



Applying Systematic Review Methods in Chemical Risk Assessment

by

Paul Alexander Whaley (MLitt)

*This dissertation is submitted for the
degree of Doctor of Philosophy*

-

January 2021

-

Lancaster Environment Centre

*To everyone who encouraged me.
Especially Miriam, who also had to live
with the consequences.*

*"Science is supposed to be cumulative, but scientists
only rarely cumulate evidence scientifically"*

Chalmers, Hedges & Cooper

Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated.

Abstract

Context

Chemical risk assessment has traditionally been dependent on “narrative” approaches for synthesising evidence about potential health harms from exposure to chemical substances. However, narrative reviews are recognised as being vulnerable to a range of methodological shortcomings which introduce bias and inconsistency into the summarisation of scientific evidence. This is likely to be a contributing factor in a number of controversies about the safety of chemical substances. The potential value of systematic review methods for improving the transparency and validity of chemical risk assessments was arguably first articulated in the mid-2000s. By 2015, the first major frameworks for conducting systematic reviews of environmental health evidence had been published. What was not well understood at the time was how systematic review, as a technically exacting methodology originally developed for evaluating the effectiveness of interventions in healthcare, might be adapted to the specific workflows and evidence streams of chemical risk assessment.

Objectives

The aim of this Thesis is to investigate how systematic review methods can be applied to the conduct of chemical risk assessment. This overall aim is broken down into four specific objectives: to identify practical challenges and knowledge gaps which impede the implementation of systematic review methods in chemical risk assessment; to define a consensus view on key recommended practices for the planning and conduct of systematic reviews in the environmental health sciences; to examine how “biological plausibility” as a concept fundamental to risk assessment is accommodated in systematic review methodologies; and to describe the role of ontologies in making evidence accessible for use in systematic chemical assessments.

Discussion

The use of systematic review methods should improve the validity, utility and transparency of chemical risk assessments. However, the successful implementation of systematic review methods hinges on addressing a number of challenges, including the

development of guidance for their conduct in environmental health contexts, and the technical development of methods where systematic review approaches need to be adapted to the specific requirements of chemical risk assessment.

In terms of developing guidance, a detailed set of recommendations for the conduct of systematic reviews in environmental health and toxicological research was developed. These “COSTER” recommendations identify 70 practices across eight performance domains that will help ensure consistent and high standards for the growing number systematic reviews on environmental health topics.

In terms of technical development of methods, “biological plausibility” is a concept used by risk assessors to describe the extent to which an experimental surrogate or knowledge of relevant biological mechanisms are informative of a systematic review conclusion. Through examination of 12 case examples it is concluded that “biological plausibility” is in fact already accommodated in the systematic review process under the assessment of the indirectness or external validity of evidence; however, the considerations which risk assessors take into account when assessing biological plausibility should be absorbed into the assessment of external validity of studies.

Finally, examination of the concept of biological plausibility demonstrates the extreme heterogeneity and volume of data which has to be accommodated in chemical risk assessments. The role of ontologies in Knowledge Organisation Systems is examined as a key enabler of scaling up of systematic review methods to handling the volume of evidence which needs to be analysed if tens of thousands of chemicals, covering potentially millions of studies, are to be reviewed systematically.

Acknowledgements

Chapter 1. Funding for the workshop was provided through the Economic & Social Science Research Council grant “Radical Futures in Social Sciences” (Lancaster University) and Lancaster Environment Centre. CH, PW, AR are grateful to Lancaster University's Faculty of Science & Technology “Distinguished Visitors” funding programme. The Royal Society of Chemistry is acknowledged for generously providing a meeting room, refreshments and facilitating the workshop proceedings. The PhD studentship of PW is partly funded through Lancaster Environment Centre. The contribution of non-author workshop participants to the development of the manuscript is also greatly appreciated.

Chapter 2. I 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.

Chapter 3. I would like to thank the GRADE Environmental Health Project Group and GRADE Working Group for their contributions to this manuscript, and the Evidence-based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health for providing funding for covering the time of PW, KT and SH in working on this manuscript. The authors would also thank the European Food Safety Authority and EBTC for organising the Scientific Colloquium, and the participants who contributed to discussions therein, which gave genesis to the concept of this manuscript

Chapter 4. I would like to thank George Woodall, Shannon Bell, Janice Lee, and Kris Thayer for their technical review, and Kristan Markey for conceptual and intellectual knowledge contributions. Funding for this study came from the U.S. Environmental Protection Agency Office of Research and Development. The work described in this article has been reviewed by the Center for Environmental and Public Health Assessment of U.S. Environmental Protection Agency and approved for publication. The views expressed in this paper are those of the authors and do not necessarily reflect

the views or policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Thesis Template. I would also thank Kayla Friedman and Malcolm Morgan of the University of Cambridge University, UK, and Charles Weir of Lancaster University, UK, for producing the Microsoft Word thesis template used to produce this document.

Overall. I would especially like to thank Crispin Halsall and Ruth Alcock for giving me the opportunity to do this PhD and their support throughout. It has opened up a world of possibilities with which I could not otherwise imagine being presented.

Contents

| | |
|---|------------|
| INTRODUCTION..... | 1 |
| Background | 1 |
| <i>Assessing risks to health posed by exposure to chemical substances.....</i> | <i>1</i> |
| <i>Systematic review as a potential solution to inconsistency in risk assessment.....</i> | <i>3</i> |
| Objectives and structure of this Thesis | 5 |
| <i>Chapter 1. Challenges and opportunities</i> | <i>6</i> |
| <i>Chapter 2. Recommended practices</i> | <i>7</i> |
| <i>Chapter 3. Biological plausibility.....</i> | <i>8</i> |
| <i>Chapter 4. Ontologies.....</i> | <i>9</i> |
| References | 10 |
| | |
| CHAPTER 1. CHALLENGES AND OPPORTUNITIES | 14 |
| | |
| CHAPTER 2. RECOMMENDED PRACTICES..... | 24 |
| | |
| CHAPTER 3. BIOLOGICAL PLAUSIBILITY | 38 |
| | |
| CHAPTER 4. ONTOLOGIES..... | 61 |
| | |
| CHAPTER 5. CONCLUSIONS AND FUTURE WORK | 88 |
| Conclusions..... | 88 |
| <i>Improving the quality of systematic reviews.....</i> | <i>88</i> |
| <i>New evidence synthesis methods for chemical risk assessment.....</i> | <i>90</i> |
| <i>The need to automate evidence synthesis.....</i> | <i>91</i> |
| <i>A radically different future.....</i> | <i>92</i> |
| Future Work: “Research Without Reading” | 93 |
| <i>Standards for complete, accurate and machine-readable research</i> | <i>93</i> |
| <i>The database technology for Knowledge Organisation Systems.....</i> | <i>94</i> |
| <i>Machine-compatible evidence analysis tools.....</i> | <i>95</i> |
| References | 97 |
| | |
| CONSOLIDATED BIBLIOGRAPHY..... | 99 |
| | |
| APPENDICES | 117 |

List of Tables

Introduction

No tables

Chapter 1. Challenges and Opportunities

No tables

Chapter 2. Recommended Practices

Table 1. *The full list of COSTER recommendations for the planning and conduct of environmental health systematic reviews*..... **p.29**

Table 2. *Explanation and elucidation of the key recommendations of COSTER* **p.32**

Chapter 3. Biological Plausibility

Table 1. *Examples of definitions of “biological plausibility”* **p.42**

Table 2. *Summary of the examples in Chapter 3 which indicate how discussion of biological plausibility maps onto the concepts of systematic review* **p.51**

Table 3. *Potential influencing factors in judging biological plausibility or external validity of study surrogates, as suggested by the examples in Chapter 3* **p.54**

Chapter 4. Ontologies

Table 1. *Demonstration of how variation in language used by study authors in title, abstract, and author keywords fields affects search results in PubMed. Database syntax is used to ensure the phrase entered is the exact one being searched for. Date of searches: 15 July 2020*..... **p. 69**

Chapter 5. Conclusions and Future Work

No tables

List of Figures

Introduction

Figure 1. *The components of a human-health risk assessment. Archetypal questions asked at each stage of the risk assessment and risk management process are included. The components are typical of wider environmental risk assessments. Adapted from World Health Organisation Chemical Risk Assessment Network (in prep)..... p.2*

Chapter 1. Challenges and Opportunities

Figure 1. *An overview to the chemical risk assessment (CRA) process, whereby risk is a function of hazard and exposure. While SR methods could in principle be applied to all steps of the CRA process, it is the view of the workshop participants that up to this point in time most attention has been focused on the hazard identification and hazard characterisation steps. There are issues around conducting a systematic review for exposure assessment which were not discussed at the workshop, such as the requirement for a very different tool for assessing risk of bias in exposure studies which may necessitate specialised knowledge of analytical/environmental chemistry..... p.17*

Box 1. *Examples of conflicting opinions from scientists and government agencies about the risks to health posed by bisphenol-A at current exposure levels p.18*

Box 2. *The use of PECO statements in the SR process p.20*

Box 3. *The potential utility of SR methods in application to REACH registrations p.21*

Chapter 2. Recommended Practices

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 p.26*

Figure 2. *Conceptual structure of COSTER with objectives for each stage of the systematic review process* **p.28**

Chapter 3. Biological Plausibility

Figure 1. *The upgrade and downgrade domains in GRADE and how they are used to determine the overall certainty in evidence for a systematic review. Adapted from (Morgan et al., 2016)* **p.43**

Figure 2. *Schematic representation of when studies of surrogates might be included in a systematic review* **p.45**

Figure 3. *The relationship between the generalisability and mechanistic aspects of biological plausibility* **p.52**

Figure 4. *How biological plausibility maps onto the processes of systematic review via the shared concept of external validity. While questions about biological mechanisms (e.g. how an exposure causes an outcome) are independent of a given systematic review, answers to those questions can be highly informative in judging the external validity of evidence*..... **p.53**

Figure 5: *Illustrations of the potential influencing factors in judging biological plausibility or external validity of study surrogates, as suggested by the examples in Chapter 3* **p.54**

Chapter 4. Ontologies

Figure 1. *The relationship between the processes involved in systematically mapping and systematically reviewing evidence. The elements which we discuss as the “information retrieval challenge” are highlighted in bold and yellow. Comprehensive evidence maps, if they represent complete inventories of the literature, should ultimately obviate the need for additional literature searches in systematic reviews conducted in response to the findings of a systematic evidence mapping exercise*..... **p.68**

Figure 2. *Illustration of how lack of knowledge of relations between concepts relevant to a research topic can result in evidence of potential importance to a given question being overlooked. In this example, awareness that DNA repair is obstructed by oxidative DNA damage allows lung cancer and leukemia to be connected to stressors which cause oxidative DNA damage to be incorporated into a cancer assessment. However, lack of awareness that replication forks regulate DNA repair may result in studies of stressors which stall replication*

forks by binding to cleavage complexes being excluded from cancer assessments
..... **p.70**

Figure 3. *The MeSH CV entry for “polycyclic aromatic hydrocarbons”, 21 July 2020*..... **p.73**

Figure 4. *The MeSH thesaurus entries for “polycyclic aromatic hydrocarbons”, 21 July 2020. For brevity, only first-level entries are shown* **p.74**

Figure 5. *The elements of an Adverse Outcome Pathway, whereby an exposure causes a Molecular Initiating Event, initiating a biological sequence of causally-related Key Events which result in a final Adverse Outcome being manifest. Experimental research can target how a challenge might affect a Key Event (Studies A, B, and C) or how one Key Event might cause another Key Event in a Key Event Relationship (Study D). Arranging biological events, exposures and the evidence around them in these sorts of AOP chains can be very valuable for integrating mechanistic evidence into chemical assessments but requires knowledge organization systems capable of reflecting the complexity and heterogeneity of the relationships and event types* **p.77**

Figure 6. *Existing biological ontologies can be used to define key events in computable terms and thereby make AOP information more interoperable with other toxicological data sources. The same can be done when describing the assays and biomarkers used to measure the key events. CHEBI = Chemical Entities of Biological Interest, PRO = Protein Ontology, GO = Gene Ontology, CL = Cell Ontology, UBERON = Uber Anatomy Ontology, MP = Mammalian Phenotype Ontology, MonDO = Mondo Disease Ontology, PCO = Population and Community Ontology, ECTO = Environment Exposure Ontology, BAO = BioAssay Ontology, EFO = Experimental Factor Ontology, SNOMED CT = SNOMED Clinical Terms, CHEAR = Children's Health Exposure Analysis Resource*..... **p.79**

Figure 7. *The workflow for matching natural language strings in research reports to a hierarchy of concepts in an ontology. Natural language information is extracted from included studies (e.g. phrases such as “increase in thyroid stimulating hormone”) into an evidence inventory (A). The terms “increase”, “thyroid”, “stimulating” and “hormone” are cleaned and mapped to ontological classes in preparation for integration with other data sets. The inventory can then be connected to other data models by mapping terminology between CVs (B). Done enough times, a large data inventory begins to accumulate* **p.81**

Chapter 5. Conclusions and Future Work

Figure 1. *The interplay between conduct standards, reporting standards, and critical appraisal tools in managing the quality of systematic review publications* **p.88**

Figure 2. *The beginnings of an approach to the mathematical description of external validity of studies included in a systematic review* **p.95**

Glossary and Abbreviations

Adverse Outcome Pathway: A way of formalizing, for risk assessment purposes, the steps by which a disease progresses from exposure through to final adverse outcome via increasing levels of biological complexity.

Bias: The systematic deviation of results or inferences from the truth.

Biological plausibility: A concept ambiguously defined in environmental health and chemical risk assessment, which generally refers to the extent to which a hypothetical association between an environmental exposure and a health outcome is grounded in existing biological knowledge.

Bisphenol-A (BPA). An organic synthetic compound which is a precursor to polycarbonates and epoxy resins, extensively used in food contact materials up to the mid-2010s and the subject of multiple controversial chemical risk assessments.

Chemical risk assessment (CRA): The determination of the probability of adverse health outcomes following exposure to chemical substances.

Chemical risk management: The process of ensuring that levels of exposure to a chemical substance do not exceed the tolerable thresholds determined by chemical risk assessment.

Consensus: General agreement, characterized by the absence of sustained opposition to any substantial issues under discussion and by a process that involves seeking to take into account the views of all parties concerned and to reconcile any conflicting arguments. Consensus need not imply unanimity.

Environmental health: the branch of public health concerned with investigating and/or mitigating factors in the environment that affect human health and disease.

European Chemicals Agency (ECHA): The agency of the European Union which manages the technical and administrative aspects of REACH.

European Food Safety Authority (EFSA): The agency of the European Union that provides scientific advice and communication on existing and emerging risks associated with the food chain.

External validity: The extent to which the results of an experiment apply to contexts outside that study, such as whether an effect observed in an experimental rat population would also be observed in a human population of concern.

Grading of Recommendations Assessment, Development and Evaluation (GRADE): A method for assessing the certainty in the evidence for effect estimates and the strength of recommendations in health care. GRADE is being adapted and applied to environmental health research.

Graph: A mathematical structure used to model pairwise relations between objects, made up of nodes which are connected by edges. In computing, a graph database uses graph structures for semantic queries. Graph databases can store statements in natural language as subject-predicate-object “triples”, with subjects and objects as nodes and predicates as edges.

Heterogeneity: Differences between studies. Heterogeneity can be statistical, referring to how studies have varying results, and methodological, referring to how studies can use varying designs to answer a given research question.

***In vitro* research:** Study models using microorganisms, cells, or biological molecules outside their normal biological context.

***In vivo* research:** Study models using whole, living organisms.

Indirectness: One of the key GRADE domains, concerned with the extent to which the evidence included in a systematic review addresses the review question.

Knowledge Organisation System (KOS): Technique for making existing information accessible to people, including ontologies, controlled vocabularies and thesauruses.

Lowest observed adverse effect level (LOAEL): The lowest concentration or amount of a substance that causes an adverse effect in a target organism, usually used as a benchmark of toxic exposure in a chemical risk assessment.

Methodological Expectations of Cochrane Intervention Reviews (MECIR): Cochrane standard for conduct of systematic reviews of healthcare interventions.

Narrative review: A broad concept which encompasses a number of different approaches to reviewing evidence, generally implying the use of methods which are based on an author's subjective judgement rather than review techniques designed to minimise bias. "Narrative" is also a technical term used in some areas of research synthesis to refer to review methods which do not deal with quantitative data. This meaning is not used in this Thesis.

National Toxicology Program Office of Health Assessment and Translation (NTP OHAT): A division of the US National Toxicology Program which conducts assessments of the potential for adverse effects on human health by chemical substances. Arguably the first government agency in the world to publish a framework for systematic review of health effects from exposure to chemical substances.

Ontology: A formal method for representing knowledge, usually within a particular knowledge domain, that relates terms or concepts to one another in a format that supports reading and searching not only for the terms themselves, but also for the relationships between those terms.

PECO statement: A mnemonic for Population Exposure Comparator Outcome statement, as a means of operationalising the formulation of questions in a systematic review

Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA): An evidence-based minimum set of items for reporting in systematic reviews and meta-analyses, focused on the reporting of reviews evaluating randomized trials, but can also be used as a basis for reporting systematic reviews of other types of research.

Recommendations for Conduct of Systematic reviews in Toxicology and Environmental health Research (COSTER): The first formally-developed set of recommendations for good practice in the conduct of environmental health systematic reviews.

Registration, Evaluation and Authorisation of Chemicals (REACH): European Union regulation addressing the production and use of chemical substances based on determination and management of the risks they pose human and environmental health.

Reporting standards for Systematic Evidence Syntheses in environmental research (ROSES): a collaborative initiative with the aim of improving the standards of reporting in evidence syntheses in environmental research. At the core of ROSES is a set of detailed forms for ensuring evidence syntheses report their methods to the highest possible standards.

Streetlight effect: The phenomenon by which research tends to be conducted in established areas of understanding rather than around novel ideas, often the result of it being easier to formulate questions around established concepts of known relevance to the problem rather than novel concepts of unknown relevance to the problem.

Systematic review (SR): a methodology for testing a research hypothesis using existing evidence, that employs techniques intended to maximise transparency of methods and minimize random and systematic error in deriving results.

The International Agency for Research on Cancer (IARC): An intergovernmental agency of the World Health Organization of the United Nations, the role of which is to conduct and coordinate research into the causes of cancer.

US Environmental Protection Agency (EPA): An independent executive agency of the United States federal government tasked with environmental protection matters

US Institute of Medicine (IOM): Renamed as the National Academy of Medicine, an American non-profit, non-governmental organisation which provides national

and international advice on issues relating to health, medicine, health policy, and biomedical science.

List of Appendices

| | |
|---|-----|
| Appendix A: Five Lessons..... | 118 |
| Appendix B: Protocols.io..... | 121 |
| Appendix C: Systematic Evidence Maps..... | 133 |
| Appendix D: Knowledge Graphs..... | 143 |
| Appendix E: NASEM Presentation..... | 158 |

Introduction

Background

Assessing risks to health posed by exposure to chemical substances

Chemical risk assessment is the determination of the probability of adverse health outcomes following exposure to chemical substances (National Research Council Committee on the Institutional Means for Assessment of Risks to Public Health, 2014). It consists of four steps: hazard identification, whereby the nature of the possible adverse health outcomes from exposure to the chemical are identified; hazard characterisation, whereby the relationship between exposure level and severity of occurrence of an outcome is determined; exposure assessment, whereby the level of the chemical to which a given population either is or can be expected to be exposed is quantified; and risk characterisation, whereby the probability of harm is calculated as a function of actual or expected exposure levels and the exposure-outcome relationship.

The results of a risk assessment process feed into risk management decisions about how to ensure levels of exposure to a substance do not exceed tolerable thresholds. Health risks from exposure to chemicals are managed through a wide variety of interventions, from placing regulatory restrictions on how much of a chemical may be used in consumer goods, through setting emissions limits on manufacturing operations, to requiring measures that limit exposure in occupational environments such as the wearing of protective equipment. The stages of a human health risk assessment are presented in Figure 1. Archetypal questions asked and addressed at each stage of the risk assessment and risk management process are also presented.

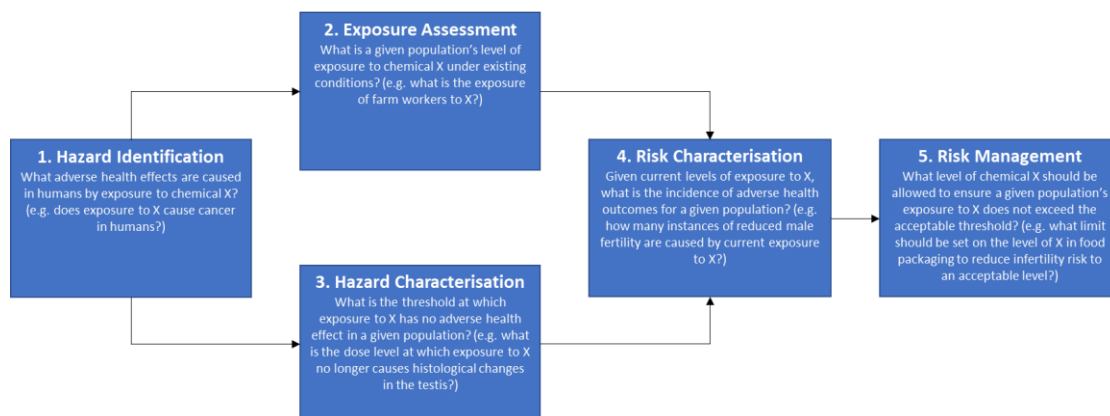


Figure 1. The components of a human-health risk assessment. Archetypal questions asked at each stage of the risk assessment and risk management process are included. The components are typical of wider environmental risk assessments. Adapted from World Health Organisation Chemical Risk Assessment Network (in prep).

With accuracy at a premium, there is a general expectation that the gathering of evidence for risk assessment be comprehensive and its evaluation be as objective as possible. This can, however, be a challenging expectation to meet, as the steps of the chemical risk assessment process draw on a range of fields of scientific research including environmental chemistry, toxicology (encompassing *in vivo*, *in vitro*, ecotoxicological and *in silico* computational methods), human epidemiology, and mathematical and statistical modelling. In spite of regulatory frameworks such as REACH (the Registration, Evaluation and Authorisation of Chemicals) emphasising the collation and analysis of all evidence relevant to evaluating risks of exposure to a given substance (Beronius et al., 2014), there has been long-standing concern about whether risk assessment processes are sufficiently scientifically robust (National Research Council, 2009).

One example which illustrates the problems with evaluating evidence of health risks from chemical exposures is in the range of contradictory opinions that expert scientists and reputable organisations have in the past held about the substance bisphenol-A (BPA). As a commonly-used food contact material, exposure to BPA had become ubiquitous by the early 20th Century. Concerns, however, were being raised about its potential to act in the body as an oestrogen (Vom Saal et al., 2012). This was heavily investigated by scientists, with almost 3,000 studies into the chemical indexed in the PubMed database by 2010 and the number of studies doubling in the following five years. Risk assessments of BPA, however, were highly inconsistent: by 2015, five

different authoritative organisations and researchers had come to incompatible conclusions about safe exposure levels to BPA, varying from “no concern for any age group” to “effects have been demonstrated [at] 1-4 magnitudes of order lower than the current LOAEL [lowest observed adverse effect level]” (Whaley et al., 2016).

These differences in conclusions have occurred in spite of each research group or agency committee ostensibly having access to the same body of scientific evidence about health risks from exposure to BPA. This should not necessarily be surprising: when a variety of expert groups interpret such a large, complex body of evidence differences in opinion should be expected. The experts will be exercising judgement from the varied backgrounds drawn on in risk assessment, with varying degrees of cognitive access to relevant information, while placing differing weight on individual studies and/or strands of evidence that they review and, when working in committee, potentially being more or less influenced by social dynamics in the group (Janis, 1983).

The problem is that when expert opinions are in conflict it can be very challenging to distinguish which conclusions are likely to represent the most valid synthesis of the totality of the available evidence. The objectivity of a process is also brought into question when it produces inconsistent results among those supposedly following it. This is not a sustainable situation for chemical risk assessment and an inadequate basis for regulatory interventions for risk management, which require consistency and certainty. The question, then, is whether it is possible to do better: can more consistency, transparency and objectivity be brought into the processes by which scientific evidence is evaluated in chemical risk assessment?

Systematic review as a potential solution to inconsistency in risk assessment

Chemical risk assessment has traditionally been dependent on what has been labelled by many as “narrative” approaches to describing what is known in answer to each question in the risk assessment process (Ågerstrand and Beronius, 2015; Beronius and Vandenberg, 2016; Rhomberg et al., 2013). As a term, “narrative” is a broad concept which encompasses a number of different approaches to reviewing evidence, from the caricature of one researcher writing about “my field, from my standpoint [...] using only my data and my ideas, and citing only my publications” (Caveman, 1999), to

thorough narrative critiques of comprehensively identified evidence as conducted by organisations such as IARC (IARC, 2019). (“Narrative” is also a term used to describe techniques for synthesising evidence without meta-analysis (Popay et al., 2006); this is not the meaning being discussed here.)

Whatever their specific type, it has been recognised that traditional narrative reviews are, to varying degrees, vulnerable to a range of methodological shortcomings which are likely to bias their summarisation of the evidence base (Chalmers et al., 2002). These include the potential for selective retrieval of evidence relevant to the review question, inconsistent interpretation of the impact of methodological shortcomings on the validity of findings of scientific studies, and often even an absence of clear review objectives (Mignini and Khan, 2006; Mulrow, 1987). As for risk assessments, when there exist multiple competing reviews, each using opaque methods, it becomes almost impossible to judge their relative merits and therefore to base decisions on the current best available evidence.

In medicine, it was increasingly clear by the early 1990s that dependence on narrative methods for evaluating the effectiveness of healthcare interventions was costing lives and wasting money (Chalmers and Glasziou, 2009). To solve this problem, the medical field began to incentivise widespread use of robust “systematic” review methods for answering questions in healthcare research. Systematic review is an approach to reviewing evidence which seeks to methodically “collate all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question,” using “explicit, systematic methods that are selected with a view to minimising bias” (Higgins et al., 2019). Systematic review has been enormously successful, rapidly becoming one of the most-cited forms of healthcare research (Patsopoulos et al., 2005), an integral step in planning research (Sutton et al., 2009) and vital to clarifying uncertainties about the effectiveness of medical interventions (Chalmers, 2010).

The potential value of systematic review methods for improving how evidence is reviewed in chemical risk assessment was arguably first articulated in the mid-2000s (Guzelian et al., 2005; Hoffmann and Hartung, 2006). This was followed by initial work at the University of California San Francisco on the *Navigation Guide* framework for conduct of systematic reviews in environmental health research (Woodruff and Sutton, 2010, 2011) and description by the European Food Safety Authority of the potential

benefits of systematic review in food and feed safety assessments (European Food Safety Authority, 2010). The first evaluation of how methods used in regulatory risk assessments compare to healthcare systematic reviews was published by the present author in 2013 (Whaley, 2013). In 2014, the Navigation Guide was formally published (Woodruff and Sutton, 2014) and in 2015 the US National Toxicology Program Office of Health Assessment and Translation issued the first government agency handbook for conduct of systematic reviews for health assessments (Rooney Andrew et al., 2014; US National Toxicology Panel, 2015).

Objectives and structure of this Thesis

Given the parallels between the challenge of evidence evaluation in chemical risk assessment as understood in 2015 and the situation in medicine which systematic review methods are intended to resolve, the overall aim of this Thesis is to investigate how systematic review methods can be applied to the conduct of chemical risk assessment. This overall aim is broken down into four specific objectives:

1. Identify practical challenges and knowledge gaps which impede the implementation of systematic review methods in chemical risk assessment;
2. Define a consensus view on key recommended practices for the planning and conduct of systematic reviews in the environmental health sciences;
3. Examine how “biological plausibility” as a concept fundamental to risk assessment is accommodated in systematic review methodologies;
4. Describe the role of ontologies in making evidence accessible for use in systematic chemical assessments.

The work in response to each objective is described in detail in four manuscripts in this Thesis. The first two manuscripts (Whaley et al., 2020, 2016) have been published in scientific journals. The third manuscript (Whaley et al., in prep) has passed the first round of the approval process for official publications of the international GRADE Working Group. The fourth manuscript (Whaley et al., submitted) has been resubmitted to a scientific journal after being revised in response to peer-review comments.

The conclusions of this Thesis are presented after the four papers. This final section describes how the broader field of systematic review methods in chemical risk assessment and environmental health research has progressed in relation to the objectives of this Thesis over the seven years since commencement of this PhD, and presents a set of research priorities which respond to that evolution.

Chapter 1. Challenges and opportunities

By 2014 systematic review had become increasingly viewed as a potentially powerful technique in assessing and communicating how likely it is that a chemical will cause health harm. However, it was not well understood at the time what various stakeholders perceived as being the main challenges in implementing systematic review methods in chemical risk assessment, nor how these challenges might practically be overcome.

The first objective of this Thesis is therefore to identify from expert practitioners the practical challenges and knowledge gaps to implementation of systematic review methods in chemical risk assessment, and to develop with them a roadmap for overcoming these obstacles and expediting the implementation of systematic methods by the various stakeholders involved in chemical risk assessment.

To achieve this, in November 2014 a one-day workshop was organised with participation of 35 scientists and researchers from the fields of medicine, toxicology, epidemiology, environmental chemistry, ecology, risk assessment, risk management and systematic review.

The workshop identified six characteristics of high quality chemical risk assessment. These included transparency of process and reasoning, validity of findings, statement of confidence in the evidence, utility and comprehensibility of assessment outputs, efficiency of use of resources, and reproducibility of results across multiple assessment teams. The limitations which traditional narrative review methods present in terms of delivering these six characteristics were contrasted with how risk assessment products might be improved if systematic methods were successfully implemented.

The workshop concluded that implementation of systematic methods in chemical risk assessment is a complex challenge, due to the multi-faceted, interdisciplinary nature of the type of work involved and the high level of heterogeneity of the evidence base

relevant to assessing health risks from exposure to chemical substances. The straightforward transferral of methods from healthcare systematic reviews is therefore not a realistic proposition. However, the participants were able to come to a consensus view on seven recommendations that would increase the likelihood of successful implementation of systematic review methods in chemical risk assessment.

Chapter 2. Recommended Practices

The second objective of this thesis responds to Recommendation #4 from Chapter 1, to contribute to the development of “a recognised ‘gold standard’ for SRs in toxicology and risk assessment”.

In 2016, while some handbooks and frameworks for conduct of systematic reviews had been published, there was no authoritative guidance written for the environmental health and chemical risk assessment community as to what criteria need to be fulfilled to render a literature review authentically systematic. While a number of handbooks, guidance and framework documents had been published, they were collectively inconsistent, individually incomplete, and sometimes made recommendations which would not necessarily be recognised by e.g. the medical community as being systematic practices.

To solve this problem, a second workshop was convened in follow-up to that which delivered Objective 1. The purpose of the second workshop was to develop an expert, cross-sector consensus view on a key set of recommended practices for the planning and conduct of systematic reviews in the environmental health sciences, including chemical risk assessments. This would serve as an authoritative guide as to what environmental health scientists and risk assessors should do if they are to conduct a review according to systematic methods.

The workshop and following consensus process yielded the *Conduct of Systematic Reviews in Toxicology and Environmental Health Research* (COSTER) recommendations, defining 70 systematic review practices across eight performance domains. The recommendations are accompanied by detailed descriptions of how the practices respond to the requirements of the environmental health and risk assessment context. As a first step in defining a widely accepted standard for conduct of systematic reviews, COSTER also proposes a set of activities which would further develop the

standard in future. Finally, the paper indicates areas in which systematic review methods have not yet been defined for environmental health contexts, so consensus on good practice cannot yet be established.

Chapter 3. Biological plausibility

Chapter 3 follows up on Recommendation #1 from Chapter 1 for “technical development of SR methodologies for CRA [chemical risk assessment] purposes” and the recommendation from Chapter 2 for work on research methods which could allow the development of “more detailed recommendations for appraising the external validity of included studies”.

To achieve this, Chapter 3 focuses on the concept of “biological plausibility” in environmental health systematic review. As a concept, “biological plausibility” is routinely used in chemical risk assessment when researchers are evaluating how confident they are in the results and inferences of a study or evidence review. When biological plausibility is high, the results of a study are more certain; when it is low, the credibility of a study is called into question and its utility in risk assessment is diminished.

Although widely-used in risk assessment, the exact definition of “biological plausibility” is ambiguous, with it being applied differently depending on the context of its use. “Biological plausibility” is purposefully not used in one of the most widely-used approaches for assessing certainty in the evidence which underpins the findings of a systematic review, the GRADE Framework (Guyatt et al., 2008; Schunemann et al., 2011). Nor is “biological plausibility” mentioned in the recommendations of COSTER.

The objective of Chapter 3 is therefore to determine whether “biological plausibility” is a concept which has been overlooked in developing systematic review methods for use in chemical risk assessment, or if the concept is already subsumed under other steps or concepts in the SR process.

