs from WHO (Descatha et al. 2018; Li et al. 2018) and Cochrane (Jefferson et al. 2014).

Secondly, COSTER recommends exclusion of studies for which there is insufficient information for risk of bias to be evaluated, to prevent the inclusion in a SR of evidence that is potentially misleading but cannot be identified as such by the reviewers (recommendation 1.3.6).

Thirdly, COSTER defines the included study itself, not documents describing the study, as the unit of evidence (recommendation 3.4). Therefore, COSTER recommends all publicly accessible study documents including conference abstracts etc. be gathered and assessed for information content as a whole, before a decision is made to exclude a study in accordance with recommendation 1.3.6. Researchers should take care not to double-count populations when combining multiple study reports, particularly when there is partial overlap between multiple documents.

Fourthly, COSTER recommends that documents should be included in a SR regardless of whether their data fit the analysis plan of the reviewers or they are in a language in which the reviewers are fluent. This is to ensure that study documents which may contain information of potential relevance to the SR's research objectives are not excluded from the data extraction step of the SR.

The authors are aware that many studies – especially epidemiological studies – cannot release detailed information on individual participants owing to privacy concerns and legal mandates. The intent of the grey literature recommendations in COSTER is not to exclude such studies, but rather to ensure that the use of study-specific findings within the larger analysis is supported by those aspects of the underlying data that are available for public scrutiny.

Recommendations	1.5.1, 1.5.2, 1.5.3
Theme	Protocol publication
Contribution of COSTER	Differentiates between protocol registration and publication as distinct steps of the methods development process

Explanation: Protocol registries such as PROSPERO (Centre for Reviews and Dissemination) and preprint repositories such as Zenodo (CERN) and the Open Science Framework (Center for Open Science 2020) allow authors to register their methods in advance of conducting a SR. However, there are no protocol registries that ensure authors have submitted sufficient information about methods that a reader can be confident a registered protocol is a complete plan for conducting a SR. Nor do such registries have capacity to peer-review protocols for soundness of the proposed methods. At most, they perform only basic quality control checks. This leads to a situation in which the value of self-registration for ensuring the comprehensiveness and validity of methods for a given protocol is unclear. Therefore, it is the view of the authors that self-registration of a protocol has value primarily as a record of intent to conduct a SR, rather than serving as a guarantee of comprehensive documentation of methods prior to conduct of a SR.

To address the limitations of protocol registration, COSTER recommends that authors of SRs take a two-step approach to protocol publication. As the first step, an outline of the proposed SR with the minimum necessary information to characterise objectives and approach should be posted on an appropriate public registry or functional equivalent thereof, over which the authors have no direct control (recommendation 1.5.1). This first draft is the permanent public record of intent to conduct a systematic review, functioning to communicate research aims and help other review teams avoid planning duplicate SRs. As the second step, this draft can then be developed in further detail as a full protocol submitted to external peer-review or other appropriate quality management process (recommendation 1.5.2), and then published either in a scientific journal or a preprint repository (recommendation 1.5.3). An example of journal publication of a protocol is provided by Mandrioli et al. (2018) and in a public repository by Martin et al. (2018). A general example of this kind of "two-stage" peer-review process, to which readers may wish to refer, is provided by the Registered Reports model of scientific publication (Chambers 2019).

Recommendations	1.4.3, 5
Theme	Internal validity assessment
Contribution of COSTER	Explicit specification of risk of bias methods for assessing internal validity of included studies

Explanation: To prevent systematic errors in included studies being transmitted through to the findings of a SR, COSTER recommends that each included study be assessed for internal validity, i.e. its potential to produce biased results. While anticipating direction and magnitude of bias is desirable in assessing the internal validity of included studies, this is often not possible or practical for SR projects; however, when feasible, evidence-based assessments of internal validity, which successfully quantify bias are consistent with COSTER.

COSTER makes no specific recommendations about which instruments should be used to assess risk of bias, leaving it to SR authors to determine which methods are best-suited to their research objectives. COSTER does, however, make a number of recommendations about the process of risk of bias assessment. This includes assessing risk of bias per outcome (recommendation 5.2) and making sure each judgement is transparent and grounded in the reviewed text (recommendation 5.6).

There is concern that risk of bias instruments may be misapplied in EH SRs, resulting in mischaracterisation of the validity of included studies (Farrah et al. 2019). The authors note that risk of bias assessment methods need to be sensitive to differences in study designs and employ suitable processes accordingly. The assessment process should balance being transparently conducted against a clear standard, whilst ensuring that potential limitations of a study are not mischaracterised by algorithmic comparison to inappropriately rigid validity criteria. Various systematic reviews and evaluations of risk of bias assessment tools are available (e.g. Wang et al. 2019; Krauth et al. 2013; Rooney et al. 2016) and a user of COSTER may wish to refer to such in deciding which tools to apply in a SR.