Chapter 3 argues that “biological plausibility” is a concept which primarily comes into play when risk assessors need to include *in vivo* and *in vitro* studies in a review because evidence from observational studies in humans is of insufficient certainty for making decisions or drawing robust enough conclusions. This is a common occurrence in

chemical risk assessment, where evidence from human populations is usually very limited.

Through a series of 12 examples that specifically reference the “biological plausibility” of an inference from an experimental model to a real-world target context of concern, Chapter 3 argues that “biological plausibility” is functionally equivalent to assessment of the indirectness of the evidence (the extent to which existing research fits with the question being posed in a systematic review) within the GRADE Framework. That is to say, the concept of biological plausibility in traditional use in chemical risk assessment maps onto concepts already in use in systematic review, meaning that systematic review methods do not need to be extended to include biological plausibility as a domain-specific concept.

However, what is clear from the 12 examples is that in risk assessment contexts there is a lot more experience in and use of highly indirect evidence than is typically encountered in the healthcare and public health contexts in which systematic review methods were developed and GRADE is normally deployed. We therefore examine how toxicologists and risk assessors judge “biological plausibility” to gather important clues as to the sort of information which should be used when assessing the indirectness of evidence in environmental health systematic reviews.

Chapter 4. Ontologies

Chapter 4 responds to Recommendation #2c of Chapter 1 for development of tools to “support extraction, analysis and sharing of data from studies included in reviews”. Over the last three years this has become increasingly recognised as a critical issue in the successful application of systematic review methods to chemical risk assessment.

The reason such tools are needed relates to the almost extreme heterogeneity of the evidence base drawn on in environmental health research, as alluded to in the 12 examples of Chapter 3 which illustrate how studies included in a systematic review are informative of, but do not directly address, the populations, health outcomes and exposures of concern in a risk assessment. Tracing how these indirectly related concepts fit together for the purpose of drawing conclusions about risks to health presented by exposure to chemical substances is a collective endeavour which exceeds the individual capacity of any one researcher or research group. To do this in a way which is efficient

and can be shared between independent research groups requires “Knowledge Organisation Systems” which capture these conceptual relations; building these systems requires the development and implementation of risk assessment “ontologies”.

The objective of Chapter 4 is to describe what are “Ontologised Knowledge Organisation Systems” and characterise how they potentially enable the vast wealth of information available about health risks posed by exposure to chemical substances to be fully available to systematic reviews.

Chapter 4 achieves this via discussion of the “streetlight effect” in information retrieval and how it challenges the conduct of systematic reviews and evidence maps. The advantages and limitations of controlled vocabularies and thesauruses are highlighted as current approaches to addressing the streetlight effect, and then contrasted with the additional retrieval power which would be permitted by wholesale implementation of ontologies in environmental health research databases. Finally, the example of Adverse Outcome Pathways, as a relatively novel innovation in chemical risk assessment, is used to both illustrate the challenges in developing Knowledge Organisation Systems for chemical risk assessment and to outline a strategy for how these challenges can be overcome.

References

- Ågerstrand, M., Beronius, A. (2015) Weight of evidence evaluation and systematic review in EU chemical risk assessment: Foundation is laid but guidance is needed. *Environ. Int.* 92-93, 590–596.
- Beronius, A., Hanberg, A., Zilliacus, J., Rudén, C. (2014) Bridging the gap between academic research and regulatory health risk assessment of Endocrine Disrupting Chemicals. *Curr. Opin. Pharmacol.* 19, 99–104.
- Beronius, A., Vandenberg, L.N. (2016) Using systematic reviews for hazard and risk assessment of endocrine disrupting chemicals. *Rev. Endocr. Metab. Disord.*
- Caveman, A. (1999) The invited review? or, my field, from my standpoint, written by me using only my data and my ideas, and citing only my publications. *J. Cell Sci.* 113, 3125–3126.
- Chalmers, I. (2010) Systematic reviews and uncertainties about the effects of treatments. *Cochrane Database Syst. Rev.* 2011, ED000004.
- Chalmers, I., Glasziou, P. (2009) Avoidable waste in the production and reporting of research evidence. *Lancet* 374, 86–89.

- Chalmers, I., Hedges, L.V., Cooper, H. (2002) A brief history of research synthesis. *Eval. Health Prof.* 25, 12–37.
- European Food Safety Authority. (2010) Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA Journal* 8, 1637.
- Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., Schünemann, H.J., GRADE Working Group (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 336, 924–926.
- Guzelian, P.S., Victoroff, M.S., Halmes, N.C., James, R.C., Guzelian, C.P. (2005) Evidence-based toxicology: a comprehensive framework for causation. *Hum. Exp. Toxicol.* 24, 161–201.
- 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.
- Hoffmann, S., Hartung, T., (2006) Toward an evidence-based toxicology. *Hum. Exp. Toxicol.* 25, 497–513.
- IARC (2019) *IARC Monographs on the Identification of Carcinogenic Hazards to Humans: Preamble*.
- Janis, I.L. (1983) *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin, Boston, USA.
- Mignini, L.E., Khan, K.S. (2006) Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. *BMC Med. Res. Methodol.* 6, 10.
- Mulrow, C.D. (1987) The medical review article: state of the science. *Ann. Intern. Med.* 106, 485–488.
- National Research Council, Division on Earth and Life Studies, Board on Environmental Studies and Toxicology, Committee on Improving Risk Analysis Approaches Used by the U.S. EPA. (2009) *Science and Decisions: Advancing Risk Assessment*. National Academies Press.
- National Research Council Committee on the Institutional Means for Assessment of Risks to Public Health. (2014) *Risk Assessment in the Federal Government: Managing the Process*. National Academies Press (US), Washington (DC).
- Patsopoulos, N.A., Analatos, A.A., Ioannidis, J.P.A. (2005) Relative citation impact of various study designs in the health sciences. *JAMA* 293, 2362–2366.
- Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., Duffy, S. (2006) *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews. A Product from the ESRC Methods Programme* 211–219.

- Rhomberg, L.R., Goodman, J.E., Bailey, L.A., Prueitt, R.L., Beck, N.B., Bevan, C., Honeycutt, M., Kaminski, N.E., Paoli, G., Pottenger, L.H., Scherer, R.W., Wise, K.C., Becker, R.A. (2013) A survey of frameworks for best practices in weight-of-evidence analyses. *Crit. Rev. Toxicol.* 43, 753–784.
- Rooney A.A., Boyles A.L., Wolfe M.S., Bucher J.R., Thayer K.A. (2014) Systematic Review and Evidence Integration for Literature-Based Environmental Health Science Assessments. *Environ. Health Perspect.* 122, 711–718.
- Board on Population Health and Public Health Practice and Institute of Medicine. Roundtable on Environmental Health Sciences, Research, and Medicine. (2014) *The Challenge: Chemicals in Today's Society*. National Academies Press (US).
- Schunemann, H., Hill, S., Guyatt, G., Akl, E.A., Ahmed, F. (2011) The GRADE approach and Bradford Hill's criteria for causation. *Journal of Epidemiology & Community Health* 65, 392–395.
- Sutton, A.J., Cooper, N.J., Jones, D.R. (2009) Evidence synthesis as the key to more coherent and efficient research. *BMC Med. Res. Methodol.* 9, 29.
- US National Toxicology Panel. (2015) *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration*. Research Triangle Park, NC, USA. Available online: <https://ntp.niehs.nih.gov/whatwestudy/assessments/noncancer/handbook/index.html>
- Vom Saal, F.S., Nagel, S.C., Coe, B.L., Angle, B.M., Taylor, J.A. (2012) The estrogenic endocrine disrupting chemical bisphenol A (BPA) and obesity. *Mol. Cell. Endocrinol.* 354, 74–84.
- Whaley, P. (2013) Systematic review and the future of evidence in chemicals policy. Available online: <http://policyfromscience.com/wp-content/uploads/2013/11/PFS-Report-Electronic-Release-Version.pdf>
- Whaley, P., Aiassa, E., Beausoleil, C., Beronius, A., Bilotta, G., Boobis, A., de Vries, R., Hanberg, A., Hoffmann, S., Hunt, N., Kwiatkowski, C.F., Lam, J., Lipworth, S., Martin, O., Randall, N., Rhomberg, L., Rooney, A.A., Schünemann, H.J., Wikoff, D., Wolffe, T., Halsall, C. (2020) Recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER). *Environ. Int.* 143, 105926.
- Whaley, P., Edwards, S.W., Kraft, A., Nyhan, K., Shapiro, A., Watford, S., Wattam, S., Wolffe, T.A.M., Angrish, M. (submitted) Knowledge Organization Systems for Systematic Chemical Assessments.
- Whaley, P., Halsall, C., Agerstrand, M., Aiassa, E., Benford, D., Bilotta, G., Coggon, D., Collins, C., Dempsey, C., Duarte-Davidson, R., FitzGerald, R., Galay-Burgos, M., Gee, D., Hoffmann, S., Lam, J., Lasserson, T., Levy, L., Lipworth, S., Ross, S.M., Martin, O., Meads, C., Meyer-Baron, M., Miller, J., Pease, C., Rooney, A., Sapiets, A., Stewart, G., Taylor, D., 2016. Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations. *Environ. Int.* 92-93, 556–564.

- Whaley, P., Piggott, T., Morgan, R.L., Wikoff, D., Hoffmann, S., Tsaïoun, K., Thayer, K., Schünemann, H.J. (in prep) “Biological plausibility” and the analysis of indirect evidence in environmental health systematic reviews: a GRADE concept paper.
- Wolffe, T.A.M., Vidler, J., Halsall, C., Hunt, N., Whaley, P. (2020) A Survey of Systematic Evidence Mapping Practice and the Case for Knowledge Graphs in Environmental Health and Toxicology. *Toxicol. Sci.* 175, 35–49.
- Wolffe, T.A.M., Whaley, P., Halsall, C., Rooney, A.A., Walker, V.R. (2019) Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. *Environ. Int.* 130, 104871.
- Woodruff, T.J., Sutton, P. (2010) Pulling Back the Curtain: Improving Reviews in Environmental Health. *Environ. Health Perspect.* 118, a326–a327.
- Woodruff, T.J., Sutton, P. (2011) An Evidence-Based Medicine Methodology To Bridge The Gap Between Clinical And Environmental Health Sciences. *Health Aff.* 30, 931–937.
- Woodruff, T.J., Sutton, P. (2014) The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. *Environ. Health Perspect.* 122, 1007–1014.
- World Health Organisation Chemical Risk Assessment Network. (in prep) *A Framework for Conduct of Systematic Reviews in Chemical Risk Assessment*. World Health Organisation.

Chapter 1. Challenges and Opportunities

This chapter was published in the journal *Environment International*. The online version of the manuscript is available at this DOI: [10.1016/j.envint.2015.11.002](https://doi.org/10.1016/j.envint.2015.11.002)

According to the Contributor Roles Taxonomy, the candidate's contribution was as follows: conceptualisation; methodology; investigation; writing (original draft); writing (review and editing); project administration; funding acquisition.

Candidate: _____ Date: _____
Mr. Paul A. Whaley

Supervisor: _____ Date: _____
Prof. Crispin J. Halsall



Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint



Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations



Paul Whaley^a, Crispin Halsall^{a,*}, Marlene Ågerstrand^b, Elisa Aiassa^d, Diane Benford^c, Gary Bilotta^e, David Coggon^f, Chris Collins^w, Ciara Dempseyⁿ, Raquel Duarte-Davidson^g, Rex FitzGerald^h, Malyka Galay-Burgos^x, David Geeⁱ, Sebastian Hoffmann^j, Juleen Lam^k, Toby Lasserson^l, Len Levy^m, Steven Lipworthⁿ, Sarah Mackenzie Ross^o, Olwenn Martinⁱ, Catherine Meads^p, Monika Meyer-Baron^q, James Miller^r, Camilla Pease^s, Andrew Rooney^t, Alison Sapiets^u, Gavin Stewart^v, David Taylorⁿ

^a Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

^b Department of Environmental Science and Analytical Chemistry, Stockholm University, SE-106 91, Stockholm, Sweden

^c Food Standards Agency, Aviation House, 125 Kingsway, London WC2B 6NH, UK

^d Assessment and Methodological Support Unit, European Food Safety Authority, Via Carlo Magno 1/a 43126, Parma, Italy

^e Aquatic Research Centre, University of Brighton, Lewes Road, Brighton BN2 4GJ, UK

^f MRC Lifecourse Epidemiology Unit, University of Southampton, MRC Lifecourse Epidemiology Unit, Southampton General Hospital, Southampton SO16 6YD, UK

^g Centre for Radiation, Chemicals and Environmental Hazards, Public Health England, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0RQ, UK

^h Swiss Centre for Applied Human Toxicology, University of Basel, Missionsstrasse 64, 4055 Basel, Switzerland

ⁱ Institute for the Environment, Health and Societies, Brunel University London, Kingston Lane, Uxbridge UB8 3PH, UK

^j Evidence-Based Toxicology Collaboration (EBTC), Stembergring 15, 33106 Paderborn, Germany

^k University of California San Francisco, Program on Reproductive Health and the Environment, San Francisco, CA, USA

^l Cochrane Editorial Unit, Cochrane Central Executive, St Albans House, 57–9 Haymarket, London SW1Y 4QX, UK

^m Institute of Environment, Health, Risks and Futures, School of Energy, Environment and Agrifood, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK

ⁿ Royal Society of Chemistry, Burlington House, Piccadilly, London W1J 0BA, UK

^o Research Department of Clinical, Educational and Health Psychology, University College London, Gower Street, London WC1E 6BT, UK

^p Health Economics Research Group, Brunel University London, Kingston Lane, Uxbridge UB8 3PH, UK

^q Leibniz Research Centre for Working Environment and Human Factors (IfADo), Neurobehavioural Toxicology, Ardeystr 67, D-44139 Dortmund, Germany

^r Centre for Ecology and Hydrology, Wallingford, Oxfordshire OX10 8BB, UK

^s Ramboll Environ, 1 Broad Gate, The Headrow, Leeds LS1 8EQ, UK

^t National Institute of Environmental Sciences (NIEHS), National Institutes of Health (NIH), Department of Health and Human Services (DHHS), Research Triangle Park, NC, USA

^u Syngenta Ltd., Jealott's Hill International Research Centre, Bracknell RG42 6EY, UK

^v Centre for Rural Economy, School of Agriculture, Food and Rural Development, University of Newcastle upon Tyne, UK

^w Department of Geography and Environmental Science, School of Archaeology, Geography and Environmental Science, University of Reading, Reading, RG6 6DW, United Kingdom

^x European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), Avenue Edmond Van Nieuwenhuysse 2 Bte 8B-1160 Brussels, Belgium

ARTICLE INFO

Article history:

Received 8 August 2015

Accepted 2 November 2015

Available online 11 December 2015

Keywords:

Risk assessment
Research synthesis
Environment
Chemicals
Systematic review
Toxicology

ABSTRACT

Systematic review (SR) is a rigorous, protocol-driven approach designed to minimise error and bias when summarising the body of research evidence relevant to a specific scientific question. Taking as a comparator the use of SR in synthesising research in healthcare, we argue that SR methods could also pave the way for a "step change" in the transparency, objectivity and communication of chemical risk assessments (CRA) in Europe and elsewhere. We suggest that current controversies around the safety of certain chemicals are partly due to limitations in current CRA procedures which have contributed to ambiguity about the health risks posed by these substances. We present an overview of how SR methods can be applied to the assessment of risks from chemicals, and indicate how challenges in adapting SR methods from healthcare research to the CRA context might be overcome. Regarding the latter, we report the outcomes from a workshop exploring how to increase uptake of SR methods, attended by experts representing a wide range of fields related to chemical toxicology, risk analysis and SR. Priorities which were identified include: the conduct of CRA-focused prototype SRs; the development of a recognised standard of reporting and conduct for SRs in toxicology and CRA; and establishing a network to facilitate research, communication and training in SR methods. We see this paper as a milestone in the creation of a research climate that fosters communication between experts in CRA and SR and facilitates wider uptake of SR methods into CRA.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

* Corresponding author.

E-mail address: c.halsall@lancaster.ac.uk (C. Halsall).

1. Introduction

Systematic review (SR) is a rigorous, protocol-driven approach to minimising error and bias¹ in the aggregation and appraisal of evidence relevant to answering a research question. SR techniques were initially developed in the fields of psychology, social science and health care and have, since the 1980s, provided a valuable tool for evidence-informed decision-making across many domains (Lau et al., 2013). In medicine, SRs have provided a valuable response to the need for consistent, transparent and scientifically-robust interpretations of the results of increasing numbers of often conflicting studies of the efficacy of healthcare interventions. SRs have taken on an increasingly fundamental role both in supporting decision-making in healthcare and, by channelling resources towards questions for which the answers are not yet known, reducing waste in research (Chalmers and Glasziou, 2009; Salman et al., 2014). It is now accepted practice in healthcare to use SR methods to assess evidence not only for the efficacy of interventions, but also on diagnostic tests, prognostics and adverse outcomes.

The extension of SR techniques to other fields is based on a mutual need across disciplines to make the best use of existing evidence when making decisions, a move for which momentum has been growing for several decades. For example, the What Works Clearinghouse was established in 2002 to apply SR techniques in support of American educational policy (US Institute of Education Sciences, 2015), and in 2000 the international Campbell Collaboration research network was convened to undertake and disseminate systematic reviews on the effects of social interventions in diverse fields such as crime and justice, education, international development and social welfare (Campbell Collaboration, 2015). Meta-analysis and SR in ecology have contributed to evidence-based environmental policy since the mid-1990s (Stewart, 2010); more recently, the Collaboration for Environmental Evidence (CEE) has been established to encourage conduct of SRs on a wide range of environmental topics (Collaboration for Environmental Evidence, 2015).

The potential advantages of adapting SR methodology to the field of chemical risk assessment (CRA) have also been recognised, with multiple research groups and organisations either developing and adopting (Woodruff and Sutton, 2014; Birnbaum et al., 2013; European Food Safety Authority, 2010; Rooney et al., 2014; Aiassa et al., 2015) or recommending (US National Research Council, 2014a, 2014b; US Environmental Protection Agency, 2013; Silbergeld and Scherer, 2013; Hoffmann and Hartung, 2006; Zoeller et al., 2015) the use of SR methods for evaluating the association between health effects and chemical exposures to inform decision-making. There are, however, a number of recognised challenges in extending SR methods to CRA, many of which derive from key differences in the evidence base between the healthcare and toxicological sciences.

SRs in medicine often focus on direct evidence for benefits and adverse effects of healthcare interventions derived from randomised controlled trials (RCTs) in humans. The evidence base for CRA is generally more complex, with a need to extrapolate from investigations in animals, *in vitro* and *in silico*, and then to synthesise findings with those from human studies if available. Furthermore, the human data tend to come from observational studies with greater and more varied potential for bias and confounding than RCTs, and the range of outcomes to be

considered is usually much wider than in the assessment of healthcare interventions. Thus, when the various types of toxicological research are combined into a single overall conclusion about the health risks posed by a chemical exposure, reviewers are challenged with integrating the results from a broad and heterogeneous evidence base.

In spite of these differences, there is reason for thinking that SR methods can be applied successfully to CRA. For example, techniques for aggregating the results of different study types are already addressed in various frameworks currently in use in toxicology. These include: International Agency of Research on Cancer (IARC) Monographs (International Agency for Research on Cancer, 2006); the Navigation Guide (Woodruff and Sutton, 2014); and the US Office for Health Assessment and Translation (OHAT) (Rooney et al., 2014; US National Toxicology Panel, 2015) – though it should be noted that none of these approaches have yet applied SR methods to the exposure assessment component of CRA. Heterogeneous sources of evidence are a familiar challenge in all domains including clinical medicine (Lau et al., 1998), and SR of observational studies has a crucial role in identifying complications and side-effects of healthcare interventions (Sterne et al., 2014; Higgins and Green, 2011). The need for SR of pre-clinical animal trials of healthcare interventions, in order to better anticipate benefits and harms to humans, is another area in which methods being developed and implemented by a number of groups including SYRCLE (Hooijmans et al., 2012; van Luijk et al., 2014) and CAMARADES (Macleod et al., 2005; Sena et al., 2014). (Stewart and Schmid, 2015) argue that research synthesis methods (including systematic review) are generic and applicable to any domain if appropriately contextualised.

Given the sometimes controversial outcomes of CRAs and the growing public and media profile of the risks that chemicals may pose to humans and the environment, SR is increasingly viewed as a potentially powerful technique in assessing and communicating how likely it is that a chemical will cause harm. SR methods add transparency, rigour and objectivity to the process of collecting the most relevant scientific evidence with which to inform policy discussions and could provide a critical tool for organising and appraising the evidence on which chemical policy decisions are based.

Consequently, in November 2014 a group of 35 scientists and researchers from the fields of medicine, toxicology, epidemiology, environmental chemistry, ecology, risk assessment, risk management and SR participated in a one-day workshop to consider the application of SR in CRA. The purpose was three-fold:

1. Identify from expert practitioners in risk assessment and SR the obstacles, in terms of practical challenges and knowledge gaps, to implementing SR methods in CRA;
2. Develop a “roadmap” for overcoming those obstacles and expediting the implementation of SR methods, where appropriate, by the various stakeholders involved in CRA;
3. Establish the foundations of a network to co-ordinate research and activities relating to the implementation of SR methods in CRA. The aim would be to support best practise in the application of SR techniques and promote the wider adoption of SR in CRA, both in Europe and elsewhere.

Participants heard seven presentations about recent developments in SR methods, their application to the risk assessment process, and their potential value to policy-makers. There were two break-out sessions in which participants were divided into three facilitated groups, firstly to discuss challenges to implementing SR methods in CRA, and then to suggest ways in which the obstacles could be overcome. These ideas were discussed in plenary before being summarised, circulated for comment, and then published in this paper. The Workshop was conducted under the “Chatham House Rule” such that participants were free to refer to the information presented and discussed, provided they did not attribute it to identifiable individuals or organisations.

¹ It is worth drawing a distinction between three sources of bias in the review process. There is potential for bias in the conduct of a review (e.g. because of inappropriate methods for identifying and selecting evidence for inclusion in the review); bias because the material available for the review is not representative of the evidence base as a whole (due to selective publication); and bias arising from flaws in the design, conduct, analysis and reporting of individual studies included in the review that can cause the effect of an intervention or exposure to be systematically under- or over-estimated. One of the major functions of SRs is to minimise bias in the conduct of a review and, as far as possible, to ensure that potential bias from selective publication and methodological flaws in the evidence are properly taken into account when drawing conclusions in response to a research question.

The purpose of this overview paper is to present the rationale for exploring the application of SR methods to CRA, the various experts' views on the challenges to implementing SR methods in CRA, and their suggestions for overcoming them. The remaining goals of the meeting are ongoing work, including the development of the roadmap concept for publication and the establishment of a network for supporting the use of SR in CRA.

2. The appeal of SR methods in CRA

Chemical risk assessment is a multi-step process leading to a quantitative characterisation of risk, which can then be used to inform the management of chemical substances so as to ensure that any risks to human health or the environment are managed optimally. CRAs entail four fundamental steps: hazard identification; hazard characterisation (often a dose-response assessment); exposure assessment; and risk characterisation (see Fig. 1). These steps draw on various fields of scientific research including environmental chemistry, toxicology (encompassing in vivo, in vitro, ecotoxicological and in silico methods), ecotoxicology, human epidemiology, and mathematical modelling.

There are many ways in which errors can occur in the interpretation of evidence from these varied disciplines, including failure to consider all relevant data, failure to allow appropriately for the strengths and limitations of individual studies, and over- or underestimating the relevance of experimental models to real-world scenarios (to name a few). Whether the appraisal of evidence is based on objective processes, or on subjective expert judgement and opinion, may also be an important factor in accurate interpretation of evidence: the assessment process always requires input from technical experts, which inevitably brings an element of subjectivity to the interpretation of the scientific evidence. Different experts may have varying degrees of practical and cognitive access to relevant information, place differing weight on individual studies and/or strands of evidence that they review and, when working in committee, may be more or less influenced by dominant personalities. This can result in misleading conclusions in which the potential for health risks is overlooked, underestimated or overstated. Furthermore, if the factors determining their assessment of evidence are undocumented, when expert opinions are in conflict it can be very

challenging to distinguish which opinion is likely to represent the most valid synthesis of the totality of available evidence.

A recent illustrative example (see Box 1) of when expert scientists and reputable organisations have come to apparently contradictory conclusions about the likelihood of a chemical causing harm is the case of bisphenol-A (BPA). BPA is a monomer used in the manufacture of the resinous linings of tin cans and other food contact materials such as polycarbonate drinks bottles. It has been banned from use in infant-feed bottles across the EU (European Commission, 1/28/2011) because of "uncertainties concerning the effect of the exposure of infants to Bisphenol A" (European Commission, 5/31/2011b).

The European Food Safety Authority (EFSA) considers that current levels of exposure to BPA present a low risk of harm to the public (European Food Safety Authority, 2015a). The French food regulator ANSES takes a seemingly different stance on the risks to health posed by BPA (French Agency for Food, Environmental and Occupational Health, and Safety, 4/7/2014), determining there to be a "potential risk to the unborn children of exposed pregnant women". On this basis, ANSES has proposed classifying BPA as toxic to reproduction in humans (French Agency for Food, Environmental and Occupational Health, and Safety, 2013), a proposal which has contributed to the French authorities' decision to implement an outright ban on BPA in all food packaging materials (France, 12/24/2012). While the ban has been challenged by some stakeholders as being disproportionate under EU law (Tošenovský, 2014, 2015; Plastics Europe, 2015), the Danish National Food Institute has argued that EFSA has overestimated the safe daily exposure to BPA and that some populations are exposed to BPA at levels higher than can be considered safe (National Food Institute, Denmark, 2015); a view reflected in the conclusions of some researchers, e.g. (Vandenberg et al., 2014) but not others, e.g. (US Food and Drug Administration, 2014).

The example of BPA illustrates the challenges in reaching consensus even when interpreting the same evidence base regarding the potential toxicity of chemical exposures, either in terms of what is known and what is uncertain about the risks to health posed by BPA, and/or what response is appropriate to managing those risks and uncertainties. It also shows how, in the absence of that consensus, there is a danger that policy on BPA may become disconnected from the evidence base, either risking harm to health through continued exposure or incurring

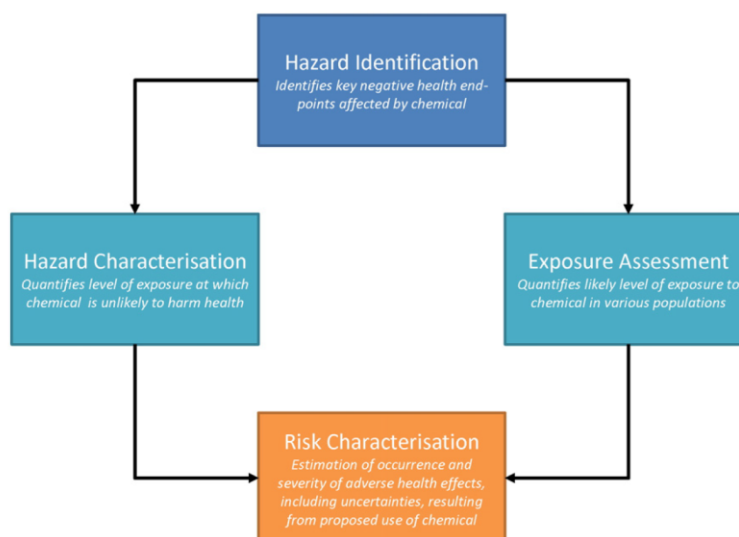


Fig. 1. An overview to the chemical risk assessment (CRA) process, whereby risk is a function of hazard and exposure. While SR methods could in principle be applied to all steps of the CRA process, it is the view of the workshop participants that up to this point in time most attention has been focused on the hazard identification and hazard characterisation steps. There are issues around conducting a systematic review for exposure assessment which were not discussed at the workshop, such as the requirement for a very different tool for assessing risk of bias in exposure studies which may necessitate specialised knowledge of analytical/environmental chemistry.

Five conflicting opinions about risks to health posed by bisphenol-A at current exposure levels

- “no health concern for any age group from dietary exposure and low health concern from aggregated exposure” (EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids (CEF) 2015)
- “The conclusions of the risk assessment show [...] a potential risk to the unborn children of exposed pregnant women. The identified effects relate to a change in the structure of the mammary gland in the unborn child, that could promote subsequent tumour development” (French Agency for Food, Environmental and Occupational Health & Safety 2013)
- “DTU evaluates that [EFSA’s TDI for BPA of] 4 µg/kg bw/day is not sufficiently protective with regards to endocrine disrupting effects of BPA. DTU finds that a TDI for BPA has to be 0.7 µg/kg bw/day or lower to be sufficiently protective” (National Food Institute, Denmark 2015)
- “BPA is safe at the current levels occurring in foods” (US Food and Drug Administration 2014)
- “we are confident that consistent, reproducible, low dose effects have been demonstrated for BPA [...] the doses that reliably produce effects in animals are 1–4 magnitudes of order lower than the current LOAEL of 50 mg/kg/day and many should be considered adverse” (Vandenberg et al. 2014)

Box 1. Examples of conflicting opinions from scientists and government agencies about the risks to health posed by bisphenol-A at current exposure levels.

unnecessary economic costs through restricting the use of a chemical which is in fact sufficiently safe. It also suggests that if the reasons for disagreement about health risks posed by a chemical are not accessible to various stakeholders in the debate, it then becomes much more difficult for regulators to credibly resolve controversies about chemical safety, potentially undermining their authority in the long term.

This example highlights the potential for differences in the interpretation of evidence when assessing chemical toxicity and the need for a process that is not only scientifically robust but also transparent, so that the reasons for any disagreement can be readily identified – including giving stakeholders greater opportunity to understand when differences in policy stem from divergent assessments of risk, and when they stem from divergent opinions as to how those risks are best managed. It also suggests the importance of the following characteristics in risk assessments that are used to inform risk management decisions:

1. *Transparency*, in that the basis for the conclusions of the risk assessment should be clear (otherwise they may not be trusted and errors may go undetected).
2. *Validity*, in that CRAs should be sufficiently (though not necessarily maximally) scientifically robust in their methodology and accurate in their estimation of risks and characterisation of attendant uncertainties as to optimise the decisions that must be made in risk management.
3. *Confidence*, providing the user with a clear statement as to the overall strength of evidence for the conclusions reached and a characterisation of the utility of the evidence for decision-making (e.g. “appropriate for hazard identification but inappropriate for identification of a reference dose”).
4. *Utility*, in that the output of the risk assessment should be in a form that is convenient and intelligible to those who will use it (outputs that are too detailed and complex to validate and readily comprehend lead to inefficiency and possibly erroneous decisions).
5. *Efficiency*, providing a clear justification of the choice of research question in the context of efficiently solving a CRA problem. Resources for CRA are often limited and it is wasteful to expend unnecessary effort on aspects of an assessment that will not be critical to decision-making (although for the purposes of transparency and validity, the reasons for focusing on a particular outcome or otherwise restricting the evaluation should be explained).

6. *Reproducibility*, in that the conclusions of the SR process when applied to the same question and data should ideally produce the same answer even when undertaken by different individuals (also described as “consistency”). In practise, different experts may reach different conclusions because they will not all make the same value judgments about the scope, quality and interpretation of evidence. Therefore, the process should be sufficiently rigorous that it is highly likely that scientific judgement would result in the same conclusion independent of the experts involved, and as a minimum the SR process should render transparent the reasons for all conclusions.

It may be perceived that the value of SR methods lies in their provision of unequivocal assessments of whether or not a chemical will induce specific harm to humans and/or wildlife in given circumstances. In practise, however, this will happen only if the evidence base is sufficiently extensive, there is unanimity in identification of the problem and in assessment of the quality of the evidence base, and also how the evidence is to be interpreted in answering the review question (without this, SRs will also produce different results). Often, the consensus and/or information may be relatively limited; in such circumstances, a SR will instead clearly state the limitations of the available data and consequent uncertainties. The value here is in the provision of a comprehensive and transparent assessment of what is *not* known and insight into the drivers of divergent opinion. From a research perspective, this yields valuable information about how research limitations and knowledge gaps contribute to ongoing uncertainty about environmental and health risks, allowing the subsequent efforts of researchers to be more clearly focused. From a policy perspective, SRs offer a transparent explanation as to why there are differences in opinion which can then be communicated to stakeholders.

Overall, SR contributes to achieving consensus not by eliminating expert judgement, nor by eliminating conflicting opinions about whether a compound should be banned (for example), but by providing a robust, systematic and transparent framework for reviewing evidence of risks, such that when there is disagreement, the reasons for it are clearly visible and the relative merits of differing opinions can be appraised. In this way, it may help to resolve controversies in the interpretation of the science which informs the risk management process.

3. SR and its application to CRA

3.1. Traditional vs. SR methods

SR methods are often contrasted with “traditional”, non-systematic narrative approaches to describing what is and is not already known in relation to a research question. In reality, the distinction between systematic and narrative review is a crude one, with narrative reviews encompassing a number of different approaches to reviewing evidence, from the caricature of one researcher writing about “my field, from my standpoint [...] using only my data and my ideas, and citing only my publications” (Caveman, 2000), to thorough narrative critiques of comprehensively identified evidence relevant to answering an explicitly articulated question, as conducted by organisations such as IARC (International Agency for Research on Cancer, 2006).

Nonetheless, it is worth noting that only relatively recently has it been recognised that traditional narrative reviews are, to varying degrees, vulnerable to a range of methodological shortcomings which are likely to bias their summarisation of the evidence base (Chalmers et al., 2002). These include selective rather than comprehensive retrieval of evidence relevant to the review topic, inconsistent interpretation of the impact of methodological shortcomings on the validity of included studies, and even an absence of clear review objectives or conclusions which are drawn directly from the strengths and limitations of the evidence base (Mulrow, 1987; Mignini and Khan, 2006).

The presence of these shortcomings seriously challenges the reader's ability to determine the credibility of a review. When there exist multiple competing reviews, each using opaque methods, it becomes almost impossible to judge their relative merits and therefore to base decisions on current best available evidence. The consequence is a proliferation of conflicting opinions about best practice that fail to take proper account of the body of research evidence. In the healthcare sciences, this was initially shown by Antman and colleagues when they found that, in comparison to recommendations of clinical experts, systematic aggregation of data from existing clinical trials of streptokinase to treat myocardial infarction would have demonstrated benefit some years before recommendations for its use became commonplace (Antman et al., 1992). More recently, cumulative meta-analyses have been shown to be more accurate in summarising current understanding of the size of effect of a wide range of healthcare interventions than researchers planning new clinical trials who have not used these methods (Clarke et al., 2014).

A SR is an approach to reviewing evidence which specifically sets out to avoid these problems, by methodically attempting “to collate all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question,” using “explicit, systematic methods that are selected with a view to minimising bias” (Higgins and Green, 2011).

In detail, this amounts to the pre-specification of the objective and methods of the SR in a written protocol, in which the aim of conducting the review is clearly stated as a structured question (for a SR of the effects of an intervention or exposure, this can establish a testable hypothesis or quantitative parameter that is to be estimated), along with the articulation of appropriate methods. The methods specified should include the techniques for identifying literature of potential relevance to the research question, the criteria for inclusion of the studies of actual relevance to the research question, how the internal validity² of the

included studies will be appraised, and the analytical techniques used for combining the results of the included studies. The purposes of the protocol are to discourage ad-hoc changes to methodology during the review process which may introduce bias, to allow any justifiable methodological changes to be tracked, and also to allow peer-review of the work that it is proposed, to help ensure the utility and validity of its objectives and methods.

The final SR itself consists of a statement of the objective, the search method, the criteria for including relevant studies for analysis, and the results of the appraisal of internal validity of the included studies, e.g. implemented as a “risk of bias” assessment in Cochrane Reviews of randomised trials (Higgins et al., 2011). The evidence is then synthesised using statistical meta-analytical techniques, narrative methods or both (depending on the extent to which meta-analysis is possible) into an overall answer to the research question. An assessment is then made of the strength of the evidence supporting the answer; in Cochrane reviews, this typically follows the GRADE methodology (Atkins et al., 2004), taking into account overall features of the evidence base including risk of bias across the included studies, publication bias in the evidence base, external validity or applicability of the evidence to the population of interest, heterogeneity of the evidence, and the overall precision of the evidence. This is finally followed by a concluding interpretation of what the SR as a whole determines is and is not known in relation to its objective.

In this, we emphasise the distinction between a SR and a meta-analysis. A meta-analysis pools the results of a number of separate studies in a single statistical analysis and may be a component of a SR; however, it does not necessarily incorporate the full set of methodological features which define the SR process (e.g. a meta-analysis may or may not include an assessment of the internal validity of included studies). While we acknowledge that some researchers use the terms “systematic review” and “meta-analysis” interchangeably, we believe the two approaches should be disambiguated. It is also worth noting that many reviews employ a combination of narrative and systematic methods; there were differing opinions among workshop participants as to the extent to which it is reasonable to expect all reviews to fully incorporate SR methods.

3.2. The current status of SR in environmental health, toxicology and CRA

While the use of SR methodologies is well established in healthcare to determine the effect of interventions on health outcomes or the accuracy of a diagnostic test, application of SR is relatively novel in the fields of toxicology and environmental health. Workshop participants heard how methods for SR of medical interventions have in the United States been adapted in both academic and federal contexts to the gathering and appraising of evidence for the effects of chemical exposures on human health: researchers at the University of California have developed the *Navigation Guide* (Woodruff and Sutton, 2014), and the US Office of Health Assessment and Translation (OHAT) at the US National Toxicology Program has developed the OHAT Framework for systematically reviewing environmental health research for hazard identification (Rooney et al., 2014).

The two approaches adapt the key elements of SR methods to questions in environmental health (which is directly relevant to the CRA process but does not include assessment of dose–response). Features that the two approaches have in common include: conducting a SR according to a pre-specified protocol; the development of a specific research question and use of “PECO” statements (see Box 2) in systematising review objectives and the methods that will be used to answer that question; an approach to appraising the internal validity of included studies adapted from the risk of bias appraisal tool developed by the Cochrane Collaboration (Higgins et al., 2011); an adaptation of the GRADE methodology (Atkins et al., 2004) for describing the certainty or strength of a body of evidence, incorporating risk of bias elements with other criteria such as for the assessment of relevance or

² “Internal validity” is a term used in Cochrane Collaboration guidance on conduct of SRs specifically intended to supersede the use of terms such as “methodological quality” or their equivalents, which are considered ambiguous (Higgins and Green, 2011). The internal validity of a piece of research is appraised in a “risk of bias” assessment. The target of the risk of bias assessment is the likelihood, magnitude and direction of systematic error in the size of an observed effect, as caused by flaws in the design, conduct, analysis and reporting of a study. Throughout this document, we follow Cochrane Collaboration conventions in using “internal validity” as a technical term in place of “methodological quality”.

“PECO” is an acronym representing: **Population** (the exposure group of interest, e.g. people of a certain age or rats in laboratory studies); **Exposure** (the compounds or exposure scenarios of interest, e.g. respiratory exposure to fine particulate matter); **Comparator** (the group to which the exposure group is being compared, e.g. vehicle-exposed controls in laboratory experiments or less exposed groups in epidemiological studies); **Outcome** (a deleterious change or marker thereof hypothesised to be brought about by the exposure). The purpose of a PECO statement is to provide a framework for developing the key question which a SR will answer, and also to determine the rationale for the inclusion and exclusion criteria that explicitly define which studies are relevant for the review.

Box 2. The use of PECO statements in the SR process.

external validity; and a methodology for combining the results of human and animal research into a statement of confidence about the hazard which a chemical poses to health.

Other tools are being developed to contribute to the systematic assessment of in vivo and ecotoxicity studies which have not been directly derived from Cochrane Collaboration methods. Presented at the Workshop was SciRAP (Science in Risk Assessment and Policy), a system developed to improve the consistency with which the relevance and reliability of studies are appraised in the context of conducting a chemical risk assessment for regulatory purposes. It is also intended to reduce the risk of selection bias in the risk assessment process by providing a mechanism for including non-standardised study methods yielding potentially valuable data (Beronius et al., 2014; SciRAP, 2014).

There are a number of other initiatives promoting and developing the use of SR methodologies in environmental and chemical risk assessment. Participants heard about how the European Food Safety Authority is integrating SR methods into its assessments of food and feed safety (European Food Safety Authority, 2015b, 2015c), and about the UK Joint Water Evidence Group methods for rapid and systematic assessments of evidence (Collins et al., 2014). Other coordinated initiatives include the Evidence-Based Toxicology Collaboration (Hoffmann and Hartung, 2006); the Collaboration for Environmental Evidence (Bilotta et al., 2014a; Land et al., 2015); and the Systematic Review Centre for Laboratory Animal Experimentation (SYRCLE).

3.3. Overcoming the challenges in implementing SR methods in CRA

Risk assessment for a chemical or group of chemicals is a multifaceted process that normally requires consideration of multiple endpoints in relation to a variety of exposure scenarios, integrating evidence from epidemiological studies, bioassays in animals, mechanistic studies and studies within the distribution and determinants of exposure by different pathways and routes. In addition to resolving methodological issues relating to underdeveloped methods (e.g. how SR methods can be used as part of dose–response assessment or how they can be applied to exposure assessment), it is important to consider how SR should fit into the CRA process. One challenge going forward is to explore the circumstances in which applying more rigorous SR methods to assess scientific evidence would be warranted, which would require insight into the practicality and cost-effectiveness of applying such methods in those situations.

In principle, it should be possible to conduct SRs in any aspect of a CRA. Given the success in employing SR methods to support evidence-based practice in healthcare, it is intuitive that SRs could address specific questions arising within toxicology, human epidemiology and environmental health (e.g. hazard assessment within a CRA) and this view appears to be gaining momentum within the environmental health literature. The SR method may also lend itself to answering questions concerning e.g. the accuracy of the reported physical-chemical properties of a substance, doses predicted by quantitative exposure

assessment, concentrations of a chemical in the environment and biota, and the derivation of a No Observed Adverse Effect Level (NOAEL) or Benchmark Dose Lower 95% confidence limit (BMDL). European Food Safety Authority (2015c) explores these issues in more detail.

Depending on scope, the resources (time and cost) to undertake an SR can be considerable. Currently there is a lack of empirical evidence relating to the resource-effectiveness of SR approaches in CRA and there was a difference of opinion among workshop participants as to whether the effort required for conducting a SR tends to be under- or overestimated. It was suggested that, where effort is likely to be substantial, efficient use of resources may be achieved by focusing on high-value questions developed through initial scoping exercises. For example, a low-dose adverse effect may be evident in animal models and supported to some extent by human epidemiology and hence a question may be formulated around this initial evidence; there may be little point, however, in pursuing a question related to non-carcinogenic toxicity in wildlife if a substantial part of the literature points towards that substance being a potential human carcinogen. There is also growing interest in rapid reviews, when full SR methods are considered overly onerous (Collins et al., 2014; Schünemann and Moja, 2015).

The priorities for expediting the adaptation of SR methods to CRA identified at the Workshop are as follows:

1. The development of a number of prototype CRA-focused SRs to explore how readily SR procedures can be integrated into the CRA process, to:
 - a. identify additional methodological challenges in adapting SR methods to the CRA context and develop techniques to address them;
 - b. acquire practical experience in managing resources when conducting SRs in CRA, including the conduct of scoping exercises for identifying high-value review questions, the further development and/or application of novel “rapid evidence review” methods (UK Civil Service, 2015), and how SR methods can be integrated into existing regulatory structures such as REACH (see Box 3) (European Chemicals Agency (2/26/2015)).
2. Technical development of SR methodologies for CRA purposes, in particular the further advancement of techniques for appraising and synthesising mechanistic, toxicological and human epidemiological studies, to include:
 - a. refining tools for more consistent and scientifically robust appraisal of the internal validity of individual studies included in a CRA and the implications for interpretation of their findings; see e.g. Bilotta et al. (2014b). This might include further development and validation of tools such as the SYRCLE methodology for assessing the internal validity of animal studies (Hooijmans et al., 2014); for SR of observational studies see e.g. Sterne et al. (2014),

Systematic review and REACH regulations

Regulations such as REACH emphasise collating at the point of registration all evidence relevant to evaluating risks to human and environmental health posed by a chemical. As yet, however, there is very little guidance on how registrants should assemble REACH-compliant dossiers, nor is there detailed guidance on how the assembled evidence is to be assessed (Beronius et al. 2014). The subsequent quality of many of the REACH registration dossiers, with 172 out of 283 compliance checks resulting in a request for further information (European Chemicals Agency 2/26/2015), suggests a need for the development of a standardised, scientifically robust approach to dossier assembly which can be consistently followed by registrants.

Box 3. The potential utility of SR methods in application to REACH registrations.

the methods employed in the NTP/OHAT and Navigation Guide protocols, and the applicability of other assessment methods such as SciRAP (Beronius et al., 2014);

- b. the development of tools for the hazard characterisation and exposure assessment components of the CRA process;
 - c. the further development of software akin to the Cochrane Collaboration's Review Manager (Nordic Cochrane Centre, 2014) and the Systematic Review Data Repository (Ip et al., 2012), and tools such as DRAGON (ICF International, 2015) and the Health Assessment Workspace Collaborative (Rusyn and Shapiro, 2013) to support extraction, analysis and sharing of data from studies included in reviews;
3. The development an empirical evidence base for the different types of bias that operate in the CRA domain, including their direction and potential magnitude, and the extent to which any methods being adopted to address them are appropriate and effective.
 4. The development of a recognised "gold standard" for SRs in toxicology and risk assessment equivalent to the Cochrane Collaboration in evidence-based medicine, to address the growing number of purported SRs of unclear validity which are increasingly prevalent in the environmental health literature.
 5. The creation of a climate of constructive discussion that fosters advancement of methods whereby chemical risk practitioners, industry, competent authorities, academic researchers and policy makers can research, discuss and evaluate SR methods and the potential advantages they can bring.
 6. The establishment of a network of scientists and CRA practitioners to pursue research into and discussion of SR methodologies and facilitate their implementation.
 7. The implementation of training programmes for risk assessment practitioners and stakeholders, focusing specifically on application

of SR methods to CRA as a complement to current courses which largely cover SR methods in healthcare.

4. Conclusions

While systematic review methods have proven highly influential in healthcare, they have yet to make widespread impact on the process of chemical risk assessment. While there is much promise in the concept of adapting SR methods to CRA to give definitive answers to specified research questions, or to enable identification of the reasons for failure to resolve debate, a number of challenges to implementing SR methods in CRA have been identified. These include particular concerns about approaches to assessing bias and confounding in observational studies, the effort involved in conducting SRs, and the subsequent benefits of conforming to SR standards. Recent experience from both regulatory agencies and academics already yields some clear recommendations which would expedite the wider implementation of SR methods in CRA, potentially increasing the efficiency, transparency and scientific robustness of the CRA process.

Disclaimer

The views expressed in this manuscript are those of the authors and do not necessarily represent the views or policies of their employers or otherwise affiliated organisations. EA is employed by the European Food Safety Authority (EFSA); however, the present article is published under her sole responsibility and may not be considered as an EFSA scientific output.

Acknowledgements

Funding for the workshop was provided through the Economic & Social Science Research Council grant "Radical Futures in Social Sciences" (Lancaster University) and Lancaster Environment Centre. CH, PW, AR are grateful to Lancaster University's Faculty of Science & Technology "Distinguished Visitors" funding programme. The Royal Society of Chemistry is acknowledged for generously providing a meeting room, refreshments and facilitating the workshop proceedings. The PhD studentship of PW is partly funded through Lancaster Environment Centre. The contribution of non-author workshop participants to the development of the manuscript is also greatly appreciated.

References

- Aiassa, E., Higgins, J.P.T., Frampton, G.K., Greiner, M., Afonso, A., Amzal, B., et al., 2015. Applicability and feasibility of systematic review for performing evidence-based risk assessment in food and feed safety. *Crit. Rev. Food Sci. Nutr.* 55 (7), 1026–1034. <http://dx.doi.org/10.1080/10408398.2013.769933>.
- Antman, E.M., Lau, J., Kupelnick, B., Mosteller, F., Chalmers, T.C., 1992. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *Treatments for myocardial infarction. JAMA* 268 (2), 240–248.
- Atkins, D., Best, D., Briss, P.A., Eccles, M., Falck-Ytter, Y., Flottorp, S., et al., 2004. Grading quality of evidence and strength of recommendations. *BMJ (Clinical research Ed.)* 328 (7454), 1490. <http://dx.doi.org/10.1136/bmj.328.7454.1490>.
- Beronius, A., Molander, L., Rudén, C., Hanberg, A., 2014. Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: a proposal to improve evaluation criteria and reporting. *Journal of Applied Toxicology: JAT* 34 (6), 607–617. <http://dx.doi.org/10.1002/jat.2991>.
- Bilotta, G.S., Milner, A.M., Boyd, I., 2014a. On the use of systematic reviews to inform environmental policies. *Environ. Sci. Pol.* 42, 67–77. <http://dx.doi.org/10.1016/j.envsci.2014.05.010>.
- Bilotta, G.S., Milner, A.M., Boyd, I.L., 2014b. Quality assessment tools for evidence from environmental science. *Environ. Evid.* 3 (1), 14. <http://dx.doi.org/10.1186/2047-2382-3-14>.
- Birnbaum, L.S., Thayer, K.A., Bucher, J.R., Wolfe, M.S., 2013. Implementing systematic review at the National Toxicology Program: status and next steps. *Environ. Health Perspect.* 121 (4), A108–A109. <http://dx.doi.org/10.1289/ehp.1306711>.
- Campbell Collaboration, 2015. The Campbell Collaboration. Available online at <http://www.campbellcollaboration.org/>, accessed 6/13/2015.
- Caveman, 2000. The invited review – or, my field, from my standpoint, written by me using only my data and my ideas, and citing only my publications. *J. Cell Sci.* 113 (Pt 18), 3125–3126.

- Chalmers, I., Glasziou, P., 2009. Avoidable waste in the production and reporting of research evidence. *Lancet* 374 (9683), 86-89. [http://dx.doi.org/10.1016/S0140-6736\(09\)60329-9](http://dx.doi.org/10.1016/S0140-6736(09)60329-9).
- Chalmers, I., Hedges, L.V., Cooper, H., 2002. A brief history of research synthesis. *Evaluation & the Health Professions* 25 (1), 12-37. <http://dx.doi.org/10.1177/016327870205001003>.
- Clarke, M., Brice, A., Chalmers, I., 2014. Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. *PLoS ONE* 9 (7), e102670. <http://dx.doi.org/10.1371/journal.pone.0102670>.
- Collaboration for Environmental Evidence (2015): The Collaboration for Environmental Evidence. Available online at <http://www.environmentalevidence.org/>, accessed 6/13/2015.
- Collins, A., Miller, J., Coughlin, D., Kirk, S., 2014. The Production of Quick Scoping Reviews and Rapid Evidence Assessments: A How to Guide (Beta Version 2). Joint Water Evidence Group. Available online at <https://sbri.innovateuk.org/documents/3058188/3918930/The+Production+of+QSRs+and+REAs+-+A+How+to+guide.pdf/45975020-be7d-4788-b74b-f3b6ed32c73a>.
- European Chemicals Agency (2/26/2015), Evaluation under REACH Progress Report 2014. Available online at: http://echa.europa.eu/documents/10162/13628/evaluation_report_2014_en.pdf. Accessed 4/11/2015.
- European Commission (1/28/2011): Directive 2011/8/EU of 28 January 2011 amending Directive 2002/72/EC as regards the restriction of use of Bisphenol A in plastic infant feeding bottles, Directive 2011/8/EU. In : Official Journal of the European Union. Available online at <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:026:0011:0014:EN:PDF>, accessed 6/15/2015.
- European Commission (5/31/2011): Bisphenol A: EU ban on baby bottles to enter into force tomorrow. Brussels. Available online at http://europa.eu/rapid/press-release_IP-11-664_en.htm, accessed 2/17/2015.
- European Food Safety Authority (2010): Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA Journal* 2010; 8(6):1637. DOI: <http://dx.doi.org/10.2903/j.efsa.2010.1637>.
- European Food Safety Authority (2015a): No consumer health risk from bisphenol A exposure. Press Release 21 Jan 2015. Parma. Available online at <http://www.efsa.europa.eu/en/press/news/150121.htm>, accessed 2/18/2015.
- European Food Safety Authority (2015b): Outcome of the targeted consultation of the EFSA Journal editorial on increasing openness, robustness and transparency of scientific assessments. Available online at <http://www.efsa.europa.eu/en/supporting/pub/785e.htm>, accessed 8/4/2015.
- European Food Safety Authority, 2015c. Principles and process for dealing with data and evidence in scientific assessments. *EFSA Journal* 13 (5), 4121. <http://dx.doi.org/10.2903/j.efsa.2015.4121>.
- France (12/24/2012): LOI no. 2012-1442 du 24 décembre 2012 visant à la suspension de la fabrication, de l'importation, de l'exportation et de la mise sur le marché de tout conditionnement à vocation alimentaire contenant du bisphénol A. *Legifrance.gouv.fr*. Available online at http://legifrance.gouv.fr/affichTexte.do;jsessionid=F6553AACC19D178279D8DF154EAC8558.tpdila17v_17cidTexte=JORFTEXT000026830015, accessed 6/15/2015.
- French Agency for Food, Environmental and Occupational Health & Safety (2013): Bisphenol A: ANSES demonstrates potential health risks and confirms the need to reduce exposure. Available online at <https://www.anses.fr/en/content/bisphenol-anses-demonstrates-potential-health-risks-and-confirms-need-reduce-exposure>.
- French Agency for Food, Environmental and Occupational Health & Safety (4/7/2014): Bisphenol A: ANSES publishes its comments in response to the EFSA draft opinion for consultation. Available online at <https://www.anses.fr/en/content/bisphenol-anses-publishes-its-comments-response-efsa-draft-opinion-consultation>, accessed 2/18/2015.
- Higgins, J.P.T., Green, S. (Eds.) (2011): Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available online at <http://handbook.cochrane.org/>, accessed 2/18/2015.
- Higgins, J.P.T., Altman, D.G., Gotzsche, P.C., Jüni, P., Moher, D., Oxman, A.D., et al., 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 343, d5928. <http://dx.doi.org/10.1136/bmj.d5928>.
- Hoffmann, S., Hartung, T., 2006. Toward an evidence-based toxicology. *Hum. Exp. Toxicol.* 25 (9), 497-513. <http://dx.doi.org/10.1191/09603227106het648oa>.
- Hooijmans, C.R., Rovers, M., de Vries, R.B., Leenaars, M., Ritskes-Hoitinga, M., 2012. An initiative to facilitate well-informed decision-making in laboratory animal research: report of the First International Symposium on Systematic Reviews in Laboratory Animal Science. *Lab. Anim.* 46 (4), 356-357. <http://dx.doi.org/10.1258/la.2012.012052>.
- Hooijmans, C.R., Rovers, M.M., de Vries, R.B.M., Leenaars, M., Ritskes-Hoitinga, M., Langendam, M.W., 2014. SYRCL's risk of bias tool for animal studies. *BMC Med. Res. Methodol.* 14, 43. <http://dx.doi.org/10.1186/1471-2288-14-43>.
- ICF International (2015): DRAGON: an online tool for systematic review. Available online at <http://www.icfi.com/insights/products-and-tools/dragon-online-tool-systematic-review>, accessed 8/4/2015.
- International Agency for Research on Cancer (2006): Preamble to the IARC Monographs. Lyon, France. Available online at <http://monographs.iarc.fr/ENG/Preamble/index.php>, accessed 9/10/2015.
- Ip, S., Hadar, N., Keefe, S., Parkin, C., Iovin, R., Balk, E.M., Lau, J., 2012. A Web-based archive of systematic review data. *Systematic Reviews* 1, 15. <http://dx.doi.org/10.1186/2046-4053-1-15>.
- Land, M., de Wit, C.A., Cousins, I.T., Herzke, D., Johansson, J., Martin, J.W., 2015. What is the effect of phasing out long-chain per- and polyfluoroalkyl substances on the concentrations of perfluoroalkyl acids and their precursors in the environment? A systematic review protocol. *Environ. Evid.* 4 (1), 3. <http://dx.doi.org/10.1186/2047-2382-4-3>.
- Lau, J., Ioannidis, J.P.A., Schmid, C.H., 1998. Summing up evidence. One answer is not always enough. *Lancet* 351 (9096), 123-127. [http://dx.doi.org/10.1016/S0140-6736\(97\)08468-7](http://dx.doi.org/10.1016/S0140-6736(97)08468-7).
- Lau, J., Rothstein, H.R., Stewart, G.B., 2013. History & progress of meta-analysis. In: Koricheva, J., Gurevitch, J., Mengersen, K. (Eds.), *Handbook of Meta-analysis in Ecology and Evolution*. Princeton University Press, Princeton Chapter 25.
- Macleod, M.R., Ebrahim, S., Roberts, I., 2005. Surveying the literature from animal experiments: systematic review and meta-analysis are important contributions. *BMJ* 331 (7508), 110. <http://dx.doi.org/10.1136/bmj.331.7508.110-b>.
- Mignini, L.E., Khan, K.S., 2006. Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. *BMC Med. Res. Methodol.* 6, 10. <http://dx.doi.org/10.1186/1471-2288-6-10>.
- Mulrow, C.D., 1987. The medical review article: state of the science. *Ann. Intern. Med.* 106 (3), 485-488.
- National Food Institute, Denmark (2015): Evaluation of EFSA's new scientific opinion on bisphenol A. Søborg, Denmark (REG-no. DK 30 06 09 46). Available online at http://www.food.dtu.dk/english/~media/Institutter/Foedevareinstituttet/Publikationer/Pub-2015/Evaluation_BisphenolA.aspx?la=da.
- Nordic Cochrane Centre (2014): Review Manager (RevMan), Version 5.3: Cochrane Collaboration. Available online at <http://tech.cochrane.org/revman>, accessed 6/18/2015.
- Plastics Europe (1/15/2015): French ban on the use of Bisphenol A (BPA) in food contact: In conflict with European law and risk assessment – severe distortion of the market – no safety benefit for consumers. Jasmin Bird. Available online at http://www.bisphenol-a-europe.org/uploads/Modules/Mediaroom/stm_re_french-bpa-ban-ban-enforced-01-01-2015.pdf, accessed 6/15/2015.
- Rooney, A.A., Boyles, A.L., Wolfe, M.S., Bucher, J.R., Thayer, K.A., 2014. Systematic review and evidence integration for literature-based environmental health science assessments. *Environ. Health Perspect.* 122 (7), 711-718. <http://dx.doi.org/10.1289/ehp.1307972>.
- Rusyn, I., Shapiro, A. (2013): Health Assessment Workspace Collaborative (HAWC). Version Solid Hammer: UNC-CH Software. Available online at <https://hawcproject.org/>, accessed 8/4/2015.
- Salman, R.A.-S., Beller, E., Kagan, J., Hemminki, E., Phillips, R.S., Savulescu, J., et al., 2014. Increasing value and reducing waste in biomedical research regulation and management. *Lancet* 383 (9912), 176-185. [http://dx.doi.org/10.1016/S0140-6736\(13\)62297-7](http://dx.doi.org/10.1016/S0140-6736(13)62297-7).
- Schünemann, H.J., Moja, L., 2015. Reviews: Rapid! Rapid! Rapid! ... and systematic. *Systematic Reviews* 4, 4. <http://dx.doi.org/10.1186/2046-4053-4-4>.
- SciRAP (2014): Science in risk assessment and policy. Department of Applied Environmental Science at Stockholm University; Institute of Environmental Medicine at Karolinska Institutet in Stockholm; MistraPharma. Available online at <http://www.sciarp.org/>, accessed 3/11/2015.
- Sena, E.S., Currie, G.L., McCann, S.K., Macleod, M.R., Howells, D.W., 2014. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *J. Cereb. Blood Flow Metab.* 34 (5), 737-742. <http://dx.doi.org/10.1038/jcbfm.2014.28>.
- Silbergeld, E., Scherer, R.W., 2013. Evidence-based toxicology: strait is the gate, but the road is worth taking. *ALTEX* 30 (1), 67-73.
- Sterne, J.A.C., Higgins, J.P.T.; Reeves, B.C. (2014): A Cochrane Risk of Bias Assessment Tool for Non-Randomized Studies of Interventions (ACROBAT-NRSI). The Cochrane Collaboration. Available online at <https://sites.google.com/site/riskofbiastool/>, accessed 9/29/2014.
- Stewart, G., 2010. Meta-analysis in applied ecology. *Biol. Lett.* 6 (1), 78-81. <http://dx.doi.org/10.1098/rsbl.2009.0546>.
- Stewart, G.B., Schmid, C.H., 2015. Lessons from meta-analysis in ecology and evolution: the need for trans-disciplinary evidence synthesis methodologies. *Research Synthesis Methods* 6 (2), 109-110. <http://dx.doi.org/10.1002/jrsm.1152>.
- Tošenovský, E. (2014): Question for written answer to the Commission, Rule 130. European Parliament, Parliamentary questions, P-008546/2014. Subject: possible negative impact on the internal market of measures concerning BPA adopted by the French authorities. 30 October 2014. Available online at <http://www.europarl.europa.eu/sides/getDoc.do?type=WQ&reference=P-2014-008546&language=EN>, accessed 6/15/2015.
- Tošenovský, E. (2015): Question for written answer to the Commission, Rule 130. European Parliament, Parliamentary questions, E-004315-15. Subject: measures concerning Bisphenol A. 17 March 2015. Available online at <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP/TEXT+WQ+E-2015-004315+0+DOC+XML+V0/EN&language=en>, accessed 6/15/2015.
- UK Civil Service (2015): What is a rapid evidence assessment? Available online at <http://www.civilservice.gov.uk/networks/gsr/resources-and-guidance/rapid-evidence-assessment/what-is>, accessed 6/13/2015.
- US Environmental Protection Agency (2013): Process for developing IRIS health assessments. Available online at <http://www.epa.gov/IRIS/process.htm>, accessed 6/16/2015.
- US Food and Drug Administration (2014): Bisphenol A (BPA): use in food contact application. Update on Bisphenol A (BPA) for Use in Food Contact Applications. Available online at <http://www.fda.gov/NewsEvents/PublicHealthFocus/ucm064437.htm>.
- US Institute of Education Sciences (2015): What works clearinghouse. Available online at <http://ies.ed.gov/ncee/www/default.aspx>, accessed 6/13/2015.
- US National Research Council, 2014a. A Framework to Guide Selection of Chemical Alternatives. The National Academies Press, Washington, D.C.
- US National Research Council, 2014b. Review of EPA's Integrated Risk Information System (IRIS) Process. The National Academies Press, Washington, D.C.
- US National Toxicology Panel (2015): Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. Available online at http://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf, accessed 1/13/2015.
- van Luijk, J., Bakker, B., Rovers, M.M., Ritskes-Hoitinga, M., de Vries, R.B.M., Leenaars, M., 2014. Systematic reviews of animal studies; missing link in translational research? *PLoS ONE* 9 (3), e89981. <http://dx.doi.org/10.1371/journal.pone.0089981>.

Applying Systematic Review Methods in Chemical Risk Assessment - Chapter 1. Challenges and Opportunities

564

P. Whaley et al. / *Environment International* 92-93 (2016) 556-564

Vandenberg, L.N., Ehrlich, S., Belcher, S.M., Ben-Jonathan, N., Dolinoy, D.C., Hugo, E.R., et al., 2014. Low dose effects of bisphenol A. *Endocrine Disruptors* 1 (1), e26490. <http://dx.doi.org/10.4161/endo.26490>.

Woodruff, T.J., Sutton, P., 2014. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into

better health outcomes. *Environ. Health Perspect.* 122 (10), 1007-1014. <http://dx.doi.org/10.1289/ehp.1307175>.

Zoeller, R.T., Bergman, Å., Becher, G., Bjerregaard, P., Bormann, R., Brandt, I., et al., 2015. A path forward in the debate over health impacts of endocrine disrupting chemicals. *Environ. Heal.* 14, 118. <http://dx.doi.org/10.1186/1476-069X-13-118>.

Chapter 2.

Recommended Practices

This chapter was published in the journal *Environment International*. The online version of this manuscript is available at this DOI: [10.1016/j.envint.2020.105926](https://doi.org/10.1016/j.envint.2020.105926)

According to the Contributor Roles Taxonomy, the candidate's contribution was as follows: conceptualisation; methodology; investigation; writing (original draft); writing (review and editing); visualisation; project administration; funding acquisition.

Candidate: _____ Date: _____
Mr. Paul A. Whaley

Supervisor: _____ Date: _____
Prof. Crispin J. Halsall



Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint



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



Paul Whaley^{a,*}, Elisa Aiassa^b, Claire Beausoleil^c, Anna Beronius^d, Gary Bilotta^e, Alan Boobis^f, Rob de Vries^g, Annika Hanberg^h, Sebastian Hoffmannⁱ, Neil Hunt^j, Carol F. Kwiatkowski^k, Juleen Lam^l, Steven Lipworth^m, Olwenn Martinⁿ, Nicola Randall^o, Lorenz Rhomberg^p, Andrew A. Rooney^q, Holger J. Schünemann^r, Daniele Wikoff^s, Taylor Wolffe^t, Crispin Halsall^u

^a Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

^b European Food Safety Authority (EFSA), Assessment and Methodological Support Unit, Via Carlo Magno 1/A, 43126 Parma, Italy

^c ANSES (French Agency for Food, Environmental and Occupational Health Safety), Risk Assessment Department, Chemical Substances Assessment Unit, F-94700 Maisons-Alfort, France

^d Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

^e School of Environment and Technology, University of Brighton, Brighton, UK

^f National Heart & Lung Institute, Imperial College London, London, UK

^g SYRCLÉ, Department for Health Evidence, Radboud Institute for Health Sciences, Radboudumc, Nijmegen, the Netherlands

^h Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

ⁱ Evidence based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health, Paderborn, Germany

^j Yordas Group, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

^k The Endocrine Disruption Exchange, P.O. Box 54, Eckert, CO 81418, USA

^l University of California, San Francisco and California State University, East Bay, 28500 Carlos Bee Blvd Room 502, Hayward, CA 94542, USA

^m Royal Society of Chemistry, Burlington House, Piccadilly, London W1J 0BA, UK

ⁿ Institute for the Environment, Health and Societies, Brunel University London, Uxbridge, UK

^o Harper Adams University, Newport, Shropshire, UK

^p One Beacon Street, 17th Floor, Boston, MA 02108, USA

^q Division of the National Toxicology Program, National Institute of Environmental Health Sciences, NC, USA

^r McGRADE Centre and Michael G De Groot Cochrane Canada Centre, Dept. of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

^s ToxStrategies, 31 College Place, Suite B118B, Asheville, NC 28801, USA

^t Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

ARTICLE INFO

Handling Editor: Adrian Covaci

Keywords:

Systematic review
Research standards
Research synthesis methods
Health assessment
Meta-analysis
Environmental health
Toxicology
Epidemiology

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

* Corresponding author.