Recommendations	1.4.5, 7.1, 7.2, 7.4, 7.5, 7.6, 7.7, 7.8, 7.10
Theme	Assessment of confidence in the overall body of evidence
Contribution of COSTER	Emphasis on evaluation of quality of evidence against pre-specified criteria known to be of importance when assessing confidence in the results of a SR

Explanation: COSTER recommends that assessment of overall confidence in the evidence included in a SR cover seven characteristics: internal validity, consistency, precision, magnitude of effect, dose-response relationship, reporting and publication bias, and external validity. While these are the same broad characteristics as those utilised in the GRADE Framework (Guyatt et al. 2008; Guyatt et al. 2011), COSTER makes no specific recommendations about which tool should be used for assessing these characteristics nor

how they should be interpreted, except that the approach should be described in the SR protocol. Appropriate methods for assessing confidence in the results of an EH SR are a matter of ongoing discussion. The GRADE Framework is under active development for the environmental health context (Morgan et al. 2016; Morgan et al. 2019). A close interpretation of GRADE has been applied to environmental health questions in SRs by NTP OHAT (Rooney et al. 2014; National Toxicology Program 2016) and the Navigation Guide (Woodruff and Sutton 2014; Johnson et al. 2016). The US EPA IRIS Program, in a recent series of epidemiology and toxicology SRs (Radke et al. 2018; Radke et al. 2019) employed a certainty assessment framework that utilises similar concepts as to those recommended in COSTER. The authors believe that a systematic approach to assessing confidence in a body evidence is a fundamental part of the SR process because readers of a SR need a trustworthy analysis of the overall trustworthiness of the evidence. The authors also note that a high-quality SR of low-quality evidence is still a trustworthy SR – even if the SR process has shown that the reader cannot necessarily put much trust in the evidence itself.

Recommendations	8.3
Theme	Making policy recommendations
Contribution of COSTER	Emphasises that recommendations about interventions are often beyond the scope of a SR of health effects from environmental exposures

Explanation: The development of environmental health policy needs to account for a wide range of issues relating to evidence of health risks, due political process, and the values and preferences of stakeholders affected by the policy. In contrast, systematic reviews ask focused questions that typically respond to only one or two of the full set of issues of importance for policy development. This is especially true for SRs of health effects of environmental exposures: while they address potential causes of adverse health outcomes (i.e. are aetiological), they would not normally also investigate evidence for the effectiveness of interventions aimed at mitigating those adverse outcomes. While identifying threshold limits to inform policy decisions is often the core business of many EH SRs, COSTER adheres to the principle that the conclusions of a SR should not reach beyond the evidence that was included within it. COSTER therefore recommends authors resist answering questions about how best to mitigate the effects of an exposure or achieve a risk threshold unless the SR has been designed to systematically locate, appraise and synthesise the relevant evidence for providing such answers.

The authors note, however, that SRs characterising adverse outcomes from environmental exposures are often conducted to support policy decisions. COSTER therefore recommends that when authors present policy implications of their SR, they do so in the form of

hypothetical frameworks. This means authors should state that if certain described conditions obtain, then a given intervention may be effective for mitigating harm. Any assumptions the authors make about values, other evidence and potential consequences of a decision should be made explicit as part of that hypothetical framework.

294

295 ***4.3 Strengths and limitations of COSTER***

296 *4.3.1 The consensus process*

297 In developing COSTER, a deliberate attempt was made to emulate formal standardisation
298 processes such as those followed by the British Standards Institution. We made a particular
299 effort to involve a full complement of stakeholder groups in direct participation in the
300 consensus process. This was to ensure coverage of a wide range of potential opinions as to what
301 might constitute good practice in conduct of EH SRs, which then needed to converge over time
302 into a consensus view. We are not aware of other research standards that have sought to do this
303 to the same extent: the IOM standards represented the views of a committee of 16 medical
304 professionals supported by a team of researchers, while MECIR was developed by a dedicated
305 Cochrane committee and finalised in response to stakeholder comments.

306 In the end, we were able to achieve consensus of 20 workshop participants, plus TW. At least
307 one representative from each of the various stakeholder groups is represented in the
308 authorship, the results of which are a comprehensive set of 70 recommendations for good
309 practice in conduct of EH SRs. The recommendations cover complex issues including protocol
310 development, risk of bias, and certainty assessment, which are inconsistently implemented
311 across the EH SR literature.