E-mail addresses: p.whaley@lancaster.ac.uk (P. Whaley), elisa.aiassa@efsa.europa.eu (E. Aiassa), claire.beausoleil@anses.fr (C. Beausoleil), anna.beronius@ki.se (A. Beronius), a.boobis@imperial.ac.uk (A. Boobis), rob.devries@radboudumc.nl (R. de Vries), annika.hanberg@ki.se (A. Hanberg), sebastian.hoffmann@seh-cs.com (S. Hoffmann), n.hunt@yordasgroup.com (N. Hunt), juleen.lam@csueb.edu (J. Lam), olwenn.martin@brunel.ac.uk (O. Martin), nrandall@harper-adams.ac.uk (N. Randall), lrhomberg@gradientcorp.com (L. Rhomberg), andrew.rooney@nih.gov (A.A. Rooney), schuneh@mcmaster.ca (H.J. Schünemann), dwikoff@toxstrategies.com (D. Wikoff), t.wolffe@lancaster.ac.uk (T. Wolffe), c.halsall@lancaster.ac.uk (C. Halsall).

<https://doi.org/10.1016/j.envint.2020.105926>

Received 2 December 2019; Received in revised form 26 May 2020; Accepted 21 June 2020

Available online 09 July 2020

0160-4120/© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license

(<http://creativecommons.org/licenses/by/4.0/>).

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 (see Fig. 1) 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).

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, 2019a, 2019b); 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

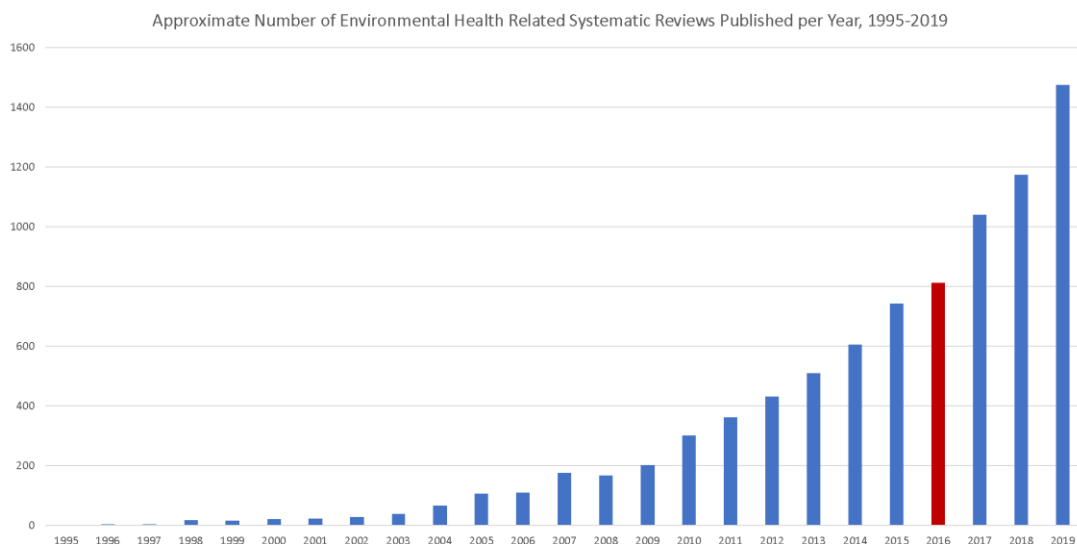


Fig. 1. Chart showing annual increase in number of publications on topics related to EH research with the term “systematic review” in the title, indexed in Web of Science. The total number of publications approximately doubled between 2016 and 2020. **Search:** TITLE: (“systematic review”), **Refined 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.

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 be reformulated? (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](https://zenodo.org) (Whaley et al., 2020).

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 Fig. 2). The recommendations within each domain are listed in Table 1. An explanation of key recommendations is provided in Table 2. 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



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

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.

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.

4.3. Strengths and limitations of COSTER

4.3.1. The consensus process

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

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

In order to improve this consensus process, and to elevate COSTER

Applying Systematic Review Methods in Chemical Risk Assessment - Chapter 2. Recommended Practices

P. Whaley, et al.

Environment International 143 (2020) 105926

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.

| 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 | <p>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.</p> <p>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.</p> <p>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.</p> <p>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.</p> |
| 1.2 Setting the research question to inform the scope of the review ("problem formulation") | <p>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.</p> <p>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).</p> <p>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:</p> <ul style="list-style-type: none"> ● 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 | <p>1.3.1 Define and justify unambiguous and appropriate eligibility criteria for each component of the objective statement.</p> <p>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)</p> <p>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.</p> <p>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.</p> <p>1.3.5 For study designs: define eligible study designs per design features rather than design labels.</p> <p>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.</p> <p>1.3.7 Include evidence which is relevant to review objectives irrespective of whether its results are in a usable form.</p> <p>1.3.8 Include relevant evidence irrespective of language.</p> <p>1.3.9 Exclude evidence which is not publicly available.</p> |
| 1.4 Planning the review methods at protocol stage | <p>1.4.1 Design sufficiently sensitive search criteria, so that studies which meet the eligibility criteria of the review are not inadvertently excluded.</p> <p>1.4.2 Design "characteristics of included studies" table.</p> <p>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.</p> <p>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.</p> <p>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.)</p> <p>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.</p> <p>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.</p> |
| 1.5 Publishing the protocol | <p>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.</p> <p>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.</p> <p>1.5.3 Publish the final version of the protocol in a public archive, prior to screening studies for inclusion in the review.</p> <p>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.</p> |

(continued on next page)

Applying Systematic Review Methods in Chemical Risk Assessment - Chapter 2. Recommended Practices

P. Whaley, et al.

Environment International 143 (2020) 105926

Table 1 (continued)

| COSTER v1.0.0: Recommendations for the planning and conduct of environmental health systematic reviews | |
|--|---|
| 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. |

(continued on next page)

Table 1 (continued)

| COSTER v1.0.0: Recommendations for the planning and conduct of environmental health systematic reviews | |
|--|---|
| 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. |

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

4.3.2. Author conflicts of interest

In order to secure cross-sector consensus, we purposely invited participants with varied interests in relation to 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

Applying Systematic Review Methods in Chemical Risk Assessment - Chapter 2. Recommended Practices

P. Whaley, et al.

Environment International 143 (2020) 105926

Table 2
Explanation and elucidation of key recommendations of COSTER.

| Explanation and elucidation of key recommendations of COSTER v1.0.0 | |
|--|--|
| Project planning: recommendations 1.1.1 through 1.5.4 | |
| Contribution of COSTER: <i>Emphasis on importance of standard practices in biomedical SRs for environmental health research</i> | <p>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).</p> <p>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.</p> |
| Disclosure and management of interests: recommendations 1.1.3, 1.1.4 | |
| Contribution of COSTER: <i>Distinction between potential and apparent conflicts of interest relating to team selection in SRs</i> | <p>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).</p> <p>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.</p> <p>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.</p> <p>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).</p> |
| Interpreting external validity of the evidence, and integrating multiple evidence streams: recommendations 1.2.2, 1.4.6, 7.8 | |
| Contribution of COSTER: <i>Adaptation of biomedical SR standards to specific context of EH research</i> | <p>Operationalising the interpretation of indirect, non-human and <i>in vitro</i> 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.</p> <p>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.</p> <p>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.</p> |
| Formulation of research objectives: recommendations 1.2.3, 1.3.3, 1.3.4, 1.3.5, 1.3.9 | |
| Contribution of COSTER: <i>Formal clarification of use of PECO-style statements in formulating SR objectives in EH research</i> | <p>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.</p> <p>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.</p> |

(continued on next page)

Applying Systematic Review Methods in Chemical Risk Assessment - Chapter 2. Recommended Practices

P. Whaley, et al.

Environment International 143 (2020) 105926

Table 2 (continued)

| Explanation and elucidation of key recommendations of COSTER v1.0.0 | |
|--|--|
| Including informally published or previously unpublished literature, regardless of usability in the planned analysis: recommendations 1.3.6 to 1.3.9, 3.4 | |
| Contribution of COSTER: Provides unambiguous rationale for exclusion of study reports due to insufficient information content | <p>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.</p> <p>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.</p> <p>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).</p> <p>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).</p> <p>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.</p> <p>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.</p> <p>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.</p> |
| Protocol publication: recommendations 1.5.1, 1.5.2, 1.5.3 | |
| Contribution of COSTER: Contribution of COSTER: Differentiates between protocol registration and publication as distinct steps of the methods development process | <p>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.</p> <p>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).</p> |
| Assessing the internal validity of included studies: recommendations 1.4.3, 5 | |
| Contribution of COSTER: Explicit specification of risk of bias methods for assessing internal validity of included studies | <p>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.</p> <p>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).</p> <p>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.</p> |

(continued on next page)

Table 2 (continued)

| Explanation and elucidation of key recommendations of COSTER v1.0.0 | |
|---|---|
| Assessment of confidence or certainty in the overall body of evidence: recommendations 1.4.5, 7.1, 7.2, 7.4, 7.5, 7.6, 7.7, 7.8, 7.10 | |
| Contribution of COSTER: Contribution of COSTER: <i>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</i> | <p>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.</p> <p>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.</p> |
| Making policy recommendations: recommendation 8.3 | |
| Contribution of COSTER: <i>Emphasises that recommendations about interventions are often beyond the scope of a SR of health effects from environmental exposures</i> | <p>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.</p> <p>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.</p> <p>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.</p> |

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. 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; nor is strong evidence for what is effective practice necessarily going to be 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, an example of which being the 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.

Declaration of Competing Interest

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.

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 <http://doi.org/10.5281/zenodo.3903115>.

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.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.envint.2020.105926>.

References

- 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. *Syst. Rev.* 5 (1), 164. <https://doi.org/10.1186/s13643-016-0337-y>.
- Arzuaga, Xabier, Smith, Martyn T., Gibbons, Catherine F., Skakkebaek, 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. *Environ. Health Perspect.* 127(6), 65001. <https://doi.org/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. *BMJ open* 7 (2), e012545. <https://doi.org/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/prospéro/> (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/prospéro/#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? *Nature* 573 (7773), 187–189. <https://doi.org/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://ccnml.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. *Environ. Int.* 119, 366–378. <https://doi.org/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. National Academies Press,

Applying Systematic Review Methods in Chemical Risk Assessment - Chapter 2. Recommended Practices

P. Whaley, et al.

Environment International 143 (2020) 105926

- Washington, D.C.
- EFSA, 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA J.* 8 (6), 1637. <https://doi.org/10.2903/j.efsa.2010.1637>.
- EFSA, 2015. Principles and process for dealing with data and evidence in scientific assessments. *EFSA J.* 13 (6), 4121. <https://doi.org/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. *Syst. Rev.* 8 (1), 280. <https://doi.org/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*. *J. Clin. Epidemiol.* 64 (4), 380–382. <https://doi.org/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. *BMJ (Clinical research ed.)* 336 (7650), 924–926. <https://doi.org/10.1136/bmj.39489.470347.AD>.
- Haddaway, Neal R., Westgate, Martin J., 2019. Predicting the time needed for environmental systematic reviews and systematic maps. *Conserv. Biol. J. Soc. Conserv. Biol.* 33 (2), 434–443. <https://doi.org/10.1111/cobi.13231>.
- Hansen, Martin Rune, Hassan, Jørs, Erik, Sandbæk, Anneli, Kolstad, Albert, Henrik, Schullehner, Jörg, Schiltnissen, Vivi, 2019. Exposure to neuroactive non-organochlorine insecticides, and diabetes mellitus and related metabolic disturbances: Protocol for a systematic review and meta-analysis. *Environ. Int.* 127, 664–670. <https://doi.org/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. *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, 2019. *IARC Monographs on the Identification of Carcinogenic Hazards to Humans: Preamble*. IARC, Lyon, France.
- IARC, 2019. Instructions for Authors for the Preparation of Drafts for IARC Monographs, updated on 6/17/2019.
- International Committee of Medical Journal Editor, 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, Mar, Del, Chris, B., Hama, Rokuro, Thompson, Matthew J., et al., 2014. Neuraminidase inhibitors for preventing and treating influenza in healthy adults and children. *Cochr. database of Syst. Rev.* 4, CD008965. <https://doi.org/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. *Environ. Int.* 92–93, 716–728. <https://doi.org/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. *Environ. Health Perspect.* 121 (9), 985–992. <https://doi.org/10.1289/ehp.1206389>.
- Lawlor, Debbie A., Tilling, Kate, Davey Smith, George, 2016. Triangulation in aetiological epidemiology. *In Int. J. Epidemiol.* 45(6), 1866–1886. <https://doi.org/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. *Environ. Int.* 119, 558–569. <https://doi.org/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. *Environ. Health Perspect.* 127 (7), 75001. <https://doi.org/10.1289/ehp.12064971>.
- Mandrioli, Daniele, Schiltnissen, 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. *Environ. Int.* 119, 174–185. <https://doi.org/10.1016/j.envint.2018.06.005>.
- Martin, O.V., Bopp, S., Ermler, S., Kienzler, A., McPhie, J., Pains, 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 Bizet, Bruno, Antignac, Jean-Philippe, Cano-Sancho, German, 2019. Associations between exposure to organochlorine chemicals and endometriosis in experimental studies: A systematic review protocol. *Environ. Int.* 124, 400–407. <https://doi.org/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*. John Wiley & Sons Ltd., Oxford, UK.
- 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*. *Environ. Int.* 128, 11–12. <https://doi.org/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. *Environ. Int.* 92–93, 611–616. <https://doi.org/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. *Environ. Int.* 121 (Pt 1), 1027–1031. <https://doi.org/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. *Nat. Hum. Behav.* 1(1), e124. <https://doi.org/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/ohat/review/index-2.html>, checked on 5/8/2019.
- Paez, Arsenio, 2017. Grey literature: An important resource in systematic reviews. *J. Evid.-Based Med.* <https://doi.org/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. *Syst. Rev.* 6 (1), 263. <https://doi.org/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. *Chemosphere* 177, 167–175. <https://doi.org/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, Sempile, 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. *Environ. Int.* 103, 73–90. <https://doi.org/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. *Environ. Int.* 121 (Pt 1), 764–793. <https://doi.org/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. *Environ. Int.* 132, 104768. <https://doi.org/10.1016/j.envint.2019.04.040>.
- Rooney, Andrew A., Boyles, Abbe L., Wolfe, Mary S., Bucher, John R., Thayer, Kristina A., 2014. Systematic review and evidence integration for literature-based environmental health science assessments. *Environ. Health Perspect.* 122 (7), 711–718. <https://doi.org/10.1289/ehp.1307972>.
- Rooney, Andrew A., Cooper, Glinda S., Jahnke, Gloria D., Lam, Juleen, Morgan, Rebecca L., Boyles, Abbe L., et al., 2016. How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. *Environ. Int.* 92–93, 617–629. <https://doi.org/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. *Regul. Toxicol. Pharmacol.* 91, 124–141. <https://doi.org/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. *Am. J. Public Health* 109 (7), 982–984. <https://doi.org/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. *Environ. Health Perspect.* 124 (6), 713–721. <https://doi.org/10.1289/ehp.1509912>.
- Stephens, Martin L., Betts, Kellyn, Beck, Nancy B., Cogliano, Vincent, Dickerson, Kay, Fitzpatrick, Suzanne, et al., 2016. The Emergence of Systematic Review in Toxicology. *Toxicol. Sci.* 152 (1), 10–16. <https://doi.org/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. *Environ. Health*

Applying Systematic Review Methods in Chemical Risk Assessment - Chapter 2. Recommended Practices

P. Whaley, et al.

Environment International 143 (2020) 105926

- Glob. Access Sci. Source 15 (1), 74. <https://doi.org/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. *Toxicol. Sci. Off. J. Soc. Toxicol.* 142 (2), 312–320. <https://doi.org/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. *Toxicol. Sci. Off. J. Soc. Toxicol.* 142 (2), 321–330. <https://doi.org/10.1093/toxsci/kfu200>.
- Vries, R. B. M. de, Hooijmans, Carljin 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. *Evid.-based Preclin. Med.* 2(1), e00007. <https://doi.org/10.1002/ebm2.7>.
- Wang, Zhicheng, Taylor, Kyla, Allman-Farinelli, Margaret, Armstrong, Bruce, Askie, Lisa, Gherzi, 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., 2020. A code of practice for the conduct of systematic reviews in toxicology and environmental health research (COSTER). <http://doi.org/10.5281/zenodo.3539002>.
- Whaley, Paul, Halsall, Crispin, Agerstrand, Marlene, Aiassa, Elisa, Benford, Diane, Bilotta, Gary, et al., 2016. Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations. *Environ. Int.* 92–93, 556–564. <https://doi.org/10.1016/j.envint.2015.11.002>.
- Woodruff, Tracey J., Sutton, Patrice, 2014. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. *Environ. Health Perspect.* 122 (10), 1007–1014. <https://doi.org/10.1289/ehp.1307175>.

Chapter 3.

Biological plausibility

This chapter is being prepared for publication as an official Concept Paper of the GRADE Working Group. It has been peer-reviewed by the GRADE Environmental Health Project Group and was presented for feedback at the GRADE Annual Meeting on 17 June 2020. The manuscript has been revised according to the comments received and is being presented for approval at the GRADE Meeting on 5 October 2020.

According to the Contributor Roles Taxonomy, the candidate's contribution was as follows: conceptualisation; methodology; investigation; writing (original draft); writing (review and editing); visualisation.

Candidate: _____ Date: _____
Mr. Paul A. Whaley

Supervisor: _____ Date: _____
Prof. Crispin J. Halsall

The definition and role of “biological plausibility” in environmental health systematic reviews: a GRADE concept paper

Paul Whaley (1,2), Thomas Piggott (3), Rebecca L. Morgan (3), Daniele Wikoff (4), Sebastian Hoffmann (2), Katya Tsaïoun (2), Kristina A. Thayer (5), Holger Schünemann* (3,6)

Affiliations

1. Lancaster Environment Centre, Lancaster University, UK
2. Evidence-based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health (EBTC)
3. Department of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main St West; Hamilton, ON L8N 3Z5, Canada
4. ToxStrategies, Inc., Asheville, NC, USA
5. U.S. Environmental Protection Agency (US EPA), Office of Research and Development, Center for Public Health and Environmental Assessment (CPHEA), Chemical Pollutant Assessment Division (CPAD), 1200 Pennsylvania Avenue, NW (8623R), Washington, DC 20460, USA
6. Michael G DeGroot Cochrane Canada and McMaster GRADE Centres; McMaster University, HSC-2C, 1280 Main St West; Hamilton, ON L8N 3Z5, Canada

Abstract

Background: “Biological plausibility” is a concept frequently referred to in environmental health when researchers are evaluating how confident they are in the results and inferences of a study or evidence review. Biological plausibility is not, however, a domain of one of the most widely-used approaches for assessing the certainty of evidence (CoE) which underpins the findings of a systematic review, the GRADE CoE Framework. Whether the omission of biological plausibility is a potential limitation of the GRADE CoE Framework is a topic that is regularly discussed, especially in the context of environmental health systematic reviews.

Objectives: Here we seek to determine whether “biological plausibility” is a concept already accommodated under the existing GRADE domains. Although evidence to support biological plausibility can come from a variety of sources and be considered in multiple contexts, here we focus on experimental animal and *in vitro* data because these are the primary types of evidence used to assess biological plausibility in human health environmental exposure assessments.

Discussion: We argue that “biological plausibility” is a concept which primarily comes into play when evidence about the effects of an exposure on a population of concern (usually humans) is lacking, i.e. it either does not exist, or it is at high risk of bias, is inconsistent, or limited in other ways. In such circumstances, researchers look toward evidence from other study designs in order to draw conclusions. We can consider experimental animal and *in vitro* evidence as “surrogates” for the target populations, outcomes and exposures of actual interest. Through discussion of 12 examples of experimental surrogates, we propose an updated working definition of “biological plausibility” as a concept with two principle aspects, a “generalisability aspect” and a “mechanistic aspect”. The “generalisability aspect” concerns the validity of inferences from experimental models to real-world scenarios, and asks the same question as assessment of external validity in systematic reviews. The “mechanistic aspect” concerns certainty in knowledge of biological mechanisms and informs judgements of the generalisability of surrogates. While both aspects are accommodated by the indirectness domain of the GRADE CoE Framework, further research is needed to determine how to use knowledge of biological mechanisms to operationalise assessment of external validity and indirectness of the evidence in systematic reviews. This research should be closely informed by experience of assessing biological plausibility in environmental health.

Introduction

In environmental and public health research, toxicology, and human health chemical risk assessment (henceforth referred to as “environmental health research”) it is rare to have definitive evidence from human populations about the potential health harms that a given environmental exposure might be causing. When it does exist, human evidence typically consists of epidemiological studies. These are not experimental in design and therefore need to be considered carefully when drawing conclusions of causality, for example with respect to considerations of exposure assessment methods and potential confounding (Braun and Gray, 2017).

The rarity of definitive evidence from studies in humans elevates the importance in environmental health research of evidence from experimental animal (*in vivo*) and *in vitro* studies. However, while evidence from *in vivo* and *in vitro* studies has the advantage that exposure can be controlled, the laboratory set-up is only indirectly representative of the real-world situation which it models - with differences between species, the use of artificial cell culture constructs to measure biological processes, and exposure regimens which are often much higher, shorter and more regimented than real-world cases (Rhomborg, 2015).

There is always, therefore, a need to translate the significance of findings from laboratory experiments to the real-world scenarios they are informing. Our ability to do this correctly is critical in successfully identifying, quantifying, and limiting health harms from environmental exposures. As systematic reviews become mainstream in environmental health (Bilotta, Milner and Boyd, 2014; Sheehan and Lam, 2015; Morgan *et al.*, 2016; Whaley *et al.*, 2016; Hoffmann *et al.*, 2017), the need for systematic approaches to translating evidence from the laboratory to the real-world context becomes increasingly important.

One concept which has long been applied in translating the findings of laboratory experiments to real-world contexts is that of “biological plausibility”. This was first introduced by Sir Austin Bradford Hill in 1965 as one of his considerations for establishing causality (Hill, 1965). Bradford Hill argued that the presence of biological plausibility can be considered to increase the likelihood that a relationship between an environmental exposure and a health outcome is a causal one. However, in spite of many definitions of “biological plausibility” having been offered (see Table 1 for some examples), exactly what constitutes biological plausibility has never been fully or finally characterised. This is particularly true for the context of conducting environmental health systematic reviews. Methodologists, including the GRADE Working Group, have regularly been challenged by environmental health practitioners about whether and how the assessment of biological plausibility is accommodated by the systematic review process (European Food Safety Authority, 2018).

| Source | Definition of "biological plausibility" |
|---|--|
| Wikipedia (Wikipedia contributors, 2014) | "A relationship between a putative cause and an outcome — that is consistent with existing biological and medical knowledge" and "one component of a method of reasoning that can establish a cause-and-effect relationship between a biological factor and a particular disease or adverse event" |
| European Food Safety Authority (Hardy <i>et al.</i> , 2017) | "Consistency between data and biological theory or mechanism" |
| Last's Dictionary of Epidemiology (International Epidemiological Association, 2001) | The "causal consideration that an observed, potentially causal association between an exposure and a health outcome may plausibly be attributed to causation on the basis of existing biomedical and epidemiological knowledge." |
| Organisation for Economic Co-operation and Development (OECD, 2016) | Being "consistent with biological knowledge" and "based on extensive previous documentation and broad acceptance" |
| US Environmental Protection Agency Cancer Guidelines (US Environmental Protection Agency, 2005) | "An inference of causality [which] tends to be strengthened by consistency with data from experimental studies or other sources demonstrating plausible biological mechanisms. A lack of mechanistic data, however, is not a reason to reject causality." |

Table 1. Examples of definitions of "biological plausibility"

GRADE and biological plausibility

The GRADE Framework, commonly used in public health and healthcare systematic reviews and increasingly in environmental health (Morgan *et al.*, 2016), contends that assessment of the certainty of evidence for answers to research questions can be successfully operationalised (i.e. conducted accurately, consistently and transparently by different researchers working in different times and places) via systematic consideration of a predefined set of eight strengths and limitations of the overall evidence base (Guyatt *et al.*, 2008). The limitation domains which reduce certainty in the results of a systematic review are risk of bias, inconsistency, indirectness, imprecision and publication bias; the strength domains which increase certainty are large effect size, presence of a dose-response relationship, and residual opposing confounding (see Figure 1).

The strength and limitation domains of the GRADE Framework are intended to be exhaustive of the concepts necessary for assessing certainty of the evidence, operationalised via a structured reasoning process designed to produce more consistent, transparent results than is achievable by direct application of the considerations of Bradford Hill. Historically, the contention has been that the role played by assessment of biological plausibility in environmental health assessments is already accommodated either in the GRADE domains or as part of the systematic review process (Schunemann *et al.*, 2011; Hultcrantz *et al.*, 2017).

| -1- Establish initial level of certainty | | -2- Consider lowering or raising level of certainty | | -3- Final rating for level of certainty |
|--|---|---|--|--|
| Study Design | Initial level of certainty in an estimate of effect | Reasons for considering lowering or raising certainty | | Certainty in an estimate of effect across those considerations |
| | | Lower if | Higher if** | |
| Randomised trials, <i>in vivo, in vitro</i> | High | Risk of bias Unexplained inconsistency Indirectness Imprecision Publication bias | Large effect | High ⊕⊕⊕⊕ |
| | Moderate | | Dose response | Moderate ⊕⊕⊕⊖ |
| Observational studies* | Low | | All plausible confounding and bias would reduce a demonstrated effect or suggest a spurious effect if no effect was observed | Low ⊕⊕⊖⊖ |
| | Very Low | | | Very low ⊕⊖⊖⊖ |

*Observational studies may start at high certainty depending on tool used for assessing risk of bias
**Upgrading criteria are usually applicable to observational studies only

Figure 1. The upgrade and downgrade domains in GRADE and how they are used to determine the overall certainty in evidence for a systematic review. Adapted from (Morgan *et al.*, 2016).

The purpose of this paper is to re-examine that contention and elucidate the role performed by the assessment of biological plausibility in environmental health systematic reviews.

We argue that biological plausibility is a concept which primarily comes into play when evidence directly concerning the populations, exposures and outcomes of concern is lacking. In these scenarios, systematic reviews may include evidence from surrogate *in vivo* and *in vitro* experimental models. We present 12 examples of the use of such surrogates in environmental health reviews to gain insight into how researchers employ the concept of biological plausibility when selecting and weighing evidence from experimental models.

We believe these 12 examples show that the concept of biological plausibility in fact consists of two connected principle aspects, which we call the “generalisability aspect” and the “mechanistic aspect”. The generalisability aspect of biological plausibility concerns the generalisability of findings in an experimental context to a target context of concern. The mechanistic aspect of biological plausibility concerns certainty in knowledge of biological mechanisms and is informative of judgements of the generalisability of a surrogate.

Since the generalisability aspect is concerned with the same issue of generalisability from experimental to target context as is the assessment of external validity in systematic reviews, and the mechanistic aspect informs judgements under the generalisability aspect, it follows that biological plausibility is accommodated under the GRADE domain of indirectness. GRADE does not therefore need to be extended to include a specific domain of biological plausibility. However, instruments for assessing the external validity of surrogates in systematic reviews have not yet been developed. We therefore recommend that the development of such tools takes into account insights from assessment of biological plausibility in environmental health research. Because GRADE has not yet been applied to assessment of certainty in knowledge of biological mechanisms, and this is an integral part

of judging the external validity of surrogates, we also recommend research be conducted into this as a new application of the GRADE CoE Framework.

“Biological plausibility” and the inclusion of surrogates in systematic reviews

Systematic review is the application of methods designed to minimise risk of systematic and random error and maximise transparency when using existing evidence to answer research questions. Systematic review questions in environmental health are generally characterised in terms of the population, exposure, comparator and outcome of concern - the PECO mnemonic (Morgan *et al.*, 2018).

The purpose of characterising environmental health questions and the objectives of systematic reviews in terms of a PECO statement is to facilitate unambiguous characterisation of the types of studies which will be considered by the authors of a systematic review to be relevant or eligible for answering their question. Studies which are more directly relevant will have PECO characteristics which closely fit with those of the systematic review; those which are less relevant will be less similar. This concept of fit between a study and the objectives of a systematic review is “external validity” - the extent to which the findings of a study can be generalised to populations, exposures and outcomes outside the context of that study (Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, 2019).

When designing a systematic review, the authors need to decide what their cut-off or threshold for external validity is going to be - where they draw the line on a study being insufficiently generalisable to their target PECO to be worth including in their review. Where the line is drawn will depend on the review objectives. However, the most efficient way of going about a systematic review is to define as eligible only those studies whose designs most directly match the PECO characterisation of the systematic review question. If the evidence from those studies is sufficiently certain then there is no need to seek out other evidence in support of the findings of the systematic review - the investigation can stop (see Figure 2A).

A classic example of this scenario is smoking causing lung cancer. Several observational studies have investigated doctors (P) and how much they smoke (E) in comparison to doctors who do not smoke (C), and assessed the relative risk of lung cancer (O) between the two groups. The studies are at relatively low risk of bias, including confounding; multiple studies of similar design give reasonably consistent results; they are in a representative population; the overall effect size is reasonably precise; there is no evidence that publication bias exaggerates the observed effect size; there is a dose-response relationship; and the effect size is large, with smoking increasing lung cancer risk by a factor of 12-24 (Doll *et al.*, 2005; Pope *et al.*, 2011).

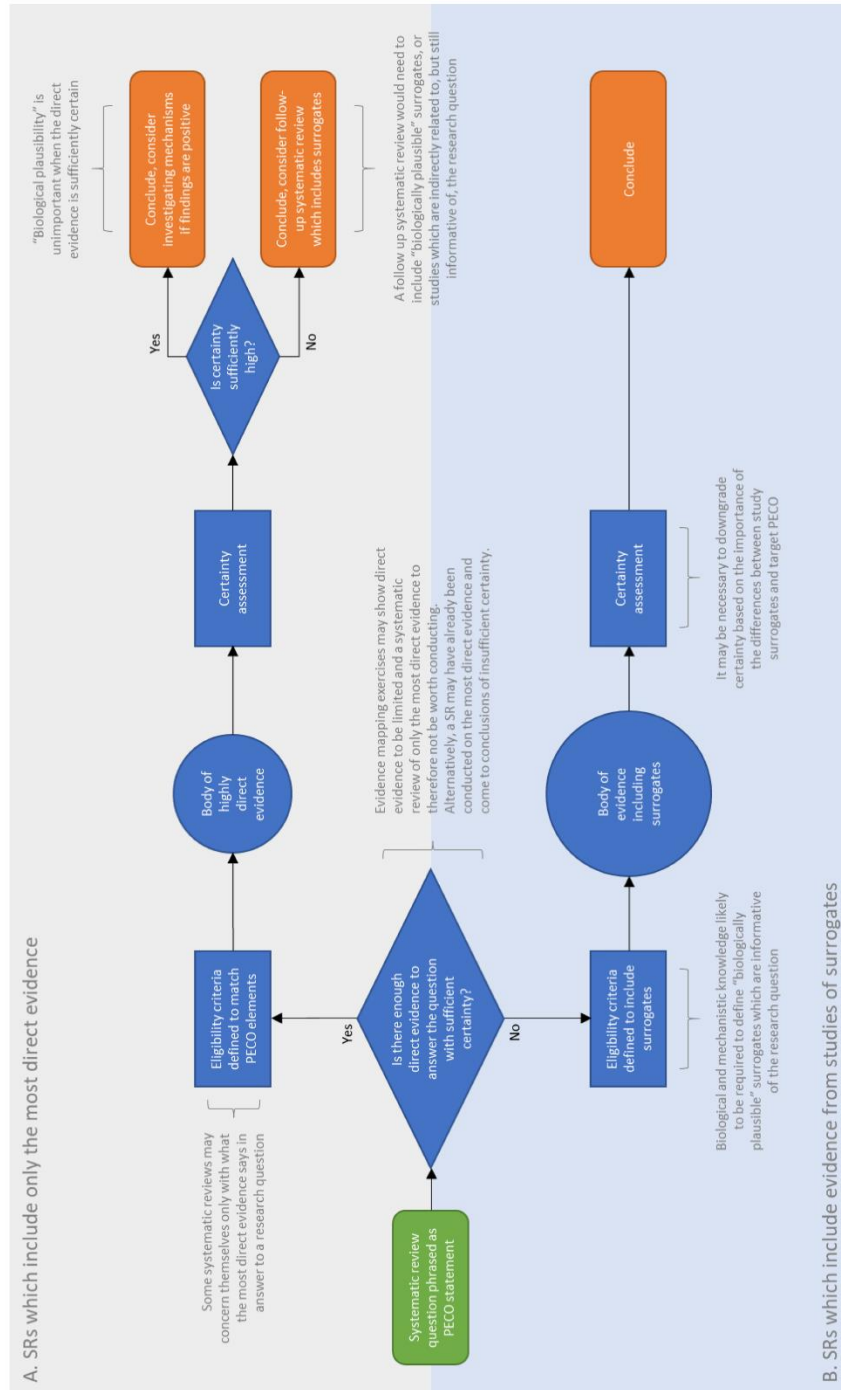


Figure 2. Schematic representation of when studies of surrogates might be included in a systematic review.

These features of the evidence can all be established, and sufficiently high certainty that a causal relationship has therefore been observed, without knowledge of the mechanism by which the exposure is having the effect. In such scenarios, the “biological plausibility” is inferred - it can be assumed there is a discoverable biological mechanism because there is high certainty that the relationship is causal, even when there is little information about the biological mechanism by which the exposure causes its outcome. Conversely, that it is not known why or how the exposure causes the outcome does not undermine certainty in the evidence for the relationship.

The challenge in environmental health research is that high certainty in direct evidence is a theoretical possibility which is only rarely realised. Usually, environmental health systematic reviews will find the human evidence which directly relates to a hypothesised exposure-outcome relationship is uncertain or even non-existent. In such circumstances, in order to further investigate and elucidate potential causal relationships between exposures and outcomes, it becomes necessary to evaluate studies of surrogates (see Figure 2B). This is done in the expectation that including in the systematic review indirect evidence from studies of surrogates will raise certainty about whether or not there is a causal relationship.

The use of surrogates is familiar in environmental health contexts, which has long been reliant on evidence where animal models stand in for target human populations, biomarkers of disease are used in place of observation of clinical health outcomes, and potential health effects of under-studied chemicals are inferred from their similarity to better-researched substances. As with any systematic process, decisions on which surrogates to include in a systematic review need to be transparent and well-reasoned, based on evidence of the validity of the decision, and defined in advance of conduct of the review. Spurious inclusion of surrogate studies is not just a waste of time and resources: if surrogates are not informative of the question but nonetheless included in the overall analysis, then the validity of the results of the systematic review may be compromised; likewise, spurious exclusion of surrogate studies which should have been included also risks false conclusions.

The “biological plausibility” of choice of surrogates

In environmental health assessments, the use of a surrogate is considered to be justifiable insofar as it provides “biologically plausible” support for the hypothesised exposure-outcome relationship in the population of concern (European Food Safety Authority, 2018). In the context of systematic reviews, the GRADE approach is to assess the importance of the indirectness of the surrogate relative to the question being asked. Evidence from surrogates which are critically indirect would be excluded from a systematic review; evidence from surrogates which is direct enough to be informative would be included but might be downgraded for certainty depending on how important the indirectness of the surrogate is determined to be (Guyatt *et al.*, 2011).

We now present a series of 12 examples of the use of surrogates in environmental health assessments which we frame in terms of the concept of biological plausibility. We then use

these examples to indicate how judgements of biological plausibility relate to the concepts of systematic review. The examples are summarised in Table 2.

Surrogate populations

Toxicology has a long history of use of animal models for investigating potential harm to health from exposure to chemical substances, due to the ethical prohibition on testing for harm in humans but still requiring evidence to inform evaluation of chemical health risks.

One example of the acceptance of surrogate animal and *in vitro* populations for predicting health outcomes in target human populations is in the assessment of the carcinogenicity of 2-nitropropane. While there is no direct evidence of carcinogenicity in humans, animal and *in vitro* evidence is considered to be sufficiently certain to justify classifying 2-nitropropane as a human carcinogen (Papameletiou *et al.*, 2017). Here it is being judged sufficiently "biologically plausible" that what is observed in animal surrogates would also be seen in humans that a conclusion of carcinogenicity can be drawn. In the language of GRADE, this would be to say that the non-human surrogate evidence is direct enough to be eligible for determining human carcinogenicity.

In a contrasting example, the US FDA has deemed rat models for assessing bladder carcinogenicity to be ineligible for assessment of saccharin as a food carcinogen. This is due to the mechanism by which saccharin causes tumour growth in rats not being present in humans (US National Resource Council, 2014). The model was originally considered to be predictive, but once the mechanism by which saccharin causes cancer in rats was determined not to be present in humans, the US FDA excluded the rat model from its assessment. That is to say, there is an absence of biological plausibility to the claim that the mechanism by which saccharin causes cancer in rats is also occurring in humans. In the language of GRADE, this would be to say the indirectness of the animal model is critical - the rat model is too indirect, does not generalise to humans, and therefore would not be eligible for inclusion in a systematic review of whether saccharin is a bladder carcinogen.

Surrogate outcomes

Surrogate outcomes are used in environmental health research because, when conducting experimental or observational studies, it is often easier or more ethical to measure biomarkers of disease than clinical outcomes of interest. This is the case when clinical outcomes may have long latency periods in the population of concern (such as for many cancer types), when observation of an outcome would entail allowing a disease to develop instead of intervening to treat it, or when the outcome may not even actually manifest in the observed population (such as when *in vitro* or certain *in vivo* models are being used).

An example of the acceptance of a surrogate outcome is in a systematic review of evidence for the developmental and reproductive toxicity of the biocide triclosan by Johnson *et al.* (2016). In this case, serum thyroxine concentrations in pregnant women were chosen as a surrogate for the neurodevelopmental health of children. The reasoning was that maternal

thyroid levels during pregnancy are sufficiently predictive of the subsequent neurodevelopmental health of the child that the indirectness of the outcome can be taken to be unimportant. These authors can be interpreted as considering there to be a sufficiently “biologically plausible” relationship between maternal serum thyroxine and neurodevelopment that it can be used as a surrogate outcome. In the language of GRADE, the indirectness of the surrogate outcome is unimportant and studies which investigate this outcome are eligible for inclusion in the systematic review..

In contrast, a systematic review of biomarkers for Alzheimer’s Disease found insufficient evidence to be able to recommend any biomarker for use as a surrogate outcome for disease progression (McGhee *et al.*, 2014). While it might seem to be “biologically plausible” that Alzheimer’s Disease results in specific changes to physical brain structure detectable in an MRI scan, there seems to be a lack of empirical evidence that this is reliably the case. The importance of the indirectness of the surrogate measures is therefore higher than might have previously been considered. This would suggest that, in GRADE, studies of anything other than clinical symptoms of Alzheimer’s Disease should be at least be downgraded for indirectness due to the uncertainty of how the biomarkers predict the final outcome of concern, if not outright excluded from systematic reviews.

Surrogate exposures

Selecting and according appropriate importance to surrogate exposures is a complex issue in environmental health systematic reviews. We briefly discuss four aspects of surrogate exposure: route of exposure; administered dose; active substance; and matrix of exposure assessment. These should provide sufficient illustration of principle, though we note that other aspects of exposure such as measurement of metabolites vs. parent compound, timing of exposure, and other issues, will need consideration in environmental health systematic reviews (Cohen Hubal *et al.*, 2020).

Route

Extrapolating from experimental routes of exposure to the exposure routes likely to be encountered by target populations is a major preoccupation of toxicological risk assessment. For example, toxicology studies which administer bisphenol-A to animal test subjects via oral gavage are considered to be of direct relevance to assessing outcomes from dietary exposure. In contrast, IV administration of bisphenol-A is typically considered to be very indirect, due to the avoidance of first-pass glucuronidation in the liver (European Food Safety Authority, 2015). However, the perceived importance of IV administration may increase if knowledge of how bisphenol-A is metabolised allows equivalent oral doses to be calculated from IV doses, as this provides what can be interpreted as a “biologically plausible” account of how the two doses are related (Taylor, Welshons and Vom Saal, 2008). In such circumstances, GRADE would consider the indirectness of the route of exposure as being of less importance. Pharmacokinetic models to aid in route-to-route extrapolation are encouraged in chemical assessments (Meek *et al.*, 2013).

Dose

In toxicological research, experiments often are conducted using high doses that are not considered environmentally or occupationally relevant. Many bioassays also merely aim at identifying a maximum tolerated dose of a chemical substance in order to provide a benchmark of toxicity. High dose regimens can raise critical concerns about the indirectness of a study, if the toxicokinetic and toxicodynamic factors by which the administered dose causes an outcome are different from those at the lower target dose level.

This is a key point of debate about the potential health effects of exposure to endocrine disrupting chemicals: if the administered high dose overwhelms the biological pathway that is involved in the endocrine activity of the active substance, then there may be critical concerns about the indirectness of the surrogate dose for determining whether the chemical of concern is an endocrine disruptor (Lagarde *et al.*, 2015). In such cases, the “biologically plausibility” of the connection between the mechanism by which the target dose causes the outcome of interest and the mechanism by which the surrogate dose is acting in the experimental environment would be very limited. In terms of GRADE, the importance of the indirectness of the evidence may be considered high when there is evidence of different toxicokinetic and toxicodynamic profiles operating at different dose levels.

In contrast, chemicals which cause cancer by a genotoxic mechanism are considered to operate according to the same mechanism of action at high and low doses (Crump, 1996). Extrapolation from across the dose range is taken as unproblematic, i.e. it is biologically plausible that the same mechanism is operating. In terms of GRADE, the importance of the indirectness of the surrogate dose would in this case be considered unimportant.

Substance

There are many chemicals to which people are potentially exposed which have very few associated toxicology studies. One means by which the potential toxicity of under-studied substances can be anticipated is by extrapolation from evidence of the toxicity of suitably similar chemicals.

For example, the UK Committee on Toxicity recently evaluated evidence of the neurotoxicity of organophosphate flame retardants (OPFRs) (UK Committee on Toxicity (COT), 2019). Part of their assessment concerned whether the neurotoxicity of OPFRs could be extrapolated from studies of the neurotoxicity of organophosphate pesticides (OPPs). COT determined that OPPs are not a good surrogate exposure for OPFRs, because OPFRs do not inhibit acetylcholinesterase to the same degree as OPPs. COT concluded that there is no “biologically plausible” explanation for how OPPs and OPFRs can cause the same effect, and therefore it is not reasonable to make inferences about the neurotoxicity of OPFRs from evidence of the neurotoxicity of OPPs. In GRADE, this absence of explanation for the mechanism by which evidence of the neurotoxicity of OPPs generalises to OPFRs would raise concerns about the importance of the indirectness of the evidence.

On the other hand, since the phase-out of consumer uses of bisphenol-A due to concerns about its potential to act in the body as an estrogen, considerable research has been conducted into whether its analogue replacements such as bisphenol-AF and bisphenol-C may have similar estrogenic potential. Enough similarities in action have been observed to suggest that, at least as a group, exposure to some bisphenols may be predictive of the effects of exposure to others (Pelch *et al.*, 2019). Similar research has been conducted into polyfluorinated compounds (Cousins *et al.*, 2020). When it is more “biologically plausible” that a surrogate and target substance share the same mechanisms by which they exert health effects, the indirectness of the surrogate becomes less important.

Biological matrix

It is often impractical or unethical to measure individual exposure to an environmental challenge. For example, air pollution is challenging to measure at the personal level (Burns *et al.*, 2020) and matrices of concern can be inaccessible, as for *in utero* exposure assessment (Gauderat *et al.*, 2017) and brain tissue (Lin, 2008). As a result, the use of surrogate exposure methods and matrices is common in environmental health research.

When using surrogate matrices, the occurrence of the exposure of concern in the matrix of interest has to be inferred from a surrogate exposure measurement in the measured matrix. For the inference to hold, it has to be “biologically plausible” that the exposure in the observed matrix represents the exposure in the target matrix - i.e. the indirectness of the surrogate matrix in relation to the target matrix should be unimportant.

In a recent series of systematic reviews of health effects of phthalate exposures, studies where phthalate levels were measured in serum rather than urine were excluded (Radke *et al.*, accepted). This was due to the high likelihood that blood samples can be contaminated with phthalates in collection and storage; these contaminants are then metabolised by enzymes in the blood such that the analysed serum gives artificially high readings of phthalate exposure levels. Since such enzymes are not in urine, phthalate contamination is not an issue: the parent compounds can be disregarded and only the metabolites measured. In the language of GRADE, phthalate levels in serum would be considered to be of greater indirectness than phthalate levels in urine because it is not “biologically plausible” that the levels seen in serum samples reflect levels in the person from whom the serum was drawn.

Cadmium is a heavy metal for which measurement in hair and toenails is an effective, non-invasive method for acquiring data about relative levels of population exposure. However, due to the complex toxicokinetics of cadmium which include its sequestration in internal organs, measurement of cadmium in hair is an unreliable indicator of whether a particular person’s current exposure level puts them at risk of damage to specific organs such as the kidney (Prozialeck and Edwards, 2010). In this case, urinary levels of β 2-microglobulin are a more direct indicator of potential renal injury from current cadmium exposure levels (Prozialeck, 2013). Since there are no “biologically plausible” explanations of how levels of cadmium in hair provide an accurate enough measure of exposure to indicate whether a

person's cadmium levels are causing renal injury, indirectness of the surrogate is important; for β 2-microglobulin the opposite is the case.

| | Surrogates of higher biological plausibility, for which indirectness is less important | Surrogates of lower biological plausibility, for which indirectness is more important |
|-------------------------------------|--|---|
| Population | Animal models for human carcinogenicity of 2-nitropropane | Rat models for human bladder carcinogenicity of saccharine |
| Exposure (dose) | Extrapolating from high-doses to low doses of genotoxic substances | Extrapolating from high doses to low doses of endocrine disrupting chemicals |
| Exposure (route) | Oral administration of bisphenol-A via gavage, or availability of a pharmacokinetic model to translate intravenous dose to oral equivalent | Intravenous administration of bisphenol-A in absence of pharmacokinetic model to translate intravenous dose to oral equivalent |
| Exposure (substance) | Inferring estrogenic potential of other bisphenols and from studies of bisphenol-A | Inferring neurotoxicity of OPFRs from studies of OPPs |
| Exposure (biological matrix) | Measurement of β 2-microglobulin levels in blood as marker of Cd exposure in the kidney | Measurement of phthalate metabolites in serum rather than urine samples; measurement of Cd in hair as marker of exposure levels in the kidney |
| Outcome | Maternal serum T4 for child neurodevelopmental outcomes | Biomarkers of Alzheimer's Disease progression in place of clinical measures |

Table 2: Summary of the examples used in this manuscript to indicate how discussion of biological plausibility maps onto the concepts of systematic review.

Discussion

The dual aspects of biological plausibility

Our examination in this manuscript of “biological plausibility” has shown it to be a complex concept which can be deployed in multiple scenarios in environmental health assessments. These potential uses extend beyond those definitions of biological plausibility which describe it exclusively in terms of biological explanations of a causal relationship between exposure and outcome (such as those in Wikipedia and in Last's Dictionary of Epidemiology, as summarised in Table 1). While heterogeneous in nature, we believe that these multiple uses of biological plausibility have two principle, related aspects in common. We call these the “generalisability aspect” and the “mechanistic aspect”.

The *generalisability aspect* of biological plausibility concerns the validity of generalisations from an experimental or observed (i.e. surrogate) population, exposure or outcome to a target population, exposure or outcome of concern. This aspect is not immediately concerned with the plausibility of causal claims about the effect of exposures on outcome, but instead about the extent to which an observation in a surrogate population plausibly generalises to a target population, a surrogate substance generalises to a target substance, etc.

While the generalisability aspect is not immediately about biology, our 12 examples do show that judgements of whether a surrogate plausibly generalises to a target context are strongly influenced by knowledge of relevant biological mechanisms. This knowledge is not always available. However, when it is, it has a large impact on judgements about the generalisability of observations in a surrogate: the higher is the certainty in the knowledge of relevant biological mechanisms, the higher is the certainty that a generalisation from a given

surrogate is valid. This certainty in biological mechanism is the *mechanistic aspect* of biological plausibility.

These two aspects are different but fundamentally linked: judgements of the plausibility of generalisations underpinned by judgements of the plausibility of mechanisms. This connection between the two aspects of biological plausibility is illustrated in Figure 3.

We believe this clarifies the concept of biological plausibility and suggests an updated working definition, as follows: *Biological plausibility is a dual-aspect concept, concerning (a) the plausibility of generalisations from research contexts to target contexts of concern, and (b) the plausibility of mechanistic explanations of biological processes. When knowledge of biological mechanisms is available, it can have considerable impact on certainty in making generalisations from research contexts to target contexts of concern.*

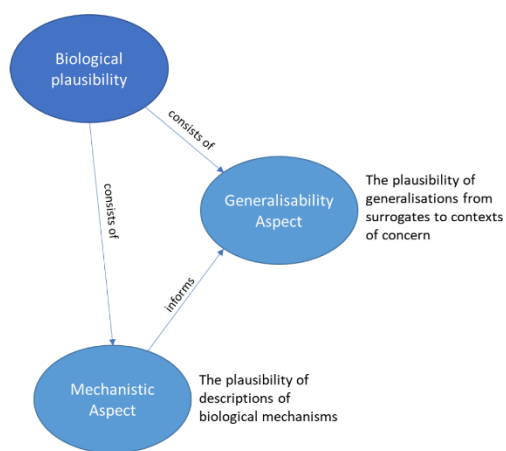


Figure 3. The relationship between the generalisability and mechanistic aspects of biological plausibility.

Implications for GRADE and systematic review

Using the example of smoking and lung cancer, we have argued that biological plausibility is inferred if direct evidence is sufficiently certain. Having identified in the concept of biological plausibility two aspects of analysis, we can further clarify this position: when direct evidence is certain, there is no need to generalise from the included evidence to the target context as defined by the review question. In such circumstances the generalisability aspect of biological plausibility does not therefore come into play, and there is no need to appeal to knowledge of biological mechanisms to inform the validity of such judgements.

However, direct evidence in environmental health research is rarely certain and very often unavailable. It is therefore frequently necessary to consider including surrogates in environmental health systematic reviews. In the systematic review process, this happens at

two stages: the defining of the eligibility criteria, when the review team decides which evidence is sufficiently generalisable to the target context to be worth including in the review; and in the assessment of indirectness, where the review team judges the extent to which the included evidence generalises to the target context.

In the language of systematic review, this assessment of generalisability is referred to as the “external validity” of evidence, i.e. the extent to which the results of an experimental or observational study apply to a target context outside of that study. This is the same issue of generalisability as that discussed when considering biological plausibility. The only difference is one of vocabulary, whereby systematic reviewers talk about the “validity” rather than “plausibility” of a generalisation. Since the generalisability aspect of biological plausibility is asking the same question as the assessment of external validity in systematic reviews, and external validity is subsumed under the GRADE domain of indirectness, it follows that there is no need to extend GRADE to accommodate this aspect of biological plausibility. How the concept of biological plausibility maps onto systematic review is illustrated in Figure 4.

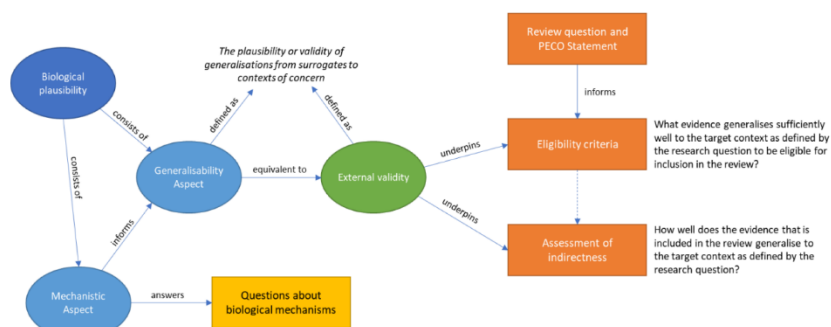


Figure 4. How biological plausibility maps onto the processes of systematic review via the shared concept of external validity. While questions about biological mechanisms (e.g. how an exposure causes an outcome) are independent of a given systematic review, answers to those questions can be highly informative in judging the external validity of evidence.

While we believe we have determined how the concept of biological plausibility maps onto systematic review and GRADE, our discussion draws further attention to two priority areas for methodological research. In both areas it would be obvious to use the experience of environmental health assessment practitioners in developing these methods.

1. How should we make judgements of external validity? The 12 examples in this manuscript show that judging external validity is complex and potentially has to be made across multiple related domains of population, exposure, outcome, and subdomains thereof. Instruments which would facilitate transparent, consistent, and accurate judgements across these domains are not yet available and should be developed.

2. How should we judge certainty in mechanism? Answering questions about biological mechanisms draws on a wide variety of information including pharmacokinetic models, information about the absorption, distribution, metabolism and excretion (“ADME”) of chemical substances, knowledge of mechanisms by which chemicals cause outcomes in both target and observed populations, and the extent to which biomarkers of disease are predictive of clinical outcomes, to name a few. If mechanistic knowledge is informative of judgements of external validity, and therefore of the indirectness domain in GRADE, then we need to develop methods for assessing certainty in mechanistic knowledge. This would be a new application of the GRADE CoE framework.

The 12 examples discussed above give some indications of the conditions under which inclusion of indirect evidence may increase certainty in the findings of a systematic review. These are outlined in Table 3 and illustrated, where feasible, in Figure 5. While these are only suggestive selections from the examples we have used in this manuscript, they do illustrate how much of this discussion is already familiar in toxicology and environmental health, and provide a robust platform of experience on which to develop answers to the two questions above.

| | Potential influencing factors in judging the biological plausibility or external validity of study surrogates |
|-----------------------------|---|
| Population | The extent to which the biological pathway connecting exposure to outcome is operating in both the surrogate population and the target population (Figure 5A) |
| Exposure – dose | The similarity of the toxicodynamic and toxicokinetic processes by which the surrogate dose acts in comparison to that of the target dose |
| Exposure – route | The similarity by which an organism absorbs and metabolises the substance of concern via the surrogate route as opposed to the target route; or the reliability with which exposure from the surrogate route can be transformed to values which match exposure from the route of interest |
| Exposure – substance | The relative affinity of the surrogate molecule for the points at which the target molecule interacts with the biological processes of interest (Figure 5C) |
| Exposure – matrix | The reliability with which levels of the substance in the observed matrix can be transformed to values in the matrix of concern (Figure 5D) |
| Outcome | The extent to which a surrogate outcome is predictive of the target outcome of concern (Figure 5B) |

Table 3: Summary of potential influencing factors in judging biological plausibility or external validity of study surrogates, as suggested by the examples in this manuscript

Finally, we note that sufficient biological knowledge to permit high-certainty judgements of mechanism and external validity is rare. Absent explanations of mechanism, evidence either ends up being excluded from a systematic review because there is no reason for considering it as eligible, or evidence would be included but its external validity would be unclear, indirectness higher as a consequence, and certainty lower overall. In these cases of low certainty due to unclear external validity of the included studies, significant mechanistic research may be required before it is possible to determine whether a given experimental model is more or less externally valid than another.

In regulatory circumstances where making decisions in the face of uncertainty is important, and mechanistic evidence to support judgements of external validity in a health assessment is limited, the external validity of a choice of surrogate is often assumed unless there is compelling evidence to the contrary (US Environmental Protection Agency, 2005).

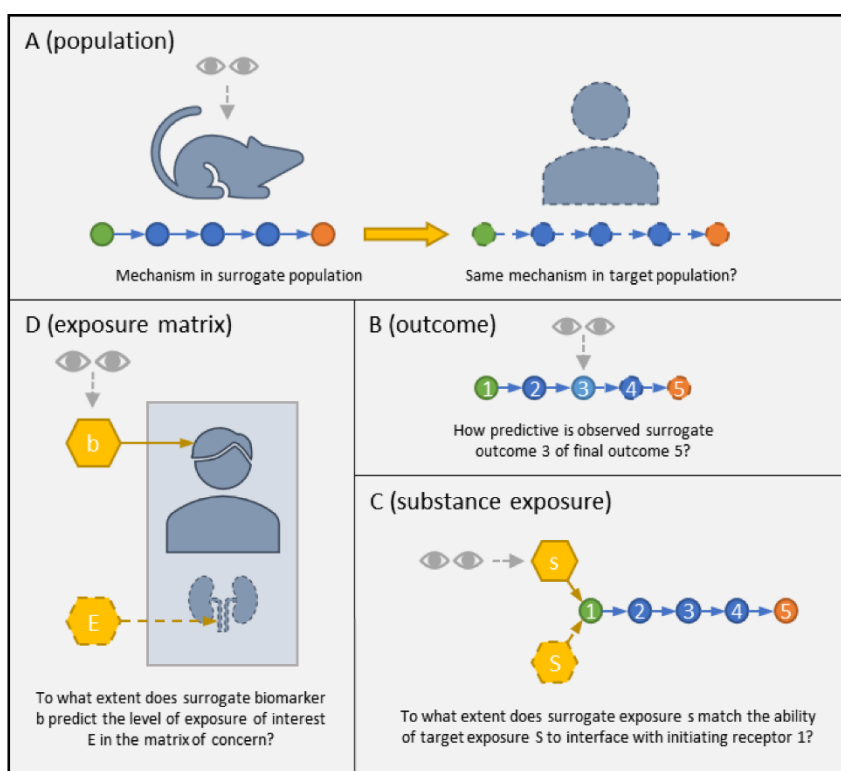


Figure 5: Illustrations of the potential influencing factors in judging biological plausibility or external validity of study surrogates, as suggested by the examples in this manuscript

Conclusion

Our analysis has elucidated Bradford Hill's proposition that establishing biological plausibility is helpful but not always necessary for a causal claim (Hill, 1965): "It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand." Biological plausibility is not necessary for determining that an exposure causes an outcome, so long as the direct evidence for the exposure-outcome relationship is sufficiently certain. Biological plausibility can nonetheless be "helpful" to establishing causation. This happens when a systematic review includes evidence from surrogates which is indirect but of sufficient external validity that it results in a more certain answer to the review question than would be yielded by inclusion of the most direct evidence alone.

Our analysis also broadens the scope of discussion in GRADE of study surrogates. Currently, GRADE guidance only explicitly addresses surrogate outcomes (Guyatt *et al.*, 2011): "Guideline developers should consider surrogate outcomes only when high-quality evidence regarding important outcomes is lacking. When such evidence is lacking [...] they should specify the important outcomes and the associated surrogates they must use as substitutes. [...] the necessity to substitute the surrogate may ultimately lead to rating down the quality of the evidence because of indirectness." Here, we have extended discussion of eligibility and potential grading of surrogate outcomes to also cover surrogate populations and surrogate exposures.

We have argued that judgements of biological plausibility, at least in their application to determining the relevance of a study to answering a research question, are accommodated by the operational procedures of systematic review and the GRADE domain of indirectness. While vocabulary and processes may differ, we feel confident that there is nothing "missing" from GRADE. What is needed, however, are means to operationalise the assessment of the indirectness of included studies and certainty in biological knowledge, the outputs of which can be used in determining the extent to which evidence should be downgraded for indirectness. Such methods would help bring shape to the amorphous nature of mechanistic evidence and aid in its exploitation in environmental health systematic reviews. Hopefully, having mapped biological plausibility onto systematic review, the development of such methods can proceed with confidence in its value for environmental health research in general - it will be of value to communities that prefer to discuss indirectness in terms of biological plausibility, and systematic reviewers can be confident that careful operationalisation of the considerations of biological plausibility will facilitate the conduct of systematic reviews.

As a final point, we observe a clear parallel between the clinical contexts in which GRADE was developed and the environmental health context in which it is now being applied. The difference is that in clinical contexts, GRADE is nearly always used to evaluate human evidence where treatments are being trialled in people, far downstream from the pre-clinical *in vitro* and animal research used to support conducting a human trial. While treatments are advanced to human trials for testing efficacy (and to large-scale observational studies for identifying other applications or potential harms) in the target population based on evidence

from preclinical studies, this evidence is usually 15-25 years old by the time a systematic review is conducted - and therefore preclinical evidence is not needed. In contrast, in environmental health contexts *in vitro* and *in vivo* research constitutes most of the evidence being dealt with. The fundamental principles for dealing with this evidence are no different, it is just the availability of human evidence that is more limited and GRADE is therefore being applied to evidence which is much further upstream than most healthcare systematic reviews have needed to worry about accommodating (though this might change as systematic review methods are taken up in the preclinical field).

Acknowledgements

The authors would like to thank the GRADE Environmental Health Project Group and GRADE Working Group for their contributions to this manuscript, and the Evidence-based Toxicology Collaboration (EBTC) at Johns Hopkins Bloomberg School of Public Health for providing funding to cover the time of PW, KT and SH in working on this manuscript. The authors would also thank the European Food Safety Authority and EBTC for organising the Scientific Colloquium, and the participants who contributed to discussions therein, which gave genesis to the concept of this manuscript (European Food Safety Authority, 2018).

Bibliography

- Bilotta, G. S., Milner, A. M. and Boyd, I. (2014) 'On the use of systematic reviews to inform environmental policies', *Environmental science & policy*, 42, pp. 67–77. doi: 10.1016/j.envsci.2014.05.010.
- Braun, J. M. and Gray, K. (2017) 'Challenges to studying the health effects of early life environmental chemical exposures on children's health', *PLoS biology*, 15(12), p. e2002800. doi: 10.1371/journal.pbio.2002800.
- Burns, J. *et al.* (2020) 'Interventions to reduce ambient air pollution and their effects on health: An abridged Cochrane systematic review', *Environment international*, 135, p. 105400. doi: 10.1016/j.envint.2019.105400.
- Cohen Hubal, E. A. *et al.* (2020) 'Advancing systematic-review methodology in exposure science for environmental health decision making', *Journal of exposure science & environmental epidemiology*. doi: 10.1038/s41370-020-0236-0.
- Cousins, I. T. *et al.* (2020) 'Strategies for grouping per- and polyfluoroalkyl substances (PFAS) to protect human and environmental health', *Environmental science. Processes & impacts*. doi: 10.1039/d0em00147c.
- Crump, K. S. (1996) 'The linearized multistage model and the future of quantitative risk assessment', *Human & experimental toxicology*, 15(10), pp. 787–798. doi: 10.1177/096032719601501001.
- Doll, R. *et al.* (2005) 'Mortality from cancer in relation to smoking: 50 years observations on British doctors', *British journal of cancer*, 92(3), pp. 426–429. doi: 10.1038/sj.bjc.6602359.
- European Food Safety Authority (2015) 'Scientific Opinion on the risks to public health

related to the presence of bisphenol A (BPA) in foodstuffs: PART II - Toxicological assessment and risk characterisation', 13(1). doi: 10.2903/j.efsa.2015.3978.

European Food Safety Authority (2018) 'EFSA Scientific Colloquium 23 – Joint European Food Safety Authority and Evidence-Based Toxicology Collaboration Colloquium Evidence integration in risk assessment: the science of combining apples and oranges 25–26 October 2017 Lisbon, Portugal', *EFSA Supporting Publications*, 15(3). doi: 10.2903/sp.efsa.2018.EN-1396.

Gauderat, G. *et al.* (2017) 'Prediction of human prenatal exposure to bisphenol A and bisphenol A glucuronide from an ovine semi-physiological toxicokinetic model', *Scientific reports*, 7(1), p. 15330. doi: 10.1038/s41598-017-15646-5.

Guyatt, G. H. *et al.* (2008) 'GRADE: an emerging consensus on rating quality of evidence and strength of recommendations', *BMJ*, 336(7650), pp. 924–926. doi: 10.1136/bmj.39489.470347.AD.

Guyatt, G. H. *et al.* (2011) 'GRADE guidelines: 2. Framing the question and deciding on important outcomes', *Journal of clinical epidemiology*, 64(4), pp. 395–400. doi: 10.1016/j.jclinepi.2010.09.012.

Hardy, A. *et al.* (2017) 'Guidance on the use of the weight of evidence approach in scientific assessments', *EFSA Journal*, 15(8). doi: 10.2903/j.efsa.2017.4971.

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (ed.) (2019) *Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019)*. Cochrane. Available at: www.training.cochrane.org/handbook.

Hill, A. B. (1965) 'The Environment and Disease: Association or Causation?', *Proceedings of the Royal Society of Medicine*, 58, pp. 295–300. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/14283879>.

Hoffmann, S. *et al.* (2017) 'A primer on systematic reviews in toxicology', *Archives of toxicology*, 91(7), pp. 2551–2575. doi: 10.1007/s00204-017-1980-3.

Hultcrantz, M. *et al.* (2017) 'The GRADE Working Group clarifies the construct of certainty of evidence', *Journal of clinical epidemiology*, 87, pp. 4–13. doi: 10.1016/j.jclinepi.2017.05.006.