312 In order to improve this consensus process , and to elevate COSTER from a set of expert
313 recommendations towards a more formal standard such as a Code of Practice (BS EN ISO
314 9001:2015; BS EN ISO 9000:2015), we suggest the following potential actions: securing greater
315 capacity to organise and participate in more face-to-face meetings; a longer process involving
316 more stakeholders to potentially allow for broader consensus on some of the more challenging
317 or controversial discussions, covering more elements of the SR process; and implementation of
318 more formal minute-taking and communication structures for making the consensus process
319 more auditable, improve transparency, and facilitate communication between participants in
320 the consensus process.

4.3.2 Author conflicts of interest

In order to secure cross-sector consensus, we invited participants because they had varied interests in developing a standard for conduct of EH SRs. We did not attempt to directly manage the interests of participants, as they were seen as desirable; instead, we sought balance across stakeholder groups and domains of expertise. We believe involvement of a broad cross-section of stakeholder groups strengthens COSTER's generalisability and broadens its acceptability, while reducing the risk that any individual interest group has had excess influence on the consensus outcome.

4.3.3 The process for developing seed recommendations for COSTER

Rather than conduct a SR of existing standards and guidance of potential relevance to seed the development of COSTER, we relied on participants' tacit knowledge of these in critiquing two established biomedical standards for SR practice. We secured participation of stakeholders with experience developing the following frameworks: the Navigation Guide (Woodruff and Sutton 2014), the National Toxicology Program Office of Health Assessment and Translation (Rooney et al. 2014); SYRINA (Vandenberg et al. 2016); the European Food Safety Authority (EFSA 2010); Cochrane's MECIR standards and the Cochrane Handbook (Higgins et al. 2011); GRADE (Morgan et al. 2016); the IARC Monographs Program (IARC 2015); and SYRCLE (Vries et al. 2015).

MECIR and IOM, as seeds for COSTER, were selected as authoritative standards likely to be comprehensive and not misleading in either what they include or omit. These two existing standards provided 80 seed criteria (see Supplemental Materials 03). While a SR of existing standards and guidelines could have extended this list, we believe it would have been a considerable task to undertake without obvious proportional benefit to a project which sought to define an initial expert consensus on basic recommended practices in EH SR. This is an element of the COSTER development methodology which could certainly be improved in future; a detailed discussion of this follows in Section 4.4 below.

4.3.4 Potential for misuse of COSTER

The value of all SRs is diminished by misuse of the term "systematic" and the publication of poor-quality SR manuscripts. COSTER seeks to avert this situation by giving authors, reviewers, editors and other stakeholders clear, comprehensive recommendations on the fundamental practices of SR. At the very least, by providing an unambiguous set of recommendations against which the conduct of a putative SR can be compared, the authors hope that it will be easier for the user to identify when phrases such as "adheres with the recommendations of COSTER" and "employed systematic review methods" are being misused.

In general, the authors recommend that readers be cautious in making any assumptions about the quality of a SR which uses or claims to have complied with COSTER. While COSTER is intended to help authors make good decisions about their EH SR methods, as a written document it has little power on its own to ensure they have been successful in making them. As is the case for any standard or set of recommendations, claims of following COSTER are open to potential abuse, either deliberate or inadvertent, as a mechanism for artificially elevating a reader's perception of the quality of a piece of research. A SR should therefore always be appraised using a valid, contextually appropriate tool before coming to any judgments about its quality.

4.4 Future development of COSTER

The recommendations of COSTER are intended as a first step in a broader research and consensus-building process, which it is hoped will eventually yield a robust, international standard for conduct of systematic reviews in environmental health research. Formal standards are typically based on both expectation and empirical evidence that the practices described in the standard contribute to a product or process being fit for purpose, combined with broad acceptance of the practices among the community that is expected to adopt the standard (British Standards Institution 2016b). Since SR methods are still relatively new in environmental health research, it follows that while the consensus view of small groups of experienced practitioners as to what they consider good practice can be secured, this view is unlikely to be universally shared and strong evidence for what is effective practice is not necessarily available. This is particularly true for areas in which SR methods are not readily portable from social science and medical contexts to environmental health, or where environmental health researchers face challenges not encountered in other fields. Broad consensus is also a challenging goal when only a small, albeit growing, part of the community is employing SR methods in conducting reviews of evidence, and practices across those SRs are inconsistent. While COSTER represents the consensus view of the authors, other expert groups may disagree with some of the recommendations of COSTER. Such disagreement is healthy: by making explicit a set of key recommended practices for SR, COSTER serves as a focal point for discussion and advancing consensus across groups.