International Epidemiological Association (2001) *A Dictionary of Epidemiology*. Oxford University Press. Available at: <https://play.google.com/store/books/details?id=nQmhQgAACAAJ>.

Johnson, P. I. *et al.* (2016) 'Application of the Navigation Guide systematic review methodology to the evidence for developmental and reproductive toxicity of triclosan', *Environment international*. The Authors, 92-93, pp. 716–728. doi: 10.1016/j.envint.2016.03.009.

Lagarde, F. *et al.* (2015) 'Non-monotonic dose-response relationships and endocrine disruptors: a qualitative method of assessment', *Environmental health: a global access science source*, 14, p. 13. doi: 10.1186/1476-069X-14-13.

Lin, J. H. (2008) 'CSF as a surrogate for assessing CNS exposure: an industrial perspective', *Current drug metabolism*, 9(1), pp. 46–59. doi: 10.2174/138920008783331077.

McGhee, D. J. M. *et al.* (2014) 'A systematic review of biomarkers for disease progression in Alzheimer's disease', *PloS one*, 9(2), p. e88854. doi: 10.1371/journal.pone.0088854.

- Meek, M. E. B. *et al.* (2013) 'Case study illustrating the WHO IPCS guidance on characterization and application of physiologically based pharmacokinetic models in risk assessment', *Regulatory toxicology and pharmacology: RTP*, 66(1), pp. 116–129. doi: 10.1016/j.yrtph.2013.03.005.
- Morgan, R. L. *et al.* (2016) 'GRADE: Assessing the quality of evidence in environmental and occupational health', *Environment international*. Elsevier Ltd, 92-93, pp. 1–6. doi: 10.1016/j.envint.2016.01.004.
- Morgan, R. L. *et al.* (2018) 'Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes', *Environment international*. Elsevier, (July), pp. 1–5. doi: 10.1016/j.envint.2018.07.015.
- OECD (2016) 'Users' Handbook supplement to the Guidance Document for developing and assessing Adverse Outcome Pathways', *Env/Jm/Mono(2016) 12*, (OECD Series on Adverse Outcome Pathways1), p. 63. doi: 10.1787/5jlv1m9d1g32-en.
- Papameletiou, D. *et al.* (2017) 'SCOEL/REC/300 2-Nitropropane - Recommendation from the Scientific Committee on Occupational Exposure Limits'. Directorate-General for Employment, Social Affairs and Inclusion (European Commission) , Scientific Committee on Occupational Exposure Limits. doi: 10.2767/841951.
- Pelch, K. E. *et al.* (2019) 'Characterization of Estrogenic and Androgenic Activities for Bisphenol A-like Chemicals (BPs): In Vitro Estrogen and Androgen Receptors Transcriptional Activation, Gene Regulation, and Binding Profiles', *Toxicological sciences: an official journal of the Society of Toxicology*. doi: 10.1093/toxsci/kfz173.
- Pope, C. A., 3rd *et al.* (2011) 'Lung cancer and cardiovascular disease mortality associated with ambient air pollution and cigarette smoke: shape of the exposure-response relationships', *Environmental health perspectives*, 119(11), pp. 1616–1621. doi: 10.1289/ehp.1103639.
- Prozialeck, W. C. (2013) 'Biomarkers for Cadmium', in Kretsinger, R. H., Uversky, V. N., and Permyakov, E. A. (eds) *Encyclopedia of Metalloproteins*. New York, NY: Springer New York, pp. 272–277. doi: 10.1007/978-1-4614-1533-6_33.
- Prozialeck, W. C. and Edwards, J. R. (2010) 'Early biomarkers of cadmium exposure and nephrotoxicity', *Biometals: an international journal on the role of metal ions in biology, biochemistry, and medicine*, 23(5), pp. 793–809. doi: 10.1007/s10534-010-9288-2.
- Radke, E. G. *et al.* (accepted) 'Application of US EPA IRIS systematic review methods to the health effects of phthalates: lessons learned and path forward', *Environment international*.
- Rhomberg, L. (2015) 'Hypothesis-Based Weight of Evidence: An Approach to Assessing Causation and its Application to Regulatory Toxicology', *Risk analysis: an official publication of the Society for Risk Analysis*, 35(6), pp. 1114–1124. doi: 10.1111/risa.12206.
- Schunemann, H. *et al.* (2011) 'The GRADE approach and Bradford Hill's criteria for causation', *Journal of Epidemiology & Community Health*, 65(5), pp. 392–395. doi: 10.1136/jech.2010.119933.
- Sheehan, M. C. and Lam, J. (2015) 'Use of Systematic Review and Meta-Analysis in Environmental Health Epidemiology: a Systematic Review and Comparison with Guidelines', *Current environmental health reports*, 2(3), pp. 272–283. doi: 10.1007/s40572-015-0062-z.

Taylor, J. A., Welshons, W. V. and Vom Saal, F. S. (2008) 'No effect of route of exposure (oral; subcutaneous injection) on plasma bisphenol A throughout 24h after administration in neonatal female mice', *Reproductive toxicology*, 25(2), pp. 169–176. doi: 10.1016/j.reprotox.2008.01.001.

UK Committee on Toxicity (COT) (2019) 'Statement on phosphate-based flame retardants and the potential for neurodevelopmental toxicity'. Available at: <https://cot.food.gov.uk/cotstatements/cotstatementsyrs/cot-statements-2019/cot-phosphate-based-flame-retardants-statement>.

US Environmental Protection Agency (2005) 'Guidelines for Carcinogen Risk Assessment'. Available at: <https://www.epa.gov/risk/guidelines-carcinogen-risk-assessment>.

US National Resource Council (2014) *Review of EPA's Integrated Risk Information System (IRIS) Process*. The National Academies Press. Available at: http://www.nap.edu/openbook.php?record_id=18764.

Whaley, P. *et al.* (2016) 'Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations', *Environment international*. The Authors, 92-93, pp. 556–564. doi: 10.1016/j.envint.2015.11.002.

Wikipedia contributors (2014) *Biological plausibility*, *Wikipedia, The Free Encyclopedia*. Available at: https://en.wikipedia.org/w/index.php?title=Biological_plausibility&oldid=614374435 (Accessed: 16 October 2019).

Chapter 4.

Ontologies

This chapter has been submitted to the journal *Environmental Health Perspectives*. It has been revised in response to peer-review comments and was resubmitted on 21 August 2020.

According to the Contributor Roles Taxonomy, the candidate's contribution was as follows: conceptualisation; methodology; investigation; writing (original draft); writing (review and editing); visualisation.

Candidate: _____ Date: _____

Mr. Paul A. Whaley

Supervisor: _____ Date: _____

Prof. Crispin J. Halsall

1 **Knowledge Organization Systems**
2 **for Systematic Chemical Assessments**

3

4 Paul Whaley^{1,2}, Stephen W Edwards³, Andrew Kraft⁴, Kate Nyhan⁵, Andrew Shapiro⁴, Sean
5 Watford⁶, Steve Wattam⁷, Taylor Wolfe², Michelle Angrish^{4*}

6 1. Evidence Based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

7 2. Lancaster Environment Centre, Lancaster University, UK

8 3. GenOmics, Bioinformatics, and Translational Research Center, RTI International, Research Triangle Park, NC
9 27709.

10 4. Center for Public Health and Environmental Assessment, Chemical Pollutant Assessment Division, US EPA, RTP,
11 Durham, NC, 27711

12 5. Harvey Cushing / John Hay Whitney Medical Library, Yale University, New Haven, CT, 06520. Environmental
13 Health Sciences, Yale School of Public Health, New Haven, 06520

14 6. Booz Allen Hamilton, Bethesda, MD, 20852

15 7. WAP Academy Consultancy Ltd, North Yorkshire, UK

16 *Corresponding Author: Michelle Angrish, angrish.michelle@epa.gov, 109 T.W. Alexander Rd., B231B, Durham, NC
17 27711

18 The authors declare they have no actual or potential competing financial or conflicts of interests.

19 **ABSTRACT**

20 **Background:** While the implementation of systematic review and evidence mapping methods
21 stands to improve the transparency and accuracy of chemical assessments, they also
22 accentuate the challenges that assessors face in locating, evaluating, and integrating evidence
23 for the health effects which an exposure might be causing.

24 **Objectives:** This manuscript uses systematic review and evidence mapping methods to
25 introduce how information retrieval in chemical assessment is challenged by conceptual and
26 semantic factors, i.e. variation in awareness of how assessment concepts are related to each
27 other and the language that is used to describe them. These factors render chemical
28 assessments vulnerable to the streetlight effect, whereby research efforts tend to focus only on
29 issues which are already relatively well-understood. We explain how controlled vocabularies,
30 thesauruses and ontologies contribute to potentially overcoming the streetlight effect in
31 information retrieval, making up the key components of Knowledge Organization Systems
32 (KOSs) which should enable much readier, more comprehensive access to assessment-
33 relevant information than is currently achievable. Finally, we use the example of Adverse
34 Outcome Pathways both to illustrate the challenges in developing KOSs for chemical
35 assessment and to indicate how these challenges can be overcome.

36 **Discussion:** Ontologies are an under-exploited element of effective knowledge organization in
37 the environmental health sciences. Agreeing on and implementing ontologies in chemical
38 assessment is a complex but tractable process with four fundamental steps. Successful
39 implementation of ontologies would not only make currently fragmented information about
40 health risks from chemical exposures vastly more accessible, it could ultimately enable
41 computational methods for chemical assessment which can take advantage of the full richness
42 of data described in natural language in primary studies.

43 **Key words:** ontologies; adverse outcome pathways; systematic review; evidence maps; systematic map; controlled
44 vocabulary; interoperability; artificial intelligence; chemical assessment; computational toxicology; in vitro and
45 alternative methods; new approach methods; knowledge organization systems

46

47 1. Introduction

48 Chemical assessment has seen significant improvement in the validity and utility of its outputs
49 over the last decades, in parallel with the introduction of an increasing variety of open source
50 and online tools and resources that facilitate communication, flexibility, access to information,
51 and inclusiveness of scope (National Research Council, 2007). However, further gains in the
52 quality and inclusivity of chemical assessment are being challenged by exponential growth in
53 the volume of risk-relevant research being published and a burgeoning array of innovative study
54 designs being developed by scientists for investigating health risks from chemical exposures. All
55 of this data has to be found, assembled into logical cause-effect frameworks, and evaluated as
56 to what it all means for health risks from chemical exposures. Continued improvement of
57 chemical assessment outputs therefore hinges on the development of new methods for data
58 acquisition, and the rapid, reproducible, and reusable identification of old and new scientific
59 information (Watford *et al.*, 2019).

60 In parallel to the increasing diversity, volume, and complexity of toxicological research has been
61 the development of systematic methods for reviewing (Woodruff and Sutton, 2011; Whaley *et*
62 *al.*, 2016; Hoffmann *et al.*, 2017) and mapping (Walker *et al.*, 2018; Wolffe *et al.*, 2019) evidence
63 relevant to assessing health risks posed by exposure to chemical substances. While systematic
64 methods improve the transparency and accuracy of chemical assessment products, they also
65 accentuate the challenge of locating, evaluating, and integrating the many types of study design
66 which provide evidence for the health effects which an exposure might be causing. This paper
67 discusses what systematic methods for literature-based chemical assessments are, the
68 challenges which current approaches to reporting and organizing data from toxicological
69 research presents to its systematic aggregation and analysis, and what can be done in terms of
70 evolving current "knowledge organization systems" so they better facilitate systematic
71 approaches to assessing health risks posed by exposure to chemical substances.

72 2. Systematic methods in chemical assessments

73 One of the major methodological innovations in chemical assessment over the last decade has
74 been the introduction of systematic methods for exploring and synthesizing evidence.
75 Systematic methods componentize the evidence assessment workflow, dividing it into a
76 modular sequence of steps (Figure 1). The approaches fall into two broad categories:
77 systematic reviews and systematic evidence maps. Systematic approaches are considered an
78 advance on traditional, expert-based narrative approaches to summarizing evidence because
79 they use explicit, discussable methods in each component, allowing the validity of decisions to
80 be scrutinized, assessed, and improved upon (Garg *et al.*, 2008).

81 2.1 Systematic reviews

82 Systematic review (SR) has traditionally been defined as “attempts to identify, appraise and
83 synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a
84 specific research question” and use “explicit, systematic methods that are selected with a view
85 aimed at minimizing bias” (Higgins *et al.* eds, 2019). The present authors favor defining it as a
86 methodology for testing a research hypothesis using existing evidence that employs techniques
87 intended to minimize random and systematic error and maximize transparency of decision-
88 making. Either way, systematic review breaks the evidence assessment process down into
89 discrete steps of specifying objectives, defining search strategies and eligibility criteria,
90 appraising the validity of each individual included study, synthesizing the evidence using
91 quantitative and narrative techniques as appropriate, and assessing certainty in the results of
92 the synthesis (Institute of Medicine, 2011; Higgins *et al.*, 2019; Whaley *et al.*, 2020). Each step
93 is thoroughly documented so the reader can assess the validity of each judgement being made
94 by the reviewers as they move from stating their research objective through to providing their
95 final conclusions.

96 While there have been several historical precursors to the approach, SR methods as currently
97 recognized were first formally introduced in the healthcare and social sciences in the late 1980s
98 and early 1990s (Chalmers *et al.*, 2002). Since then, SR has become a fundamental technique
99 for evaluating existing evidence of the efficacy of interventions in healthcare, education, criminal
100 justice, and other fields (Farrington and Ttofi, 2009; Braga *et al.*, 2012; Roberts *et al.*, 2017).

101 The potential value of SR methods for similarly advancing toxicology and chemical risk
102 assessment was first mooted in the published literature around the mid-2000s (Guzelian *et al.*,
103 2005; Hoffmann and Hartung, 2006). By 2014, the first SR frameworks for chemical risk
104 assessment had been published (European Food Safety Authority, 2010; Rooney *et al.*, 2014;
105 Woodruff and Sutton, 2014), with subsequent rapid uptake from regional (Schaefer and Myers,
106 2017), national (Yost *et al.*, 2019), and international agencies (Descatha *et al.*, 2020; Orellano *et*
107 *al.*, 2020).

108 2.2 Systematic evidence maps

109 SR methods function best when responding to focused questions posed in “confirmatory mode”
110 research contexts (Nosek *et al.*, 2018), where researchers are testing a hypothesis or
111 quantifying a specific exposure-outcome relationship using existing evidence in lieu of
112 conducting an experiment. However, many research contexts are not confirmatory but
113 exploratory, generating new hypotheses which might need to be tested and identifying novel
114 issues which may warrant further investigation. In these contexts the methods of systematic

115 review, developed for narrowly-defined questions, rapidly become unwieldy and demand
116 interrogation of evidence at a level of detail at odds with the broader objectives of an exploratory
117 research exercise (Wolffe *et al.*, 2020). In response to the limitations of SR methods for
118 exploratory research, systematic evidence maps (SEMs - also known as “evidence maps” or
119 “systematic maps”) have been developed.

120 SEMs are designed to apply the same principles of comprehensiveness and transparency of
121 systematic review; however, instead of answering specific research questions they result in
122 queryable databases of evidence which catalogue research of relevance to an open question,
123 theme, or policy area, developed to support a broad range of decision-making contexts (James
124 *et al.*, 2016). In a chemical assessment, the characteristics summarized in a SEM will vary
125 depending on decision-making context but will usually consist of study type, chemical or test
126 substance, population, outcome, summary results, and (potentially) indicators of the validity of a
127 study. This is much less information than required for a systematic review, with the bare
128 minimum of information required for priority-setting being extracted and stored in the map
129 database. The resulting inventory of studies and findings allows a user to make screening-level
130 decisions based on regulatory needs, outcomes of regulatory concern, research questions, etc.

131 In essence, SEMs are the application of systematic methods to scoping reviews, providing an
132 evidence-based approach to deciding when to conduct new SRs (when a confluence of
133 sufficiently high-quality evidence suggests a need for a regulatory exposure limit to be revised),
134 new primary studies (when sufficiently high-quality data required for a decision may be absent),
135 or not do anything at all (when a new confluence of data would not lead to a change in exposure
136 values). Although SEMs are one of the newest innovations in evidence synthesis methods, they
137 are already seeing uptake in the environmental and social sciences (Cheng *et al.*, 2019),
138 environmental economics (Fagerholm *et al.*, 2016) and healthcare (El Idrissi *et al.*, 2019),
139 among others. Examples from environmental health include SEMs of evidence for
140 transgenerational inheritance of health effects from environmental exposures (Walker *et al.*,
141 2018), health effects of exposure to acrolein (Keshava *et al.*, 2020), and protocols for health
142 effects of PFAS exposure (Pelch *et al.*, 2019) and interventions to reduce traffic-related air
143 pollution (Sanchez *et al.*, 2020).

144 3. The information retrieval challenge

145 Systematic methods are a natural fit for chemical assessments, providing a mechanism for
146 meeting the expectation that an assessment fully and transparently utilizes all relevant evidence
147 in the course of analyzing health risks posed by exposure to chemical substances. However, in
148 spite of the contribution made by the various on-line research platforms, databases and

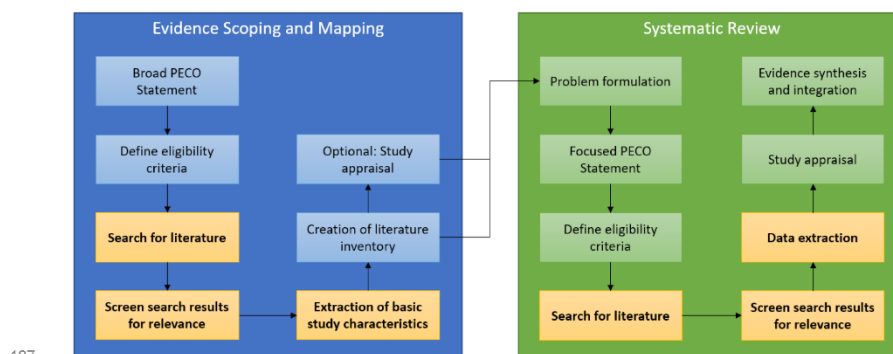
149 indexing systems which have emerged over the last three decades, the extent to which
150 evidence relevant to chemical assessments can be systematically accessed remains heavily
151 constrained by current approaches to storing and cataloguing scientific knowledge.

152 The formal record of scientific research is almost exclusively the written study report.
153 Researchers report their methods and findings in manuscripts which are published in scientific
154 journals. These documents are stored in multiple siloed databases and are retrieved using
155 complex, sensitive queries which require detailed understanding of the varying data schemas
156 and search interfaces employed by each database. Because each database is siloed, covers
157 different areas of the total literature, and stores documents in its own unique manner, these
158 searches have to be redesigned and reconducted multiple times in order to ensure all relevant
159 documents are retrieved. The searches also return a large proportion of false-positive results
160 which have to be screened out to identify the documents of true relevance to the objectives of
161 the reviewing or mapping exercise. Finally, the data in the relevant documents has to be
162 manually read and extracted into an appropriate format for analysis in the systematic review or
163 evidence map.

164 The result is a lengthy location and extraction process that may still inadvertently exclude
165 potentially large numbers of relevant records because of the "streetlight effect". This is the
166 phenomenon by which research tends to be conducted in established areas of understanding
167 rather than around novel ideas (Kaplan, 1973; Battaglia and Atkinson, 2015). While there are
168 multiple causes of the streetlight effect, database queries are affected in two principle ways of
169 relevance to our discussion. Firstly, in most databases, the content of stored documents is
170 represented using only a relatively limited selection of keywords in comparison to the full set of
171 concepts actually discussed in the documents in question, plus the words in the title and
172 abstract. This means that queries can only retrieve certain results: records where the search
173 terms happen to be concepts deemed by the database designers as important enough to be
174 cataloged in the database's keywords; and records where the search terms happen to match
175 the words used by the authors in the manuscript's title, abstract, and author keywords.
176 Secondly, only information known by the searchers or coded into the database as conceptually
177 related to the research problem can be retrieved.

178 Overcoming the streetlight effect and quickly and accurately locating and extracting relevant
179 data in scientific documents is what we refer to as the "information retrieval challenge". To
180 address this challenge, we first need to understand the difficulties which retrieving information in
181 written documents presents to developers of databases. We explain this in terms of two root
182 factors: firstly, the "semantic factor", whereby natural variation in language presents an obstacle
183 to identifying relevant research; and, secondly, the "conceptual factor", which describes how

184 limits on knowledge of the conceptual relations between research topics makes it difficult to
185 access research on themes (or in “domains”) which are related to, but not directly about, the
186 immediate topic of interest.



187

188 **Figure 1.** The relationship between the processes involved in systematically mapping and systematically reviewing
189 evidence. The elements which we discuss as the “information retrieval challenge” are highlighted in bold and yellow.
190 Comprehensive evidence maps, if they represent complete inventories of the literature, should ultimately obviate the
191 need for additional literature searches in systematic reviews conducted in response to the findings of a systematic
192 evidence mapping exercise.

193 3.1 The semantic factor

194 The language that scientists use to describe their work can be quite varied, with researchers
195 using different words for the same things (synonyms) and the same words for different things
196 (homographs and polysemes). Because meaning is a function of the relationships between
197 words and the context in which they are presented (Gasparri and Marconi, 2019), scientists can
198 even use incorrect words to describe their activities and still successfully get their meaning
199 across to a sufficiently fluent reader.

200 The flexibility of language allows it to evolve over time and enables us to use familiar words to
201 talk about new things in our changing physical and intellectual worlds (Sorensen, 2018).
202 However, this variation and evolution in natural language also presents significant challenges to
203 the information retrieval process: not only do databases have to be engineered to accommodate
204 such variation, because approaches to accommodating the variation differ from one database to
205 the next, a database user has to be aware of both the variation in the way language is being
206 employed by authors of the documents in which they are interested, and also how this variation
207 is handled by the database itself, in order to design searches which maximize the amount of
208 relevant literature being retrieved.

209 This is why complex search strings need to be used in querying research databases, to cover
210 the many different ways of expressing the same concepts. It is also why the strings have to be
211 different for each database. There is no one correct way of solving the problem of variation in
212 language, just different optimizations - hence, the designers of each database end up
213 implementing different solutions fashioned with different priorities in mind given the database's
214 intended use.

215 If an information retrieval strategy does not include all the words that have been or are being
216 used for the concepts of interest, in a way which responds to the individual characteristics of the
217 database being searched, then relevant documents will be overlooked. This is one of the
218 reasons why information specialists are needed for SR projects (Rethlefsen *et al.*, 2015). An
219 example of how linguistic variation can affect the number of results retrieved for a search
220 concept is illustrated in Table 1, where different terms for the same concept can return different
221 results within and across databases.

222 **Table 1.** Demonstration of how variation in language used by study authors in title, abstract, and author keywords
223 fields affects search results in PubMed. Database syntax is used to ensure the phrase entered is the exact one being
224 searched for. Date of searches: 15 July 2020.

| PubMed Query | Results |
|---|---------|
| "PAHs"[Title/Abstract] OR "PAHs"[Other Term] | 15,912 |
| "PAH"[Title/Abstract] OR "PAH"[Other Term] | 22,605 |
| "polycyclic aromatic hydrocarbon"[Title/Abstract] OR "polycyclic aromatic hydrocarbon"[Other Term] | 4,545 |
| "aromatic polycyclic hydrocarbons"[Title/Abstract] OR "aromatic polycyclic hydrocarbons"[Other Term] | 59 |
| "polycyclic aromatic hydrocarbons"[Title/Abstract] OR "polycyclic aromatic hydrocarbons"[Other Term] | 19,311 |

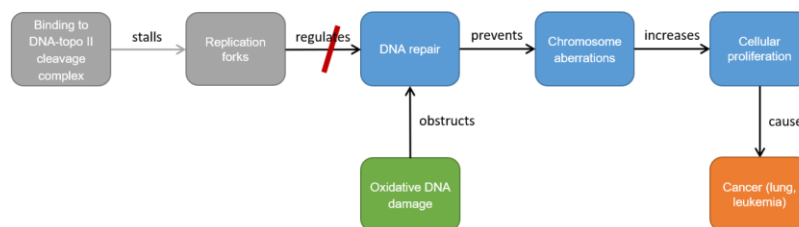
225

226 3.2 The conceptual factor

227 For any given domain of interest, there will be an expansive network of related concepts and
228 sub-concepts of relevance to a SR or SEM exercise. Having a complete map of the
229 relationships between these concepts is necessary if the full body of relevant information is to
230 be retrieved (Figure 2); however, expert knowledge is finite, which means that important
231 relationships outside the knowledge sphere of the expert conducting a review are always at risk
232 of being missed. This is illustrated in Figure 2. Here, an expert might be aware that DNA strand
233 breaks are related to inadequate DNA repair, and is consequently able to include exposures

234 which increase oxidative DNA damage in a cancer assessment. However, if the expert is not
235 aware that DNA strand breaks are also related to collapse of stalled replication forks, then
236 research into this event and others which are related to it may be overlooked in the assessment.

237 Other examples of conceptual relationships which are often of relevance in synthesizing
238 evidence to answer a research question but might not be known to researchers conducting a
239 SR or SEM exercise include comparable chemicals, where the known effects of exposure to
240 one substance can be informative of the potential effects of another; biologically-comparable
241 species, where an animal might serve as a model for disease in humans; and surrogate
242 outcomes, where an upstream biomarker of health effects might be a strong predictor of a final
243 health outcome. Some of these conceptual relationships will be known and some speculative;
244 however, unless they are accounted for in an information retrieval strategy, evidence of potential
245 importance for answering a given question may be overlooked.



246

247 **Figure 2.** Illustration of how lack of knowledge of relations between concepts relevant to a research topic can result in
248 evidence of potential importance to a given question being overlooked. In this example, awareness that DNA repair is
249 obstructed by oxidative DNA damage allows lung cancer and leukemia to be connected to stressors which cause
250 oxidative DNA damage to be incorporated into a cancer assessment. However, lack of awareness that replication
251 forks regulate DNA repair may result in studies of stressors which stall replication forks by binding to cleavage
252 complexes being excluded from cancer assessments.

253 The conceptual factor is fundamental to the streetlight problem in information retrieval: without
254 measures to augment search terms with related concepts, search strategies can only find
255 information on concepts which the searcher already knows to be related to those specified in
256 the research question. Addressing the challenge of finding what is relevant, but not necessarily
257 known by the searcher as relevant, is a central element of modern information retrieval
258 strategies.

259 **3.3 The conceptual and semantic factors in SRs and SEMs**

260 One strategy for addressing the conceptual factor in the information retrieval challenge, which at
261 least ensures saturation of concepts in relation to a research question, is simply to narrow the
262 topic of the review. A narrower question entails fewer related concepts, which makes it easier to
263 ensure all relevant concepts are known and accounted for. This is a fundamental component of
264 current practice in SR, whereby a tightly focused research objective is a common
265 recommendation (Institute of Medicine, 2011; Morgan *et al.*, 2018) and helps ensure that a
266 finitely-resourced research project provides comprehensive coverage of the topics it needs to
267 include in order to answer its question.

268 The problem with topic-narrowing as a strategy is that it deliberately excludes evidence which
269 may be relevant to the review question, on the assumption that the excluded evidence is going
270 to be insufficiently informative to materially alter the conclusions of the review. This may be a
271 reasonable assumption for SRs where the knowledge objective is very specific. However, it is
272 much less available as a strategy for SEM exercises, where the purpose is to map domain
273 topics and the evidence associated with them in a broad thematic or policy area rather than in
274 relation to a specific question (Miake-Lye *et al.*, 2016; Saran and White, 2018).

275 Whether narrow or broad, the same structural issue is faced by SRs and SEMs. Researchers
276 need to access a universe of information but because they only know or are able to recall a
277 certain proportion of terms for and linkages between concepts, they only have partial access to
278 the full universe of information they need. This situation can be improved by groups of experts
279 working together using effective elicitation strategies; however, their view of the evidence will
280 still be biased by what they can collectively access - the streetlight might be larger, but it still
281 offers only partial illumination. The information which users can actually access is only a
282 fragmented representation of all that which is actually known. To reduce this fragmentation, to
283 allow movement across conceptual linkages which are unknown to particular individuals or
284 groups, requires systems which make those linkages accessible without the end-user having to
285 be aware of them.

286 We call these approaches to providing access to information "Knowledge Organization
287 Systems" (KOS) and discuss how their evolution, particularly the introduction of ontologies, is
288 fundamental to the ongoing modernization of chemical assessments.

289 4. Knowledge Organization Systems

290 Here, three KOS approaches are discussed: controlled vocabularies, thesauruses, and
291 ontologies. While controlled vocabularies and thesauruses are well-established KOSs in
292 chemical assessment, the value of broader adoption of ontologies is highlighted.

293 4.1 Controlled vocabularies

294 A controlled vocabulary (CV) is a defined list of words and phrases used to tag content in a
295 database, to make that content retrievable via navigation or search. It is a type of metadata
296 (data about data) which provides an interpretive layer between the user of a database and the
297 content in the database. CVs can be used in tools that expand, translate, or map user queries to
298 the terminology used to classify content in the database, and sometimes to map additional entry
299 terms (synonyms) which the user may not have applied but the CV defines as being
300 semantically equivalent to the terms in the user query (Ashburner *et al.*, 2000; Stearns *et al.*,
301 2001; Fragoso *et al.*, 2004).

302 In its simplest form, a CV is a consistent labeling system in which the same concept is always
303 given the same name (e.g. "PAH" → "polycyclic aromatic hydrocarbons", "polycyclic aromatic
304 hydrocarbon" → "polycyclic aromatic hydrocarbons", "aromatic polycyclic hydrocarbons" →
305 "polycyclic aromatic hydrocarbons"). In a database which tags all records about a concept with
306 the same CV label, the user is able to retrieve all documents known to the system as discussing
307 that concept, independent of the author's terminology, simply by searching for the CV label
308 ("polycyclic aromatic hydrocarbons"). The CV allows the user to do this without needing to
309 specify each individual synonymous term, the full range of which the user may not have access
310 to. This utility is illustrated by the CV terms of the Medical Subject Headings (MeSH) used to
311 index research in the Medline database (see Figure 3).

312 CVs are one approach to addressing the semantic factor in information retrieval, increasing the
313 recall of queries by augmenting users' search terms with a set of synonyms. They can also
314 improve the precision of a query by disambiguating word senses (e.g. 'bank' as a mound of
315 earth, rather than as a place to deposit money) and reducing false positives (a paper about the
316 use of pesticides in the home will not be indexed as occupational exposure). CVs can, however,
317 reduce recall if the user is expecting to find a concept not included in the CV, if indexers (human
318 or machine) fail to assign relevant terms, or if some records are not tagged with CV terms at all.

319 The main limitation of CVs, in terms of their function as part of a KOS, is they capture only one
320 type of logical relationship between concepts, i.e. an equivalence relation where one thing is
321 defined as being the same as another thing (e.g. PAH → polycyclic aromatic hydrocarbons).

322 While capture of synonyms that are unknown to a system user is valuable, there are other types
323 of relationship which, if they can be built into a KOS, go further in overcoming the semantic
324 factor in information retrieval.

325 **Figure 3.** The MeSH CV entry for “polycyclic aromatic hydrocarbons”, 21 July 2020.

| Controlled vocabulary terms for “polycyclic aromatic hydrocarbons” in the Medical Subject Headings (MeSH) index (NCBI 2020) |
|---|
| Polycyclic Aromatic Hydrocarbons Aromatic hydrocarbons that contain extended fused-ring structures. Year introduced: 2017(1996) |
| Tree Number(s): D02.455.426.559.847, D04.615 MeSH Unique ID: D011084 |
| Entry Terms: <ul style="list-style-type: none">• Aromatic Hydrocarbons, Polycyclic• Hydrocarbons, Polycyclic Aromatic• Polynuclear Aromatic Hydrocarbons• Aromatic Hydrocarbons, Polynuclear• Hydrocarbons, Polynuclear Aromatic• Polycyclic Hydrocarbons, Aromatic• Aromatic Polycyclic Hydrocarbons• Hydrocarbons, Aromatic Polycyclic |

326

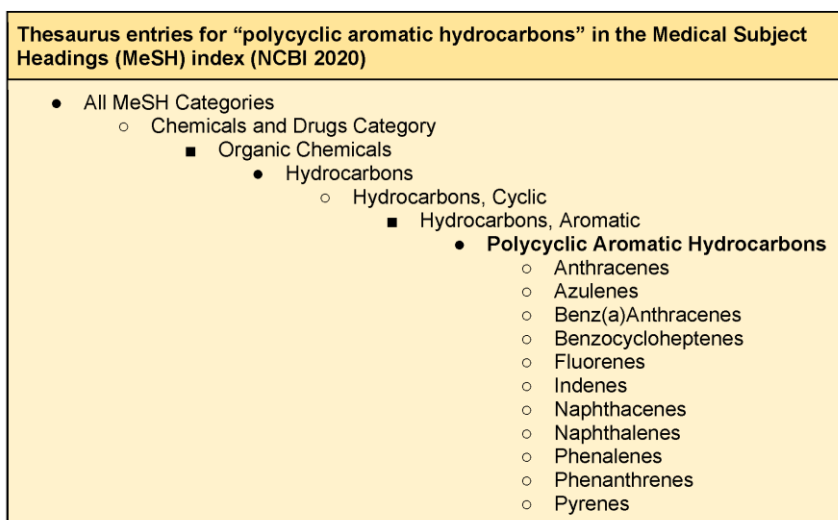
327 [4.2 Thesauruses](#)

328 Thesauruses expand beyond the equivalence relation of synonymy by introducing an
329 overarching conceptual hierarchy in which the CV terms are organized and related. Such
330 hierarchies are valuable for a KOS because they allow information consisting of related but non-
331 equivalent concepts to be defined as relevant to a user’s search term. By organizing concepts in
332 terms of how they are related, rather than simply in terms of when two terms refer to the same
333 concept, the introduction of a thesaurus begins to address the conceptual factor in information
334 retrieval.