As community experience in conducting EH SRs develops over the next period, the authors suggest that future development of COSTER adopt the framework for development of reporting guidelines for health research presented in Moher et al. (2014). This framework emphasises four steps:

1. a systematic review of existing standards and guidelines;

2. a systematic review of the prevalence of current research practices;

3. the critical appraisal of existing guidelines and current research practices for completeness, face validity, and construct validity;

4. a process to determine community consensus on best practices and the criteria for a guideline.

Steps 1 and 2 would result in a larger seed-set of potential recommendations than was provided by selecting the MECIR and IOM standards as the basis for the current consensus. However, such a SR could be a significant undertaking, as it requires a decision as to what is relevant (e.g. should nutrition and public health standards be included?) and potentially interpreting the implied standards in several large handbooks, a large number of reporting standards and guidelines, and potentially even individual SR study reports as well. This is a major challenge for qualitative analysis and requires appropriate resources.

Steps 3 and 4, as a broad discussion and consensus process, would provide a community view of where current practices fall short of expectation or need, or where specific processes might exceed what the community views as strictly necessary for conduct of a robust EH SR. For future versions of COSTER, it is important that the consensus process be extended beyond the 21 people it was possible to involve here. Care will need to be taken to maintain stakeholder balance as numbers of participants are increased.

The authors recommend COSTER be re-assessed according to the above methodology, with a view to an updated set of recommendations being published around 2025. Some examples of recent methodological innovations in EH SR which should be considered for inclusion in future versions of COSTER include:

- more detailed recommendations for handling of specific types of evidence, including mechanistic and *in vitro* study designs, observational studies and controlled trials in humans;

- the handling of evidence of the efficacy of EH interventions, e.g. health benefits from introducing low-smoke cookstoves (e.g. Quansah et al. 2017);

- more advanced evidence integration techniques such as triangulation (e.g. Lawlor et al. 2016) and meta-regression (e.g. Phung et al. 2017);

- the prespecification of exposure assessment criteria in risk of bias assessment, where COSTER currently only explicitly mentions confounders;

- more detailed recommendations for appraising the external validity of included studies.

5. Conclusion

COSTER presents the recommendations of a diverse group of expert practitioners, reflecting their consensus view on good practice in the planning and conduct of environmental health systematic reviews. COSTER is intended as the first step in a broader consensus-building process which should lead to the eventual development of robust standards for conduct of EH SRs, while in the near-term providing recommendations on good practice as guidance for EH SR stakeholders.

Acknowledgements

We would like to thank Kate Jones and the Royal Society of Chemistry for hosting the workshop, and Lancaster University Faculty of Science and Technology and Lancaster Environment Centre for providing funding to run the workshop. Funding was also provided by the UK's Economic & Social Research Council (ESRC) "Radical Futures" programme and the Engineering & Physical Science Research Council (EPSRC) "Impact Acceleration Award" EP/K50421X/1 for developing systematic review methodology for environmental health.

The authors declare no competing financial interests. Further details on potential COIs have been provided in the DOI forms (see supplemental materials). The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of their respective employers or organisations. Previous submitted versions of the manuscript have been archived at https://doi.org/10.5281/zenodo.3560978.

We would also like to thank the following for their contribution to the workshop discussions: Sarah Bull (Royal Society of Chemistry); Richard Brown (World Health Organization); Kurt Straif (ret.) and Kathryn Guyton (International Agency for Research on Cancer); Julian Higgins (University of Bristol); Toby Lasserson (Cochrane Editorial Unit); Jennifer McPartland (Environmental Defense Fund); Sharon Munn (EU Joint Research Centre); Angelika Tritscher (World Health Organization); Christopher Weiss (US National Institute of Environmental Health Sciences). TW did not participate in the workshop but contributed to the consensus development calls and the manuscript.

1 Publication bibliography

Adams, Jean; Hillier-Brown, Frances C.; Moore, Helen J.; Lake, Amelia A.; Araujo-Soares, Vera; White, Martin; Summerbell, Carolyn (2016): Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies. In *Systematic reviews* 5 (1), p. 164. DOI: 10.1186/s13643-016-0337-y.

Arzuaga, Xabier; Smith, Martyn T.; Gibbons, Catherine F.; Skakkebæk, Niels E.; Yost, Erin E.; Beverly, Brandiese E. J. et al. (2019): Proposed Key Characteristics of Male Reproductive Toxicants as an Approach for Organizing and Evaluating Mechanistic Evidence in Human Health Hazard Assessments. In *Environmental health perspectives* 127 (6), p. 65001. DOI: 10.1289/EHP5045.

Borah, Rohit; Brown, Andrew W.; Capers, Patrice L.; Kaiser, Kathryn A. (2017): Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. In *BMJ open* 7 (2), e012545. DOI: 10.1136/bmjopen-2016-012545.

Campbell Collaboration (2014): Methodological Expectations of Campbell Collaboration Intervention Reviews (MEC2IR), Conduct Standards, version 1.0. Available online at http://www.campbellcollaboration.org/news_/Campbell_adopts_MEC2IR_guidelines.php, checked on 3/8/2017.

Center for Open Science (2020): Open Science Framework. Available online at https://osf.io/, checked on 5/24/2020.

Centre for Reviews and Dissemination: PROSPERO: International prospective register of systematic reviews. University of York. Available online at https://www.crd.york.ac.uk/prospero/, checked on 7/21/2019.

Centre for Reviews and Dissemination (2020): About PROSPERO. University of York. Available online at https://www.crd.york.ac.uk/prospero/#aboutpage, updated on 2020, checked on 5/24/2020.

CERN: Zenodo. European Organization for Nuclear Research. Available online at https://zenodo.org/, checked on 7/21/2019.

Chambers, Chris (2019): What's next for Registered Reports? In *Nature* 573 (7773), pp. 187–189. DOI: 10.1038/d41586-019-02674-6.

Chandler, Jackie; Churchill, Rachel; Higgins, Julian; Lasserson, Toby; Tovey, David (2013): Methodological standards for the conduct of new Cochrane Intervention Reviews. Cochrane Editorial Unit.

Columbia University (2004): Responsible Conduct of Research: Conflicts of Interest. Available online at http://ccnmtl.columbia.edu/projects/rcr/rcr_conflicts/foundation/index.html, checked on 1/3/2018.

Descatha, Alexis; Sembajwe, Grace; Baer, Michael; Boccuni, Fabio; Di Tecco, Cristina; Duret, Clément et al. (2018): WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on stroke. In *Environment international* 119, pp. 366–378. DOI: 10.1016/j.envint.2018.06.016.

Eden, Jill; Levit, Laura A.; Berg, Alfred O.; Morton, Sally C. (2011): Finding what works in health care. Standards for systematic reviews. Washington, D.C.: National Academies Press.

EFSA (2010): Application of systematic review methodology to food and feed safety assessments to support decision making. In *EFSA Journal* 8 (6), p. 1637. DOI: 10.2903/j.efsa.2010.1637.

EFSA (2015): Principles and process for dealing with data and evidence in scientific assessments. In *EFSA Journal* 13 (6), p. 4121. DOI: 10.2903/j.efsa.2015.4121.

EPA (2018): Application of Systematic Review in TSCA Risk Evaluations. US EPA Office of Chemical Safety and Pollution Prevention (EPA Document # 740-P1-8001). Available online at https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/application-systematic-review-tsca-risk-evaluations, checked on 5/8/2019.

Farrah, Kelly; Young, Kelsey; Tunis, Matthew C.; Zhao, Linlu (2019): Risk of bias tools in systematic reviews of health interventions: an analysis of PROSPERO-registered protocols. In *Systematic reviews* 8 (1), p. 280. DOI: 10.1186/s13643-019-1172-8.

Guyatt, Gordon H.; Oxman, Andrew D.; Schünemann, Holger J.; Tugwell, Peter; Knottnerus, Andre (2011): GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. In *Journal of clinical epidemiology* 64 (4), pp. 380–382. DOI: 10.1016/j.jclinepi.2010.09.011.

Guyatt, Gordon H.; Oxman, Andrew D.; Vist, Gunn E.; Kunz, Regina; Falck-Ytter, Yngve; Alonso-Coello, Pablo; Schünemann, Holger J. (2008): GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. In *BMJ (Clinical research ed.)* 336 (7650), pp. 924–926. DOI: 10.1136/bmj.39489.470347.AD.

Haddaway, Neal R.; Westgate, Martin J. (2019): Predicting the time needed for environmental systematic reviews and systematic maps. In *Conservation biology : the journal of the Society for Conservation Biology* 33 (2), pp. 434–443. DOI: 10.1111/cobi.13231.