335 An illustration of this is the MeSH thesaurus, which organizes MeSH CV terms in a parent-child
336 hierarchy (see Figure 4). This “is a class of” type of logical relationship can be exploited for
337 greater recall in search results than allowed by an equivalence relation. For example, a PubMed
338 search for “polycyclic aromatic hydrocarbons” using MeSH headings will return citations which
339 have not only been indexed with terms synonymous with polycyclic aromatic hydrocarbons, but
340 contain terms which are subclasses thereof - such as anthracenes, fluorenes, and pyrenes.

341 This is not possible in a CV alone because while a pyrene is a type of polycyclic aromatic
342 hydrocarbon, it is not equivalent to one: it is false to state “pyrene → polycyclic aromatic
343 hydrocarbon”. Since CVs are restricted to the equivalence relation, they have no mechanism to
344 describe the relationship between pyrenes and polycyclic aromatic hydrocarbons, and therefore
345 have to treat them as unrelated entities. When being queried, a system employing only a CV
346 thus requires the user to enter terms for each subclass of polycyclic aromatic hydrocarbons. If
347 the user does not know all the subclasses, then citations which are about e.g. pyrenes but do
348 not use the term “polycyclic aromatic hydrocarbon” would be missing from the search results,
349 even though they are relevant to the user’s information needs.

350 **Figure 4.** The MeSH thesaurus entries for “polycyclic aromatic hydrocarbons”, 21 July 2020. For brevity, only first-
351 level entries are shown.



352

353 In developing a more comprehensive taxonomy of the concepts which have been labelled by
354 the CV, simply by adding the “is a class of” relationship via a thesaurus, MeSH greatly increases
355 the conceptual coverage of a user’s search for PAHs without the user needing to account for
356 related, but not synonymous, terms in their search.

357 4.3 Ontologies

358 Thesauruses, as hierarchical taxonomies, are a powerful strategy in KOS development. When
359 implemented comprehensively and fully exploited by a user, they make a significant contribution

360 to addressing the semantic and conceptual factors in the information retrieval challenge.
361 However, being able to codify more information about the relationships between the concepts of
362 the CV than simple hierarchies can further increase the information retrieval capacity of a KOS.
363 After all, there are many more types of relationship than "is a class of", however powerful that
364 relationship is as a general organizing principle.

365 When a taxonomy moves beyond a hierarchy toward a representation of the properties of and
366 the relations between concepts, it becomes an ontology. An ontology is a formal method for
367 representing knowledge, usually within a particular knowledge domain, that relates terms or
368 concepts to one another in a format that supports reading and searching not only for the terms
369 themselves, but also for the relationships between those terms (Whetzel *et al.*, 2011). Using an
370 ontology allows knowledge to be stored in a mathematical graph, which is a well-studied
371 structure that has many useful properties in terms of searching and/or querying.

372 Returning to our example of cancer and DNA damage, Figure 2 provides a visual representation
373 of the richer way in which an ontology can relate concepts to each other in a graphical schema,
374 with concepts (nodes) related to each other via edges. The ontology is not restricted to being
375 hierarchical, as both nodes (the things in the database) and the edges between nodes (the
376 relationships between them) can be the object of a controlled vocabulary and carry semantic
377 value. This allows highly specific relationships such as "stalls" and "regulates" to be represented
378 in the KOS, enabling information about those relationships, or things related by those
379 relationships, to be retrieved. Queries can be written which trace a path through the graph, in
380 principle returning information about e.g. replication forks and oxidative damage in relation to
381 DNA repair, whether or not the user is aware of any relationships between the concepts.

382 5. Building an ontologized KOS

383 A KOS which incorporates ontologies can be used for much more complex information retrieval
384 tasks than one which only incorporates thesauruses because the ontologized KOS is able to
385 represent complex connections between units of information. This is particularly valuable for
386 making systematically accessible information which is indirectly related to an exposure-outcome
387 relationship of concern but nonetheless informative for a chemical assessment.

388 The challenge with the development and implementation of ontologies is that, while they provide
389 a formal way of representing knowledge in a domain, they rely on the existing knowledge within
390 that domain to determine how that knowledge is organized: the system needs to be known in
391 order to be described, yet needs to be described in order to be known. We now use the

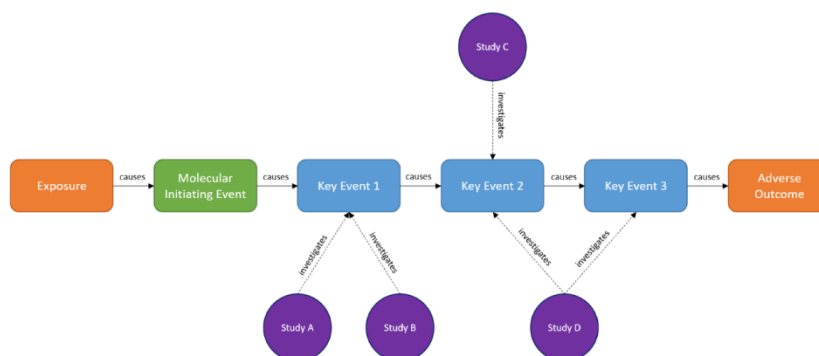
392 development of Adverse Outcome Pathways (AOPs) as an example to illustrate this challenge
393 of building ontologies and indicate how it can be solved.

394 5.1 Adverse Outcome Pathways as an Example of an Ontologized KOS

395 AOPs are a way of formalizing, for risk assessment purposes, the steps by which a disease
396 progresses from exposure through to final adverse outcome via increasing levels of biological
397 complexity (Knapen *et al.* 2018). The AOP framework was designed to provide a consistent
398 description of toxicological mechanisms across differing levels of biological organization and to
399 account for gaps in our knowledge concerning these mechanisms. They are of interest in
400 chemical assessments because they provide a means of integrating different assays targeting
401 varied components of a biological system and organizing the evidence to identify data gaps. As
402 such, they provide a means for incorporating mechanistic data into chemical assessments
403 (Arzuaga *et al.*, 2019a; Arzuaga *et al.*, 2019b).

404 Specific interpretations vary as the concept is still under development, but AOPs are essentially
405 logic models which connect an initial exposure to an outcome via a sequence of biological
406 events (Villeneuve *et al.* 2018). The sequence of events begins with a Molecular Initiating Event
407 (MIE), where a stressor initiates a biological change at the molecular level in a cell in an
408 organism. Activation of the MIE initiates progression through a sequence of Key Events (KEs)
409 occurring at increasing levels of biological complexity - from subcellular to cellular to organ, to
410 whole organism, to population. The final event in the chain of Key Events is the Adverse
411 Outcome (AO).

412 Although nascent, the AOP framework is an example of an ontologized KOS. By connecting
413 relevant literature to Key Events, and Key Events to each other using logical relationships
414 (known as "Key Event Relationships" or KERs), an AOP allows the full evidence space around
415 an exposure-outcome relationship to be accessed from any single entry-point. This allows
416 system users who lack any prior knowledge of the AOP to access connected evidence within
417 that space. For example, as shown in Figure 3, it is possible to move upstream from the AO to
418 Assays A and B via KEs 3, 2 and 1 – thereby incorporating information in the chemical
419 assessment which might otherwise have been excluded by searches or inclusion criteria
420 focusing on the AO alone (National Research Council, 2007; Schwarzman *et al.*, 2015).



421

422 **Figure 5** The elements of an Adverse Outcome Pathway, whereby an exposure causes a Molecular Initiating Event,
423 initiating a biological sequence of causally-related Key Events which result in a final Adverse Outcome being
424 manifest. Experimental research can target how a challenge might affect a Key Event (Studies A, B, and C) or how
425 one Key Event might cause another Key Event in a Key Event Relationship (Study D). Arranging biological events,
426 exposures and the evidence around them in these sorts of AOP chains can be very valuable for integrating
427 mechanistic evidence into chemical assessments but requires knowledge organization systems capable of reflecting
428 the complexity and heterogeneity of the relationships and event types.

429 5.2 How the conceptual and semantic factors challenge the building of AOPs

430 Contemporary methods for development of AOPs rely exclusively on human expert knowledge
431 of the mechanisms and biological pathways from which the AOP is ultimately derived. As such,
432 AOPs lack transparency and are highly vulnerable to both the semantic and conceptual factors
433 in information retrieval, and therefore unlikely to be based on an evaluation of the complete
434 evidence base which is relevant to their development.

435 Currently, an AOP author will define the key events associated with an AOP based on their
436 expert knowledge of the mechanisms based on one or more prototypical stressors. The author
437 ties assays and biomarkers that are associated with each of the steps leading towards the
438 adverse outcome of interest to the underlying biological events they represent. The AOP author
439 then assembles the literature that supports the linkages between each pair of key events and
440 evaluates the overall strength of the evidence supporting each linkage based on guidance
441 provided by the OECD AOP Development Programme (OECD, 2016).

442 This process is challenged by the breadth of knowledge required to fully understand the entire
443 toxicological pathway, covering literature from molecular, physiological, clinical, and
444 epidemiological domains. With the overwhelming number of publications in the scientific

445 literature, it is impossible for a small group of experts to be fully aware of the complete evidence
446 base and, therefore, the entire universe of biological concepts relevant to the AOP from across
447 all related knowledge domains.

448 In theory, this problem should be solvable using systematic methods. One should be able to
449 systematically map the scientific literature (the evidence base) to develop a model of the current
450 known biology and identify candidate KEs. Systematic review methods could then be used to
451 evaluate the relationship between each pair of candidate KEs by considering the upstream key
452 event as the "exposure" and the downstream key event as the "outcome". Those candidate KEs
453 which attain a sufficiently high level of certainty as being causally related would be elevated to
454 formal KEs and become part of the approved AOP.

455 The problem is, this map-and-review approach is not practically feasible. Literature databases
456 currently only represent a minority of AOP concepts in their controlled vocabularies, while
457 representation of the relationships between the concepts is more limited still. While these issues
458 can to some degree be mitigated by running large numbers of complex, iterated searches which
459 spider out to related concepts and terms for those concepts, such searches are challenging and
460 time-consuming to develop and their completeness difficult to validate. They are still dependent
461 on expert knowledge and painstaking analysis of the literature to map the relevant components
462 of the biology when developing AOPs.

463 In response to the challenges of mapping and reviewing such a complex evidence base, AOPs
464 have generally been developed in the publicly-accessible AOP Wiki (<https://aopwiki.org/>), a
465 resource that facilitates crowd-sourcing while also implementing some controlled vocabularies
466 and descriptors of AOP components. However, the number of experts who can realistically
467 contribute tends to in fact be quite small, and the system is still vulnerable to the streetlight
468 effect. According to the AOP Wiki, only 16 AOPs of the 306 in development have been
469 endorsed. The 306 in development represent only a fraction of the thousands of biological
470 processes we know we could be evaluating.

471 5.3 Escaping the streetlight

472 Recent developments in AOPs illustrate four core steps in an overall general strategy for
473 addressing the streetlight problem and overcoming the semantic and conceptual factors in
474 retrieving information about health risks posed by exposure to chemical substances. These
475 steps need further development in the AOP sphere and apply in general to the development and
476 implementation of ontologized KOSs in toxicology and environmental health.

477 **Step 1. Enumerate AOP-relevant entities, how they are related, and specify the**
 478 **vocabulary for labeling them.** The first step in developing an ontologized KOS is to define the
 479 things which are to be covered by the ontology (the entities), the ways in which those things are
 480 related (the relationships between the entities), and the terms which will be used to label the
 481 entities and relationships (the controlled vocabulary). This is a bootstrapping exercise of
 482 iteratively defining, mapping and refreshing the conceptual framework which constitutes the
 483 ontology. It is based on expert knowledge and active surveillance of the literature. In at least its
 484 initial phase it is conducted manually before computationally assisted approaches can be
 485 applied later.

486 An AOP ontology has already been developed within the international AOP KnowledgeBase
 487 (<https://aopkb.oecd.org/>) and incorporates terms from existing biological ontologies into the AOP
 488 descriptions within the AOP KnowledgeBase (Ives *et al.*, 2017). Some existing ontologies and
 489 how they relate to levels of cellular organization in an AOP are shown in Figure 6, indicating
 490 options for how the AOP ontology might be extended in the future. There has also been work on
 491 semantically defining AOPs (Wang *et al.*, 2019; Wang, 2020), which may also inform these
 492 efforts in the future. The Gene Ontology Causal Activity Model (Thomas *et al.*, 2019) is
 493 suggestive of an approach to defining the relationships between events in an AOP.

| | Adverse Outcome Pathway Level | | | | | | |
|---|-------------------------------|----------------------------|----------------|--------------|-------------|----------------------------|----------------------------|
| | Exposure | Molecular Initiating event | Cellular Event | Tissue Event | Organ Event | Individual Adverse Outcome | Population Adverse Outcome |
| Biological Information Ontologies | | | | | | | |
| CHEBI | x | x | | | | | |
| PRO | | x | | | | | |
| GO | | x | x | | | | |
| CL | | | x | | | | |
| UBERON | | | | x | x | | |
| MP | | | | x | x | x | |
| MonDO | | | | | | x | |
| PCO | | | | | | | x |
| Measurement Information Ontologies | | | | | | | |
| ECTO | x | | | | | | |
| BAO | | x | x | x | | | |
| EFO | | | x | x | x | | |
| SNOMED CT | | | | x | x | x | |
| CHEAR | | | | | | x | x |

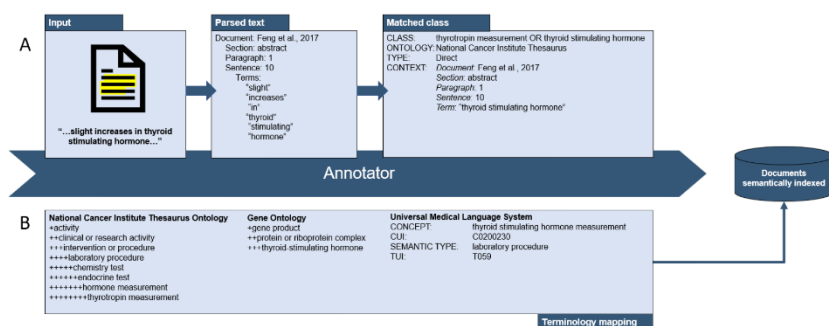
494
 495 **Figure 6.** Existing biological ontologies can be used to define key events in computable terms and thereby make
 496 AOP information more interoperable with other toxicological data sources. The same can be done when describing
 497 the assays and biomarkers used to measure the key events. CHEBI = Chemical Entities of Biological Interest, PRO =
 498 Protein Ontology, GO = Gene Ontology, CL = Cell Ontology, UBERON = Uber Anatomy Ontology, MP = Mammalian
 499 Phenotype Ontology, MonDO = Mondo Disease Ontology, PCO = Population and Community Ontology, ECTO =

500 Environment Exposure Ontology, BAO = BioAssay Ontology, EFO = Experimental Factor Ontology, SNOMED CT =
501 SNOMED Clinical Terms, CHEAR = Children's Health Exposure Analysis Resource.

502 **Step 2. Catalog the evidence for hypothetical relationships.** In developing an AOP, at least
503 some minimum evidence for the existence of an entity or a relationship needs to be identified in
504 order for something to be put forward as a candidate Key Event or Key Event Relationship. This
505 can be as little as a speculatively hypothesized relationship in a single document (even if the
506 relationship proves false, this is still part of the knowledge which the ontology is being used to
507 map and would need to be catalogued). If enough evidence with the appropriate agreed
508 characteristics accumulates, candidate events and relationships can be elevated to being Key.

509 Evidence is put behind relationships by tagging natural language expressions in relevant
510 research documents with authorized terms from controlled vocabularies, a process illustrated in
511 Figure 7. This ensures the ontology developed via the expert process of Step 1 is associated
512 with the real-world knowledge which the ontology is intended to describe. It also allows spurious
513 relationships and factually non-existent entities to be discarded. Both manual and automated
514 methods are required for tagging the literature with concepts from the ontology. In the early
515 stages, the process is almost exclusively manual, with an essential role for editors and
516 biocurators in annotating documents. This is well documented in, for example, the Gene
517 Ontology (Poux and Gaudet, 2017).

518 Because the rate-limiting step in creating annotations is the physical process of reading and
519 tagging the scientific literature, it is necessary to automate the annotation of documents in order
520 to scale the application of the ontology to the growing volume of new research. Here, the results
521 of manual annotation exercises can be used as training data for automated methods for tagging
522 free text with controlled vocabularies. Conducting SEMs and SRs provides an opportunity to do
523 this: with the right tools and training, data extractors should in principle be able to annotate the
524 documents included in their map or review. Natural Language Processing (NLP) techniques
525 including Named Entity Recognition for tagging entities and sentiment analysis for identifying
526 relationships will be central to automation (Marshall and Wallace, 2019; O'Connor *et al.*, 2020).
527 Various other machine learning applications could drastically reduce the time needed to review
528 and vet evidence (Witwehr *et al.*, 2020). The use of semantic authoring tools that would render
529 new studies machine-readable (Eldesouky *et al.*, 2016; Oliveira *et al.*, 2017; Oldman and
530 Tanase, 2018) would obviate many of the challenges in annotating research documents and
531 should be explored for toxicology and environmental health contexts.



532

533 **Figure 7.** The workflow for matching natural language strings in research reports to a hierarchy of concepts in an
534 ontology. Natural language information is extracted from included studies (e.g. phrases such as "increase in thyroid
535 stimulating hormone") into an evidence inventory (A). The terms "increase", "thyroid", "stimulating" and "hormone" are
536 cleaned and mapped to ontological classes in preparation for integration with other data sets. The inventory can then
537 be connected to other data models by mapping terminology between CVs (B). Done enough times, a large data
538 inventory begins to accumulate.

539 **Step 3. Integrate different systems.** In the case of the biological mechanisms that underlie the
540 indirect evidence supporting a chemical assessment, we have not one but many domains of
541 knowledge. In addition, we have gaps in that knowledge even for the most well-studied
542 toxicological mechanisms. As a result, a framework is needed that can incorporate and
543 represent biological knowledge in an interoperable (the ability for systems to exchange and use
544 information) network of resources including visualizations, workflows, and computational
545 pipelines that are on-line, interactive, and automatically updated. They must make intelligible to
546 the user the knowledge from the many domains and explicitly account for missing information.

547 Illustrative examples of such systems include the Health Assessment Workplace Collaborative
548 (HAWC, <https://hawcprd.epa.gov/portal/>), the US EPA Chemicals Dashboard (Williams *et al.*,
549 2017), the US EPA ChemView Portal (<https://chemview.epa.gov/chemview>), and the AOP
550 KnowledgeBase; however, while these are functional and interactive depots for aggregating
551 toxicological information, they are not yet interoperable. Achieving interoperability will require
552 data management and stewardship which promotes the FAIR principles of information
553 findability, accessibility, interoperability, and reusability (Wilkinson *et al.*, 2016; Watford *et al.*,
554 2019).

555 **Step 4. Apply and evaluate.** Making research machine-readable by tagging free text with
556 controlled vocabulary terms from an ontology enables the use of computationally intelligent tools
557 and applications for semi-automating a literature-based chemical assessment. Some predictive
558 toxicology applications based on AOPs have already been attempted (Burgoon, 2017). Any

559 computational approach using input from a complex KOS needs to be evaluated and tested, to
560 see if it produces better results for the same task as a different method. A recent example of this
561 is the comparison of *in silico* approaches to *in vivo* assays and human data for identifying skin
562 sensitizers (Luechtefeld *et al.*, 2018; Golden *et al.*, 2020).

563 6. Conclusion

564 Chemical assessment largely involves the analysis of evidence which is only indirectly related to
565 the target populations, exposures and outcomes of concern. Surrogate populations are used
566 because experimental toxicology is unethical in humans, so animal and *in vitro* models need to
567 be used instead. For many chemicals (and by definition for novel substances) few studies have
568 been conducted, requiring their potential toxicity to be inferred from suitably similar chemicals
569 whose characteristics are better understood. Evidence of health outcomes may also be sparse.
570 This is especially the case for diseases with long latency periods, such as certain brain cancers,
571 or those which cannot be observed in a test system, such as when an *in vitro* model is being
572 used for an apical outcome.

573 However, because it is only indirectly related to the target population, exposure and outcome of
574 concern, such evidence currently requires expert knowledge to locate and is therefore
575 vulnerable to the semantic and conceptual factors in information retrieval. The use of formal
576 systems such as ontologies promise to allow us to organize and unify this disparate information,
577 overcoming the streetlight effect, and making scientific knowledge generally accessible for use
578 in chemical assessment.

579 Acknowledgements

580 The authors would like to thank George Woodall, Shannon Bell, Janice Lee, and Kris Thayer for
581 their technical review. The authors would also like to thank Kristan Markey for conceptual and
582 intellectual knowledge contributions. Funding for this study came from the U.S. Environmental
583 Protection Agency Office of Research and Development. The work described in this article has
584 been reviewed by the Center for Environmental and Public Health Assessment of U.S.
585 Environmental Protection Agency and approved for publication. The views expressed in this
586 paper are those of the authors and do not necessarily reflect the views or policies of the U.S.
587 Environmental Protection Agency. Mention of trade names or commercial products does not
588 constitute endorsement or recommendation for use.

References

- Arzuaga, X., Smith, M. T., *et al.* (2019a) 'Proposed Key Characteristics of Male Reproductive Toxicants as an Approach for Organizing and Evaluating Mechanistic Evidence in Human Health Hazard Assessments', *Environmental health perspectives*, 127(6), p. 65001. doi: 10.1289/ehp5045.
- Arzuaga, X., Walker, T., *et al.* (2019b) 'Use of the Adverse Outcome Pathway (AOP) framework to evaluate species concordance and human relevance of Dibutyl phthalate (DBP)-induced male reproductive toxicity', *Reproductive toxicology*. doi: 10.1016/j.reprotox.2019.06.009.
- Ashburner, M. *et al.* (2000) 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium', *Nature genetics*, 25(1), pp. 25–29. doi: 10.1038/75556.
- Battaglia, M. and Atkinson, M. A. (2015) 'The streetlight effect in type 1 diabetes', *Diabetes*, 64(4), pp. 1081–1090. doi: 10.2337/db14-1208.
- Braga, A., Papachristos, A. and Hureau, D. (2012) 'Hot spots policing effects on crime', *Campbell Systematic Reviews*, 8(1), pp. 1–96. doi: 10.4073/csr.2012.8.
- Burgoon, L. D. (2017) 'The AOPOntology: A Semantic Artificial Intelligence Tool for Predictive Toxicology', *Applied In Vitro Toxicology*. Mary Ann Liebert, Inc., publishers, 3(3), pp. 278–281. doi: 10.1089/aivt.2017.0012.
- Chalmers, I., Hedges, L. V. and Cooper, H. (2002) 'A brief history of research synthesis', *Evaluation & the health professions*, 25(1), pp. 12–37. doi: 10.1177/0163278702025001003.
- Cheng, S. H. *et al.* (2019) 'A systematic map of evidence on the contribution of forests to poverty alleviation', *Environmental Evidence*, 8(1), p. 221. doi: 10.1186/s13750-019-0148-4 M4 - Citavi.
- Descatha, A. *et al.* (2020) 'The effect of exposure to long working hours on stroke: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury', *Environment international*, 142, p. 105746. doi: 10.1016/j.envint.2020.105746.
- Eldesouky, B. *et al.* (2016) 'Seed, an End-User Text Composition Tool for the Semantic Web: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I', in Groth, P. *et al.* (eds) *The Semantic Web – ISWC 2016*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 218–233. doi: 10.1007/978-3-319-46523-4_14.
- El Idrissi, T., Idri, A. and Bakkoury, Z. (2019) 'Systematic map and review of predictive techniques in diabetes self-management', *International journal of information management*, 46, pp. 263–277. doi: 10.1016/j.ijinfomgt.2018.09.011 M4 - Citavi.
- European Food Safety Authority (2010) 'Application of systematic review methodology to food and feed safety assessments to support decision making', *EFSA Journal*, 8(6), p. 1637. doi: 10.2903/j.efsa.2010.1637.
- Fagerholm, N. *et al.* (2016) 'A systematic map of ecosystem services assessments around European agroforestry', *Ecological indicators*, 62, pp. 47–65. doi: 10.1016/j.ecolind.2015.11.016 M4 - Citavi.
- Farrington, D. P. and Ttofi, M. M. (2009) 'School-Based Programs to Reduce Bullying and Victimization', *Campbell Systematic Reviews*, 5(1), pp. i–148. doi: 10.4073/csr.2009.6.

- Fragoso, G. *et al.* (2004) 'Overview and utilization of the NCI thesaurus', *Comparative and functional genomics*, 5(8), pp. 648–654. doi: 10.1002/cfg.445.
- Garg, A. X., Hackam, D. and Tonelli, M. (2008) 'Systematic Review and Meta-analysis: When One Study Is Just not Enough', *Clinical journal of the American Society of Nephrology: CJASN*, 3(1), pp. 253–260. doi: 10.2215/CJN.01430307.
- Gasparri, L. and Marconi, D. (2019) 'Word Meaning', in Edward, N. Z. (ed.) *The Stanford Encyclopedia of Philosophy*. Fall 2019. Metaphysics Research Lab, Stanford University.
- Golden, E. *et al.* (2020) 'Evaluation of the global performance of eight in silico skin sensitization models using human data', *ALTEX*. doi: 10.14573/altex.1911261.
- Guzelian, P. S. *et al.* (2005) 'Evidence-based toxicology: a comprehensive framework for causation', *Human & experimental toxicology*, 24(4), pp. 161–201. doi: 10.1191/0960327105ht5170a.
- Higgins, JPT *et al.* (2019) *Methodological Expectations of Cochrane Intervention Reviews (MECIR)*. Cochrane. Available at: <https://community.cochrane.org/mecir-manual>.
- Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (ed.) (2019) *Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019)*. Cochrane. Available at: www.training.cochrane.org/handbook.
- Hoffmann, S. *et al.* (2017) 'A primer on systematic reviews in toxicology', *Archives of toxicology*, 91(7), pp. 2551–2575. doi: 10.1007/s00204-017-1980-3.
- Hoffmann, S. and Hartung, T. (2006) 'Toward an evidence-based toxicology', *Human & experimental toxicology*, 25(9), pp. 497–513. doi: 10.1191/0960327106het6480a.
- Institute of Medicine, Board on Health Care Services and Committee on Standards for Systematic Reviews of Comparative Effectiveness Research (2011) *Finding What Works in Health Care: Standards for Systematic Reviews*. Washington, D.C.: National Academies Press. doi: 10.17226/13059.
- Ives, C. *et al.* (2017) 'Creating a Structured AOP Knowledgebase via Ontology-Based Annotations', *Appl In Vitro Toxicol*, 3(4), pp. 298–311. doi: 10.1089/aivt.2017.0017.
- James, K. L., Randall, N. P. and Haddaway, N. R. (2016) 'A methodology for systematic mapping in environmental sciences', *Environmental Evidence*, 5(1), p. 7. doi: 10.1186/s13750-016-0059-6.
- Kaplan, A. (1973) *The conduct of inquiry*. Transaction Publishers. Available at: <https://play.google.com/store/books/details?id=ks8wuzH5Ks8C>.
- Keshava, C. *et al.* (2020) 'Application of systematic evidence mapping to assess the impact of new research when updating health reference values: A case example using acrolein', *Environment international*, 143, p. 105956. doi: 10.1016/j.envint.2020.105956.
- Luechtefeld, T. *et al.* (2018) 'Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility', *Toxicological sciences: an official journal of the Society of Toxicology*, (July). doi: 10.1093/toxsci/kfy152.
- Marshall, I. J. and Wallace, B. C. (2019) 'Toward systematic review automation: a practical guide to using machine learning tools in research synthesis', *Systematic reviews*. Systematic Reviews, 8(1), p. 163. doi: 10.1186/s13643-019-1074-9.

- Miake-Lye, I. M. *et al.* (2016) 'What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products', *Systematic reviews*, 5, p. 28. doi: 10.1186/s13643-016-0204-x.
- Morgan, R. L. *et al.* (2018) 'Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes', *Environment international*. Elsevier, (July), pp. 1–5. doi: 10.1016/j.envint.2018.07.015.
- National Research Council (2007) *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: The National Academies Press, p. 216. doi: 10.17226/11970.
- Nosek, B. A. *et al.* (2018) 'The preregistration revolution', *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), pp. 2600–2606. doi: 10.1073/pnas.1708274114.
- O'Connor, A. M. *et al.* (2020) 'A focus on cross-purpose tools, automated recognition of study design in multiple disciplines, and evaluation of automation tools: a summary of significant discussions at the fourth meeting of the International Collaboration for Automation of Systematic Reviews (ICASR)', *Systematic reviews*, 9(1), p. 100. doi: 10.1186/s13643-020-01351-4.
- OECD (2016) 'Users' Handbook supplement to the Guidance Document for developing and assessing Adverse Outcome Pathways', *Env/Jm/Mono(2016) 12*, (OECD Series on Adverse Outcome Pathways1), p. 63. doi: 10.1787/5jlv1m9d1g32-en.
- Oldman, D. and Tanase, D. (2018) 'Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II', in Vrandečić, D. *et al.* (eds) *The Semantic Web – ISWC 2018*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 325–340. doi: 10.1007/978-3-030-00668-6_20.
- Oliveira, E. C. *et al.* (2017) 'Ontology-Based CMS News Authoring Environment', in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pp. 264–265. doi: 10.1109/ICSC.2017.91.
- Orellano, P. *et al.* (2020) 'Short-term exposure to particulate matter (PM10 and PM2.5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis', *Environment international*, 142, p. 105876. doi: 10.1016/j.envint.2020.105876.
- Pelch, K. E. *et al.* (2019) 'PFAS health effects database: Protocol for a systematic evidence map', *Environment international*, 130, p. 104851. doi: 10.1016/j.envint.2019.05.045.
- Poux, S. and Gaudet, P. (2017) 'Best Practices in Manual Annotation with the Gene Ontology', *Methods in molecular biology*, 1446, pp. 41–54. doi: 10.1007/978-1-4939-3743-1_4.
- Rethlefsen, M. L. *et al.* (2015) 'Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews', *Journal of clinical epidemiology*, 68(6 PG - 617-626), pp. 617–626. doi: 10.1016/j.jclinepi.2014.11.025.
- Roberts, D. *et al.* (2017) 'Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth', *Cochrane database of systematic reviews*, (3). doi: 10.1002/14651858.CD004454.pub3.
- Rooney, A. *et al.* (2014) 'Systematic Review and Evidence Integration for Literature-Based Environmental Health Science Assessments', *Environmental health perspectives*, 122(7), pp. 711–718. doi: 10.1289/ehp.1307972.