Hansen, Martin Rune Hassan; Jørs, Erik; Sandbæk, Annelli; Kolstad, Henrik Albert; Schullehner, Jörg; Schlünssen, Vivi (2019): Exposure to neuroactive non-organochlorine insecticides, and diabetes mellitus and related metabolic disturbances: Protocol for a systematic review and meta-analysis. In *Environment international* 127, pp. 664–670. DOI: 10.1016/j.envint.2019.02.074.

Higgins, J.P.T.; Thomas, J.; Chandler, J.; Cumpston, M.; Li, T.; Page, M. J.; Welch, V. A. (Eds.) (2019): Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane. Available online at www.training.cochrane.org/handbook.

Higgins, Julian P. T.; Altman, Douglas G.; Gøtzsche, Peter C.; Jüni, Peter; Moher, David; Oxman, Andrew D. et al. (2011): The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. In *BMJ (Clinical research ed.)* 343, d5928.

IARC (2015): IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Preamble. World Health Organization International Agency for Research on Cancer. Lyon, France.

IARC (2019a): IARC Monographs on the Identification of Carcinogenic Hazards to Humans: Preamble. IARC. Lyon, France.

IARC (2019b): Instructions for Authors for the Preparation of Drafts for IARC Monographs, updated on 6/17/2019.

International Committee of Medical Journal Editors (2013): ICMJE Form for Disclosure of Potential Conflicts of Interest. Available online at http://www.icmje.org/conflicts-of-interest/, checked on 1/3/2018.

ISO/IEC (2004): ISO/IEC 2:2004 Standardization and related activities -- General vocabulary. ISO/IEC. Switzerland. Available online at https://www.iso.org/standard/39976.html, checked on 7/20/2017.

Jefferson, Tom; Jones, Mark A.; Doshi, Peter; Del Mar, Chris B.; Hama, Rokuro; Thompson, Matthew J. et al. (2014): Neuraminidase inhibitors for preventing and treating influenza in healthy adults and children. In *The Cochrane database of systematic reviews* (4), CD008965. DOI: 10.1002/14651858.CD008965.pub4.

Johnson, Paula I.; Koustas, Erica; Vesterinen, Hanna M.; Sutton, Patrice; Atchley, Dylan S.; Kim, Allegra N. et al. (2016): Application of the Navigation Guide systematic review methodology to the evidence for developmental and reproductive toxicity of triclosan. In *Environment international* 92-93, pp. 716–728. DOI: 10.1016/j.envint.2016.03.009.

Krauth, David; Woodruff, Tracey J.; Bero, Lisa (2013): Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. In *Environmental health perspectives* 121 (9), pp. 985–992. DOI: 10.1289/ehp.1206389.

Lawlor, Debbie A.; Tilling, Kate; Davey Smith, George (2016): Triangulation in aetiological epidemiology. In *International journal of epidemiology* 45 (6), pp. 1866–1886. DOI: 10.1093/ije/dyw314.

Li, Jian; Brisson, Chantal; Clays, Els; Ferrario, Marco M.; Ivanov, Ivan D.; Landsbergis, Paul et al. (2018): WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on ischaemic heart disease. In *Environment international* 119, pp. 558–569. DOI: 10.1016/j.envint.2018.06.022.

Luderer, Ulrike; Eskenazi, Brenda; Hauser, Russ; Korach, Kenneth S.; McHale, Cliona M.; Moran, Francisco et al. (2019): Proposed Key Characteristics of Female Reproductive Toxicants as an Approach for Organizing and Evaluating Mechanistic Data in Hazard Assessment. In *Environmental health perspectives* 127 (7), p. 75001. DOI: 10.1289/EHP4971.

Mandrioli, Daniele; Schlünssen, Vivi; Ádám, Balázs; Cohen, Robert A.; Colosio, Claudio; Chen, Weihong et al. (2018): WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of occupational exposure to dusts and/or fibres and of the effect of occupational exposure to dusts and/or fibres on pneumoconiosis. In *Environment international* 119, pp. 174–185. DOI: 10.1016/j.envint.2018.06.005.

Martin, O. V.; Bopp, S.; Ermler, S.; Kienzler, A.; McPhie, J.; Paini, A. et al. (2018): Protocol For A Systematic Review Of Ten Years Of Research On Interactions In Chemical Mixtures Of Environmental Pollutants.

Matta, Komodo; Ploteau, Stéphane; Coumoul, Xavier; Koual, Meriem; Le Bizec, Bruno; Antignac, Jean-Philippe; Cano-Sancho, German (2019): Associations between exposure to organochlorine chemicals and endometriosis in experimental studies: A systematic review protocol. In *Environment international* 124, pp. 400–407. DOI: 10.1016/j.envint.2018.12.063.

Moher, David; Altman, Douglas G.; Schulz, Kenneth F.; Simera, Iveta; Wager, Elizabeth (Eds.) (2014): Guidelines for Reporting Health Research: A User's Manual. Oxford, UK: John Wiley & Sons, Ltd.

Morgan, Rebecca L.; Beverly, Brandy; Ghersi, Davina; Schünemann, Holger J.; Rooney, Andrew A.; Whaley, Paul et al. (2019): GRADE guidelines for environmental and occupational health: A

new series of articles in Environment International. In *Environment international* 128, pp. 11–12. DOI: 10.1016/j.envint.2019.04.016.

Morgan, Rebecca L.; Thayer, Kristina A.; Bero, Lisa; Bruce, Nigel; Falck-Ytter, Yngve; Ghersi, Davina et al. (2016): GRADE: Assessing the quality of evidence in environmental and occupational health. In *Environment international* 92-93, pp. 611–616. DOI: 10.1016/j.envint.2016.01.004.

Morgan, Rebecca L.; Whaley, Paul; Thayer, Kristina A.; Schünemann, Holger J. (2018): Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. In *Environment international* 121 (Pt 1), pp. 1027–1031. DOI: 10.1016/j.envint.2018.07.015.

Munafò, Marcus R.; Nosek, Brian A.; Bishop, Dorothy V. M.; Button, Katherine S.; Chambers, Christopher D.; Du Percie Sert, Nathalie et al. (2017): A manifesto for reproducible science. In *Nat Hum Behav* 1 (1), e124. DOI: 10.1038/s41562-016-0021.

NAS (2014): Review of EPA's Integrated Risk Information System (IRIS) Process. Washington (DC).

NAS (2017): Application of Systematic Review Methods in an Overall Strategy for Evaluating Low-Dose Toxicity from Endocrine Active Chemicals. Washington (DC).

National Toxicology Program (2016): Monograph on Immunotoxicity Associated with Exposure to Perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS). National Toxicology Program. Research Triangle Park, NC. Available online at https://ntp.niehs.nih.gov/ntp/ohat/pfoa_pfos/pfoa_pfosmonograph_508.pdf.

NTP OHAT (2019): Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. US National Toxicology Program Office of Health Assessment and Translation. Available online at https://ntp.niehs.nih.gov/pubhealth/hat/review/index-2.html, checked on 5/8/2019.

Paez, Arsenio (2017): Grey literature: An important resource in systematic reviews. In *Journal of evidence-based medicine*. DOI: 10.1111/jebm.12265.

Page, Matthew J.; Moher, David (2017): Evaluations of the uptake and impact of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement and extensions. A scoping review. In *Systematic reviews* 6 (1), p. 263. DOI: 10.1186/s13643-017-0663-8.

Phung, Dung; Des Connell; Rutherford, Shannon; Chu, Cordia (2017): Cardiovascular risk from water arsenic exposure in Vietnam: Application of systematic review and meta-regression analysis in chemical health risk assessment. In *Chemosphere* 177, pp. 167–175. DOI: 10.1016/j.chemosphere.2017.03.012.

BS EN ISO 9000:2015, September 2015: Quality management systems - Fundamentals and vocabulary.

BS EN ISO 9001:2015, September 2015: Quality management systems - Requirements.

Quansah, Reginald; Semple, Sean; Ochieng, Caroline A.; Juvekar, Sanjar; Armah, Frederick Ato; Luginaah, Isaac; Emina, Jacques (2017): Effectiveness of interventions to reduce household air pollution and/or improve health in homes using solid fuel in low-and-middle income countries: A systematic review and meta-analysis. In *Environment international* 103, pp. 73–90. DOI: 10.1016/j.envint.2017.03.010.

Radke, Elizabeth G.; Braun, Joseph M.; Meeker, John D.; Cooper, Glinda S. (2018): Phthalate exposure and male reproductive outcomes: A systematic review of the human epidemiological evidence. In *Environment international* 121 (Pt 1), pp. 764–793. DOI: 10.1016/j.envint.2018.07.029.

Radke, Elizabeth G.; Galizia, Audrey; Thayer, Kristina A.; Cooper, Glinda S. (2019): Phthalate exposure and metabolic effects: a systematic review of the human epidemiological evidence. In *Environment international* 132, p. 104768. DOI: 10.1016/j.envint.2019.04.040.

Rooney, Andrew A.; Boyles, Abee L.; Wolfe, Mary S.; Bucher, John R.; Thayer, Kristina A. (2014): Systematic review and evidence integration for literature-based environmental health science assessments. In *Environmental health perspectives* 122 (7), pp. 711–718. DOI: 10.1289/ehp.1307972.

Rooney, Andrew A.; Cooper, Glinda S.; Jahnke, Gloria D.; Lam, Juleen; Morgan, Rebecca L.; Boyles, Abee L. et al. (2016): How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. In *Environment international* 92-93, pp. 617–629. DOI: 10.1016/j.envint.2016.01.005.

Schaefer, Heather R.; Myers, Jessica L. (2017): Guidelines for performing systematic reviews in the development of toxicity factors. In *Regulatory toxicology and pharmacology : RTP* 91, pp. 124–141. DOI: 10.1016/j.yrtph.2017.10.008.

Singla, Veena I.; Sutton, Patrice M.; Woodruff, Tracey J. (2019): The Environmental Protection Agency Toxic Substances Control Act Systematic Review Method May Curtail Science Used to

Inform Policies, With Profound Implications for Public Health. In *American journal of public health* 109 (7), pp. 982–984. DOI: 10.2105/AJPH.2019.305068.

Smith, Martyn T.; Guyton, Kathryn Z.; Gibbons, Catherine F.; Fritz, Jason M.; Portier, Christopher J.; Rusyn, Ivan et al. (2016): Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. In *Environmental health perspectives* 124 (6), pp. 713–721. DOI: 10.1289/ehp.1509912.

Stephens, Martin L.; Betts, Kellyn; Beck, Nancy B.; Cogliano, Vincent; Dickersin, Kay; Fitzpatrick, Suzanne et al. (2016): The Emergence of Systematic Review in Toxicology. In *Toxicological Sciences* 152 (1), pp. 10–16. DOI: 10.1093/toxsci/kfw059.

Vandenberg, Laura N.; Ågerstrand, Marlene; Beronius, Anna; Beausoleil, Claire; Bergman, Åke; Bero, Lisa A. et al. (2016): A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. In *Environmental health : a global access science source* 15 (1), p. 74. DOI: 10.1186/s12940-016-0156-6.

Villeneuve, Daniel L.; Crump, Doug; Garcia-Reyero, Natàlia; Hecker, Markus; Hutchinson, Thomas H.; LaLone, Carlie A. et al. (2014a): Adverse outcome pathway (AOP) development I: strategies and principles. In *Toxicological sciences : an official journal of the Society of Toxicology* 142 (2), pp. 312–320. DOI: 10.1093/toxsci/kfu199.

Villeneuve, Daniel L.; Crump, Doug; Garcia-Reyero, Natàlia; Hecker, Markus; Hutchinson, Thomas H.; LaLone, Carlie A. et al. (2014b): Adverse outcome pathway development II: best practices. In *Toxicological sciences : an official journal of the Society of Toxicology* 142 (2), pp. 321–330. DOI: 10.1093/toxsci/kfu200.

Vries, R. B. M. de; Hooijmans, Carlijn R.; Langendam, Miranda W.; van Luijk, Judith; Leenaars, Marlies; Ritskes-Hoitinga, Merel; Wever, Kimberley E. (2015): A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies. In *Evidence-based Preclinical Medicine* 2 (1), e00007. DOI: 10.1002/ebm2.7.

Wang, Zhicheng; Taylor, Kyla; Allman-Farinelli, Margaret; Armstrong, Bruce; Askie, Lisa; Ghersi, Davina et al. (2019): A systematic review: Tools for assessing methodological quality of human observational studies.

Whaley, Paul; Aiassa, Elisa; Beausoleil, Claire; Beronius, Anna; Bilotta, Gary; Boobis, Alan et al. (2019): A code of practice for the conduct of systematic reviews in toxicology and environmental health research (COSTER).

669 Whaley, Paul; Halsall, Crispin; Agerstrand, Marlene; Aiassa, Elisa; Benford, Diane; Bilotta, Gary et
670 al. (2016): Implementing systematic review techniques in chemical risk assessment: Challenges,
671 opportunities and recommendations. In *Environment international* 92-93, pp. 556–564. DOI:
672 10.1016/j.envint.2015.11.002.

673 Woodruff, Tracey J.; Sutton, Patrice (2014): The Navigation Guide systematic review
674 methodology: a rigorous and transparent method for translating environmental health science
675 into better health outcomes. In *Environmental health perspectives* 122 (10), pp. 1007–1014. DOI:
676 10.1289/ehp.1307175.

677