- Sanchez, K. A. *et al.* (2020) 'Urban policy interventions to reduce traffic emissions and traffic-related air pollution: Protocol for a systematic evidence map', *Environment international*, 142, p. 105826. doi: 10.1016/j.envint.2020.105826.
- Saran, A. and White, H. (2018) 'Evidence and gap maps: a comparison of different approaches', *Campbell Systematic Reviews*, 14(1), pp. 1–38. doi: 10.4073/cmdp.2018.2.
- Schaefer, H. R. and Myers, J. L. (2017) 'Guidelines for performing systematic reviews in the development of toxicity factors', *Regulatory toxicology and pharmacology: RTP*, 91, pp. 124–141. doi: 10.1016/j.yrtph.2017.10.008.
- Schwarzman, M. R. *et al.* (2015) 'Screening for Chemical Contributions to Breast Cancer Risk: A Case Study for Chemical Safety Evaluation', *Environmental health perspectives*, 123(12), pp. 1255–1264. doi: 10.1289/ehp.1408337.
- Sorensen, R. (2018) 'Vagueness', in Edward, N. Z. (ed.) *The Stanford Encyclopedia of Philosophy*. Summer 2018. Metaphysics Research Lab, Stanford University.
- Stearns, M. Q. *et al.* (2001) 'SNOMED clinical terms: overview of the development process and project status', *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 662–666.
- Thomas, P. D. *et al.* (2019) 'Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems', *Nature genetics*, 51(10), pp. 1429–1433. doi: 10.1038/s41588-019-0500-1.
- Villeneuve, D. L. *et al.* (2014) 'Adverse outcome pathway (AOP) development I: strategies and principles', *Toxicological sciences: an official journal of the Society of Toxicology*, 142(2), pp. 312–320.
- Walker, V. R. *et al.* (2018) 'Human and animal evidence of potential transgenerational inheritance of health effects: An evidence map and state-of-the-science evaluation', *Environment international*, 115, pp. 48–69. doi: 10.1016/j.envint.2017.12.032.
- Wang, R.-L. (2020) 'Semantic characterization of adverse outcome pathways', *Aquatic toxicology*, 222, p. 105478. doi: 10.1016/j.aquatox.2020.105478.
- Wang, R.-L., Edwards, S. and Ives, C. (2019) 'Ontology-based semantic mapping of chemical toxicities', *Toxicology*, 412, pp. 89–100. doi: 10.1016/j.tox.2018.11.005.
- Watford, S. *et al.* (2019) 'Progress in data interoperability to support computational toxicology and chemical safety evaluation', *Toxicology and applied pharmacology*, 380, p. 114707. doi: 10.1016/j.taap.2019.114707.
- Whaley, P. *et al.* (2016) 'Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations', *Environment international*, 92-93, pp. 556–564. doi: 10.1016/j.envint.2015.11.002.
- Whaley, P. *et al.* (2020) 'Recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER)', *Environment international*, 143, p. 105926. doi: 10.1016/j.envint.2020.105926.
- Whetzel, P. L. *et al.* (2011) 'BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications', *Nucleic acids research*, 39(Web Server issue), pp. W541–5. doi: 10.1093/nar/gkr469.

- Wilkinson, M. D. *et al.* (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3, p. 160018. doi: 10.1038/sdata.2016.18.
- Williams, A. J. *et al.* (2017) 'The CompTox Chemistry Dashboard: a community data resource for environmental chemistry', *Journal of cheminformatics*, 9(1), pp. 61–61. doi: 10.1186/s13321-017-0247-6.
- Wittwehr, C. *et al.* (2020) 'Artificial Intelligence for chemical risk assessment', *Computational Toxicology*, 13, p. 100114. doi: 10.1016/j.comtox.2019.100114.
- Wolffe, T. A. M. *et al.* (2019) 'Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management', *Environment international*, 130, p. 104871. doi: 10.1016/j.envint.2019.05.065.
- Wolffe, T. A. M. *et al.* (2020) 'A Survey of Systematic Evidence Mapping Practice and the Case for Knowledge Graphs in Environmental Health and Toxicology', *Toxicological sciences: an official journal of the Society of Toxicology*, 175(1), pp. 35–49. doi: 10.1093/toxsci/kfaa025.
- Woodruff, T. J. and Sutton, P. (2014) 'The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes', *Environmental health perspectives*, 122(10), pp. 1007–1014. doi: 10.1289/ehp.1307175.
- Woodruff, T. J., Sutton, P. and The Navigation Guide Work Group (2011) 'An Evidence-Based Medicine Methodology To Bridge The Gap Between Clinical And Environmental Health Sciences', *Health affairs*, 30(5), pp. 931–937. doi: 10.1377/hlthaff.2010.1219.
- Yost, E. E. *et al.* (2019) 'Hazards of diisobutyl phthalate (DIBP) exposure: A systematic review of animal toxicology studies', *Environment international*, 125, pp. 579–594. doi: 10.1016/j.envint.2018.09.038.

Chapter 5.

Conclusions and Future Work

Conclusions

Improving the quality of systematic reviews

Chapter 1 concluded that systematic review methods “have yet to make widespread impact on the process of chemical risk assessment” and identified several challenges to implementing systematic methods in chemical risk assessment which would need to be overcome if its potential is to be realised. These included the need for technical methodological work to improve the validity and utility of the outputs of systematic reviews, for tools which would reduce the amount of effort and resource required to conduct systematic reviews, and for clear standards for good conduct to help address the issue of the suspect quality of many of the environmental health systematic reviews being published at the time. Chapters 2, 3 and 4 deliver some of that work.

What was not anticipated during the writing of Chapter 1 was the sudden acceleration in uptake of systematic methods (or at least, attempts at such) that would be seen after 2016: by the end of 2019, roughly as many systematic reviews had been published since the writing of Chapter 1 as had ever been published before it (see Chapter 2, Figure 1). This suggests that some of the practical barriers to uptake of systematic methods, in particular the resources required for their conduct, were perhaps not as important as the authors had expected. On the other hand, the explosion in publication of environmental health systematic reviews accentuates other challenges identified in Chapter 1, in particular the need for clear guidance on good practices in the conduct of systematic reviews.

Chapter 2 represents a response to that need, establishing the consensus view of a representative selection of stakeholders as to a set of recommended practices in the conduct of environmental health systematic reviews. Due to only recently being published, it is not yet possible to gauge the research community's reaction to the recommendations or its effectiveness as an intervention for improving the quality of published systematic reviews. What has become increasingly clear, however, is that interventions such as the development of reporting standards and recommendations for conduct of systematic reviews are only individual elements of a broader strategy which is needed for improving the quality of environmental health systematic reviews.

The need for a more integrated strategy is suggested by a growing body of evidence that individual quality control interventions are ineffective when taken in isolation. For example, in spite of widespread endorsement among medical journals of the PRISMA standard for reporting biomedical systematic reviews, there is little evidence that journals which endorse PRISMA publish systematic reviews of higher quality than journals which do not (Stevens et al., 2014). Overall, publishing standards for systematic reviews have remained largely unchanged in spite of the widespread introduction of reporting standards and attempts by journals to implement processes which are expected to raise the quality of the systematic reviews they are publishing (Page et al., 2016).

The importance of the interplay between conduct standards, reporting standards, and critical appraisal tools in improving the quality of published systematic reviews was initially underestimated in Chapter 1. It was first outlined in the editorial for the Special Issue in which Chapter 1 was published (see Appendix A) and referenced in earlier versions of Chapter 2 before the manuscript was simplified in response to peer-review comments (Whaley et al., 2019). This interplay is shown in Figure 1 below. If reporting standards and conduct standards are to be more effective in improving the quality of published systematic reviews, the relationship between conduct standards, reporting standards (which usually only imply certain practices) and the use of critical appraisal tools needs to be further clarified and exploited.

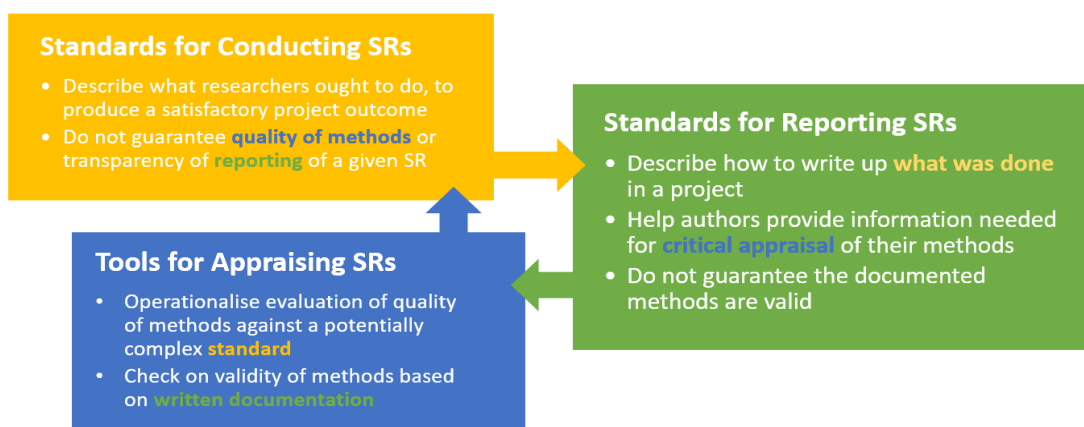


Figure 1. The interplay between conduct standards, reporting standards, and critical appraisal tools in managing the quality of systematic review publications.

New evidence synthesis methods for chemical risk assessment

Chapter 3 responds to the call in Chapter 1 for more work on adapting systematic review methods from biomedicine to the risk assessment context, and the recommendation from Chapter 2 for detailed work specifically on assessing the external validity of evidence. Chapter 3 also further demonstrates the value of interdisciplinary collaboration between research methodologists with a background in public health and biomedicine, and researchers working in toxicology, risk assessment and environmental health. This allowed both the identification of new methodological approaches for systematic review in chemical risk assessment, and also the feeding back of those new approaches into potential methodological innovations for biomedical and public health systematic reviews.

In Chapter 3 we were able to resolve the controversy about the value of “biological plausibility” in making causal inferences in public and environmental health. We did this by engaging in the long-running debate about how best to interpret Austin Bradford Hill’s intuition that while the availability of a biological explanation for association between an exposure and an outcome seems to be helpful in determining whether the relationship is causal, it does not seem to be necessary to have such knowledge to make such a determination. Through examining examples drawn almost exclusively from chemical risk assessment, we were able to tease out the role which biological knowledge has in informing researchers’ confidence in whether an exposure-outcome association is causal or not.

Because the role which biological knowledge has in judging the causality of associations in risk assessment turns out to be equivalent to assessing external validity in systematic reviews, we were then able to show not only how the systematic review process already accommodates the processes being followed by risk assessors, but therefore also how the processes being followed by risk assessors can inform the operationalisation of external validity in systematic reviews. This potentially closes one of the gaps between systematic review methods and risk assessment without needing radical change to the risk assessment process, but instead simply careful operationalisation of assessment of external validity which can serve both the risk assessment and public and environmental health communities equally.

The need to automate evidence synthesis

Chapter 3 also illustrates the challenge presented by the sheer volume of evidence which, although only indirectly related to the research question being asked, needs to be accounted for in a systematic review in order for the review to provide sufficiently certain estimates of health risks posed by exposure to chemical substances.

The conventional approach in systematic review, as inherited from its origin in biomedical and public health research, is simply to disregard indirect evidence: most systematic reviews are designed around tightly-focused questions which include only the most directly informative evidence (path A of figure 2 in Chapter 3). This is generally a viable strategy in healthcare and public health reviews, as there is usually a sufficiently substantial body of human evidence that the results of a systematic review can be usefully informative of a policy decision. Unfortunately, this is a much less viable strategy in environmental health research and chemical risk assessment, where there is usually very little evidence of direct relevance to a systematic review question.

The challenge is, once one starts broadening the eligibility criteria of a systematic review, the volume of evidence which has to be handled increases exponentially. For example, one might wish to include surrogate exposures in a systematic review and therefore extend the eligibility criteria of the review to chemicals which are structurally one or two steps removed from the substance of concern. This might increase the number of eligible exposures to ten or twenty substances. If each chemical has 10-20 studies associated with it, the number of studies to be included might increase from a

handful to hundreds. The same is true for including animal studies for surrogate populations. Given the number of potentially informative surrogate outcomes being studied using *in vitro* models, the amount of evidence that may need to be handled could end up being vast.

There are two obstacles to the incorporation of indirect evidence in systematic reviews discussed in this Thesis. As summarised in Table 3 of Chapter 3, relationships between surrogates are determined by features such as similarity of biological pathways in populations, relative affinity of molecules for points at which a substance interacts with relevant biological processes in an organism, and the predictivity of surrogate outcomes for outcomes of concern, among others. The problem, as discussed in Chapter 4, is that knowledge of how different types of surrogate are related to each other is not captured in existing research databases; as such, indirect evidence is very difficult to consistently and reliably retrieve. The second obstacle is simply the sheer volume of the evidence which needs to be synthesised: with nearly one million citations now being added to MEDLINE every year (National Library of Medicine, 2020), there is significantly more research being conducted than can realistically be manually summarised.

Both of these obstacles are overcome with the same solution: the automatic population of large evidence databases with data from scientific studies. These are the Knowledge Organisation Systems of Chapter 4. In the course of developing this Thesis I have come to the conclusion that Knowledge Organisation Systems are the natural next step in the evolution of systematic review. When combined with Artificial Intelligence techniques for summarising and synthesising research, they stand to radically change the way in which evidence synthesis is conducted.

A radically different future

I would personally speculate that once Knowledge Organisation Systems of reasonable scale and power have been implemented, the character of systematic review and evidence synthesis will undergo radical change. The steps of systematic review, of setting inclusion criteria for studies based on narrowly-defined PECO statements, of searching databases using sensitive keyword term combinations to try and achieve conceptual coverage of the topics of interest to the review, of manually screening studies for relevance and extracting relevant data for synthesis: these steps are all

determined by the need to do good research while working within the constraints imposed by small groups of people with finite recall manually analysing data. The problem is, these constraints mean we are only exploiting a fraction of the vast wealth of scientific knowledge we are generating every year.

These constraints disappear when we replace the reading of PDFs with the databases of Knowledge Organisation Systems. The vast wealth of human knowledge is no longer stored in individual, separate documents which have to be read in order for the information within them to be made accessible to the research team; instead, the knowledge encoded within them is represented directly in large-scale semantic databases. Evidence synthesis stops being about individual researchers making sense of how a small handful of studies fit together and becomes about the querying of the Knowledge Organisation System, using semantic reasoners and big data techniques to interpret how the range of information around the concepts of interest answers the questions we are asking. The limits become computational rather than merely practical, and the methods for research synthesis will change accordingly.

Future Work: “Research Without Reading”

I would identify three broad research themes which could be developed to facilitate the transition from where we are now, whereby evidence synthesis is a small-scale, manual activity which exploits only a fraction of available knowledge, to scientific research being represented in large-scale Knowledge Organisation Systems.

Standards for complete, accurate and machine-readable research

Systematic reviews have repeatedly demonstrated that primary research is very often both poorly conducted and incompletely reported, routinely overlooking practices such as the blinding of investigators and randomisation of subjects to the exposure and control arms of a study, and failing to adequately describe the methods used in sufficient detail to allow the credibility of the study’s results to be assessed (de Vries et al., 2014; Ritskes-Hoitinga et al., 2014). Systematic reviews have value even if they are only able to identify where a body of evidence has collectively uncertain results; however, they would have even more value if the evidence being analysed was of consistently higher quality.

Building on Chapters 1, 2 and 4, there is a need for development of more effective standards for conduct and reporting of research which is also machine readable. General improvement in the quality of conduct and documentation of research would raise the standard of the stock material for evidence syntheses, because reporting would be more complete and the results of studies would be of higher validity. Making research machine readable (meaning that data about study methods and results can be piped directly into Knowledge Organisation Systems instead of being presented in an isolated PDF) by tagging study reports with metadata including the ontological classes of Chapter 4 would help remove the bottleneck of manual reading which prevents the implementation of large-scale databases of scientific knowledge.

As an example of how this might work, I have been using the [Protocols.io](https://protocols.io) platform to prototype systematic review protocol templates which close the gap between standards for reporting and conducting research. I am doing this by interpreting COSTER from Chapter 2 into an explicit, step-by-step “recipe” for conduct of a systematic review which can be followed by a researcher (see Appendix B). Because the “recipe” prompts the researcher to describe how they fulfilled each step, and is directly derived from a comprehensive set of good practice recommendations, the result of following the protocol template should be complete documentation of each important step of a systematic review. It should also result in more valid results overall, because the scientist is prompted to follow recommended practices they might otherwise have overlooked. Finally, because each step is essentially an object which can be named and given various attributes, this approach becomes the first step in making a research report directly machine readable.

The database technology for Knowledge Organisation Systems

Suitable database technology which could underpin large Knowledge Organisation Systems still needs to be developed and implemented. In itself, the value of databases summarising the methods and findings of environmental health studies is nothing new and already well recognised. The Health Assessment Workspace Collaborative (HAWC) (<https://hawcproject.org/>) is arguably the first platform which has been purpose-built for supporting health assessments. However, as a relational database it struggles with accommodating new study designs and can be expected to become computationally inefficient once the number of records it contains exceeds a certain

threshold. Relational databases also find it notoriously difficult to cope with unstructured, semantic data such as textual information about study methods (Robinson et al., 2013).

Instead of relational databases, it would make sense to explore how graph databases can be used to represent the knowledge which is codified in scientific documents. Appendices C and D show some of my initial work on exploring evidence mapping methods and posits how graph databases, with their “on-read” rather than “on-write” schema, are better suited to the challenges of representing scientific knowledge in a database and making it readily accessible to users. The ontologies of Chapter 4 provide an interpretive layer to the database, to make research about the concepts in the database readily accessible to the user rather than, as currently has to be done, the user having to manually retrieve information for themselves which is buried in PDFs of manuscripts. A larger-scale exploration of how graph databases can be used for storing environmental health knowledge should be conducted, in particular as it relates to functioning as a repository for the outputs of AI-driven automated data extraction and machine-readable study reports.

Machine-compatible evidence analysis tools

Chapter 3 anticipates an external validity instrument for systematic review. Initially the tool will be designed for use by people; I would speculate that it will involve the assessment of the biological similarity of observed experimental PECO to the target PECO of the systematic review which is being conducted. However, the analysis will be complex and increasingly information-heavy, and therefore likely to only be conducted in a simplistic way when done by people. This intuition is reinforced by how complex it is to collate and analyse the evidence which is needed for mapping biological processes in Adverse Outcome Pathways (as discussed in Chapter 4). Nonetheless, because detailed biological information is needed in order for judgements of external validity to be properly grounded, it seems inevitable that computational processes will be required for its identification and interpretation.

A further complicating factor is that, if computational methods for analysing evidence are to be acceptable from a regulatory perspective, it seems likely that the processes for analysing the evidence will need to be in principle human-understandable: black box

processes are probably not an option, at least in the medium-term. In order for computational approaches to evidence analysis to perform as well as, then better than, people, in a way which is nonetheless understandable to people, requires human-level reasoning processes to produce outputs which are interpretable by machines (i.e. processes which use human concepts in a way which can be described in numbers).

A potential solution to this is the development of a tool for interpreting the external validity of a study included in a systematic review in terms of its distance in n dimensions from a fixed point of origin in information space defined by the PECO statement of the systematic review (see Figure 2). Initial distances in that space can be established by asking domain experts to describe numerically (such as by using a Likert scale) the extent to which they consider the PECO elements of various studies to be similar to each other. A proposal for how this might work is outlined in Appendix E. If the information about how experts are making judgements can be enriched with data from a Knowledge Organisation System (Chapter 4), it should be possible for machines to make the same type of calculations as humans but using much more data than people can realistically process. Hopefully, this would be a sufficiently white-box process that it can be used for predicting health risks in a fashion acceptable to regulators, while exploiting the vast increase in information-processing capacity granted by the use of computational approaches.

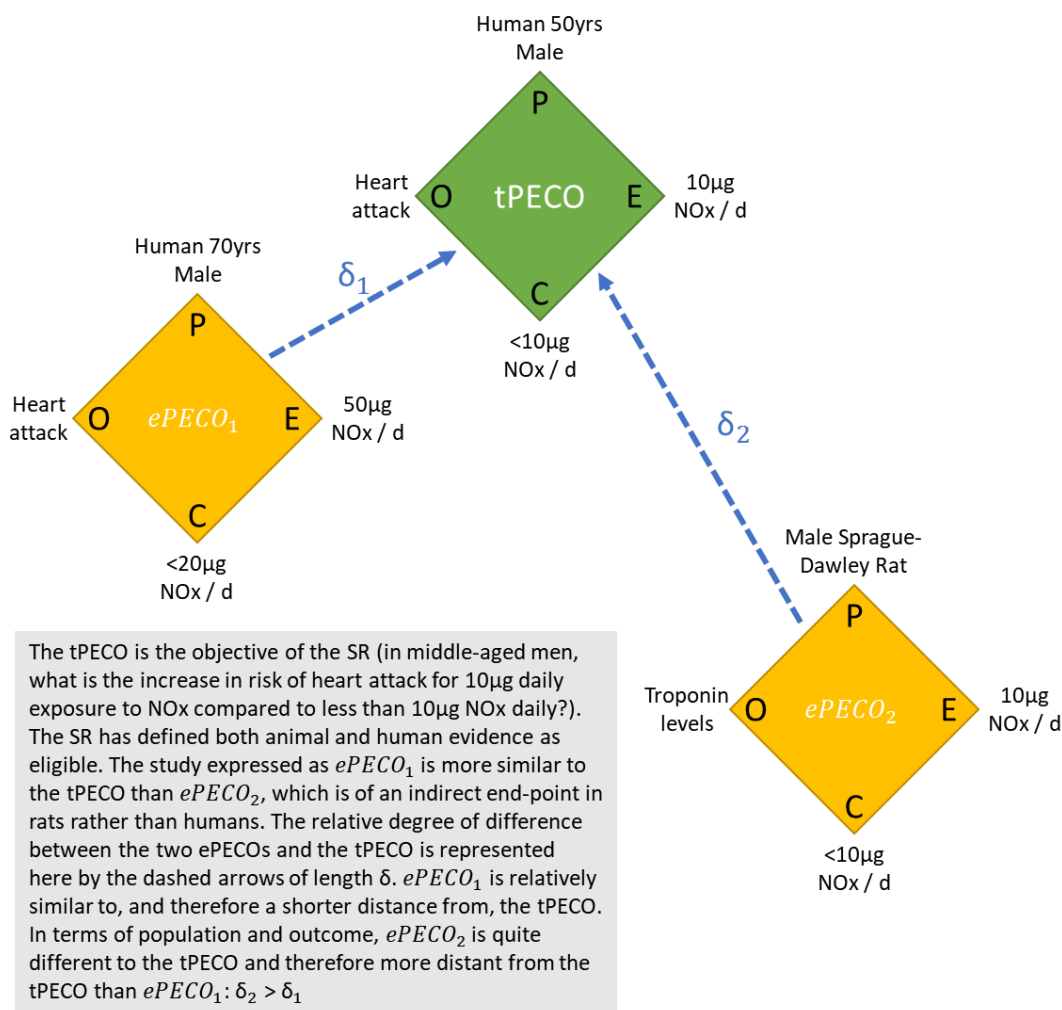


Figure 2. The beginnings of an approach to the mathematical description of external validity of studies included in a systematic review.

References

- de Vries, R.B.M., Wever, K.E., Avey, M.T., Stephens, M.L., Sena, E.S., Leenaars, M. (2014) The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *ILAR J.* 55, 427–437.
- National Library of Medicine. (2020) Citations Added to MEDLINE by Fiscal Year [WWW Document]. URL https://www.nlm.nih.gov/bsd/stats/cit_added.html (accessed 9.9.19).
- Page, M.J., Shamseer, L., Altman, D.G., Tetzlaff, J., Sampson, M., Tricco, A.C., Catalá-López, F., Li, L., Reid, E.K., Sarkis-Onofre, R., Moher, D. (2016) Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. *PLoS Med.* 13, e1002028.

- Ritskes-Hoitinga, M., Leenaars, M., Avey, M., Rovers, M., Scholten, R. (2014) Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. *Cochrane Database Syst. Rev.* ED000078.
- Robinson, I., Webber, J., Webber, J., Eifrem, E. (2013) *Graph Databases*. United States: O'Reilly Media.
- Stevens, A., Shamseer, L., Weinstein, E., Yazdi, F., Turner, L., Thielman, J., Altman, D.G., Hirst, A., Hoey, J., Palepu, A., Schulz, K.F., Moher, D. (2014) Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *BMJ* 348, g3804.
- Whaley, P., Aiassa, E., Beausoleil, C., Beronius, A., Bilotta, G., Boobis, A., Vries, R., Hanberg, A., Hoffmann, S., Hunt, N., Kwiatkowski, C., Lam, J., Lipworth, S., Martin, O., Randall, N., Rhomberg, L., Rooney, A.A., Schünemann, H.J., Wikoff, D., Wolffe, T., Halsall, C. (2019). A code of practice for the conduct of systematic reviews in toxicology and environmental health research (COSTER). [preprint]

Consolidated Bibliography

- Adams, J., Hillier-Brown, F.C., Moore, H.J., Lake, A.A., Araujo-Soares, V. and White, M. *et al.* (2016) Searching and synthesising ‘grey literature’ and ‘grey information’ in public health: critical reflections on three case studies. *Systematic Reviews*, 5(1), 979. Available at: doi:10.1186/s13643-016-0337-y.
- Ågerstrand, M. and Beronius, A. (2016) Weight of evidence evaluation and systematic review in EU chemical risk assessment: Foundation is laid but guidance is needed. *Environment International*, 92-93, 590–596. Available at: doi:10.1016/j.envint.2015.10.008.
- Aiassa, E., Higgins, J.P.T., Frampton, G.K., Greiner, M., Afonso, A. and Amzal, B. *et al.* (2015) Applicability and feasibility of systematic review for performing evidence-based risk assessment in food and feed safety. *Critical Reviews in Food Science and Nutrition*, 55(7), 1026–1034. Available at: doi:10.1080/10408398.2013.769933.
- Al-Shahi Salman, R., Beller, E., Kagan, J., Hemminki, E., Phillips, R.S. and Savulescu, J. *et al.* (2014) Increasing value and reducing waste in biomedical research regulation and management. *Lancet (London, England)*, 383(9912), 176–185. Available at: doi:10.1016/S0140-6736(13)62297-7.
- Antman, E.M. (1992) A Comparison of Results of Meta-analyses of Randomized Control Trials and Recommendations of Clinical Experts. *JAMA*, 268(2), 240. Available at: doi:10.1001/jama.1992.03490020088036.
- Arzuaga, X., Smith, M.T., Gibbons, C.F., Skakkebak, N.E., Yost, E.E. and Beverly, B.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. *Environmental Health Perspectives*, 127(6), 65001. Available at: doi:10.1289/EHP5045.
- Arzuaga, X., Walker, T., Yost, E.E., Radke, E.G. and Hotchkiss, A.K. (2019) Use of the Adverse Outcome Pathway (AOP) framework to evaluate species concordance and human relevance of Dibutyl phthalate (DBP)-induced male reproductive toxicity. *Reproductive Toxicology*. Available at: doi:10.1016/j.reprotox.2019.06.009.

- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H. and Cherry, J.M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), 25–29. Available at: doi:10.1038/75556.
- Atkins, D., Best, D., Briss, P.A., Eccles, M., Falck-Ytter, Y. and Flottorp, S. *et al.* (2004) Grading quality of evidence and strength of recommendations. *BMJ*, 328(7454), 1490. Available at: doi:10.1136/bmj.328.7454.1490.
- Battaglia, M. and Atkinson, M.A. (2015) The Streetlight Effect in Type 1 Diabetes. *Diabetes*, 64(4), 1081–1090. Available at: doi:10.2337/db14-1208.
- Beronius, A. and Vandenberg, L.N. (2015) Using systematic reviews for hazard and risk assessment of endocrine disrupting chemicals. *Reviews in Endocrine and Metabolic Disorders*, 16(4), 273–287. Available at: doi:10.1007/s11154-016-9334-7.
- Beronius, A., Hanberg, A., Zilliacus, J. and Rudén, C. (2014) Bridging the gap between academic research and regulatory health risk assessment of Endocrine Disrupting Chemicals. *Current Opinion in Pharmacology*, 19, 99–104. Available at: doi:10.1016/j.coph.2014.08.005.
- Beronius, A., Molander, L., Rudén, C. and Hanberg, A. (2014) Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: a proposal to improve evaluation criteria and reporting. *Journal of Applied Toxicology : JAT*, 34(6), 607–617. Available at: doi:10.1002/jat.2991.
- Bilotta, G.S., Milner, A.M. and Boyd, I. (2014) On the use of systematic reviews to inform environmental policies. *Environmental Science & Policy*, 42, 67–77. Available at: doi:10.1016/j.envsci.2014.05.010.
- Bilotta, G.S., Milner, A.M. and Boyd, I.L. (2014) Quality assessment tools for evidence from environmental science. *Environmental Evidence*, 3(1), 14. Available at: doi:10.1186/2047-2382-3-14.
- Birnbaum, L.S., Thayer, K.A., Bucher, J.R. and Wolfe, M.S. (2013) Implementing Systematic Review at the National Toxicology Program: Status and Next Steps. *Environmental Health Perspectives*, 121(4). Available at: doi:10.1289/ehp.1306711.
- Borah, R., Brown, A.W., Capers, P.L. and Kaiser, K.A. (2017) Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*, 7(2), e012545. Available at: doi:10.1136/bmjopen-2016-012545.
- Braga, A., Papachristos, A. and Hureau, D. (2012) Hot spots policing effects on crime. *Campbell Systematic Reviews*, 8(1), 1–96. Available at: doi:10.4073/csr.2012.8.
- Braun, J.M. and Gray, K. (2017) Challenges to studying the health effects of early life environmental chemical exposures on children’s health. *PLOS Biology*, 15(12), e2002800. Available at: doi:10.1371/journal.pbio.2002800.
- Burgoon, L.D. (2017) The AOPontology: A Semantic Artificial Intelligence Tool for Predictive Toxicology. *Applied In Vitro Toxicology*, 3(3), 278–281. Available at: doi:10.1089/aivt.2017.0012.

- Burns, J., Boogaard, H., Polus, S., Pfadenhauer, L.M., Rohwer, A.C. and van Erp, A.M. *et al.* (2020) Interventions to reduce ambient air pollution and their effects on health: An abridged Cochrane systematic review. *Environment international*, 135, 105400. Available at: doi:10.1016/j.envint.2019.105400.
- Caveman (2000) The invited review - or, my field, from my standpoint, written by me using only my data and my ideas, and citing only my publications. *Journal of Cell Science*, 113(Pt 18), 3125–3126.
- Chalmers, I. (2010) Systematic reviews and uncertainties about the effects of treatments. *The Cochrane Database of Systematic Reviews*, 2011, ED000004. Available at: doi:10.1002/14651858.ED000004.
- Chalmers, I. and Glasziou, P. (2009) Avoidable waste in the production and reporting of research evidence. *Obstetrics and Gynecology*, 114(6), 1341–1345. Available at: doi:10.1097/AOG.0b013e3181c3020d.
- Chalmers, I., Hedges, L.V. and Cooper, H. (2002) A Brief History of Research Synthesis. *Evaluation & the Health Professions*, 25(1), 12–37. Available at: doi:10.1177/0163278702025001003.
- Chambers, C. (2019) What's next for Registered Reports? *Nature*, 573(7773), 187–189. Available at: doi:10.1038/d41586-019-02674-6.
- Cheng, S.H., MacLeod, K., Ahlroth, S., Onder, S., Perge, E. and Shyamsundar, P. *et al.* (2019) A systematic map of evidence on the contribution of forests to poverty alleviation. *Environmental Evidence*, 8(1). Available at: doi:10.1186/s13750-019-0148-4.
- Clarke, M., Brice, A., Chalmers, I. and Gluud, L.L. (2014) Accumulating Research: A Systematic Account of How Cumulative Meta-Analyses Would Have Provided Knowledge, Improved Health, Reduced Harm and Saved Resources. *PLoS ONE*, 9(7), e102670. Available at: doi:10.1371/journal.pone.0102670.
- Cohen Hubal, E.A., Frank, J.J., Nachman, R., Angrish, M., Deziel, N.C. and Fry, M. *et al.* (2020) Advancing systematic-review methodology in exposure science for environmental health decision making. *Journal of Exposure Science & Environmental Epidemiology*. Available at: doi:10.1038/s41370-020-0236-0.
- Collins, A., Miller, J., Coughlin, D. and Kirk, S. (2015) *The production of quick scoping reviews and rapid evidence assessments: a how to guide*. Available at: doi:10.1007/0-387-28098-7_14.
- Cousins, I.T., DeWitt, J.C., Glüge, J., Goldenman, G., Herzke, D. and Lohmann, R. *et al.* (2020) Strategies for grouping per- and polyfluoroalkyl substances (PFAS) to protect human and environmental health. *Environmental Science: Processes & Impacts*, 22(7), 1444–1460. Available at: doi:10.1039/D0EM00147C.
- Crump, K.S. (1996) The linearized multistage model and the future of quantitative risk assessment. *Human & Experimental Toxicology*, 15(10), 787–798. Available at: doi:10.1177/096032719601501001.
- Descatha, A., Sembajwe, G., Baer, M., Boccuni, F., Di Tecco, C. and Duret, C. *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. *Environment International*, 119, 366–378. Available at: doi:10.1016/j.envint.2018.06.016.
- Descatha, A., Sembajwe, G., Pega, F., Ujita, Y., Baer, M. and Boccuni, F. *et al.* (2020) The effect of exposure to long working hours on stroke: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. *Environment international*, 142, 105746. Available at: doi:10.1016/j.envint.2020.105746.
- Dessimoz, C. and Škunca, N. (Eds.) (2017) *The Gene Ontology Handbook*. Springer New York: New York, NY.
- Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2005) Mortality from cancer in relation to smoking: 50 years observations on British doctors. *British Journal of Cancer*, 92(3), 426–429. Available at: doi:10.1038/sj.bjc.6602359.
- Eden, J., Levit, L., Berg, A. and Morton, S. (2011) *Finding What Works in Health Care*. National Academies Press: Washington, D.C.
- EL Idrissi, T., Idri, A. and Bakkoury, Z. (2019) Systematic map and review of predictive techniques in diabetes self-management. *International Journal of Information Management*, 46, 263–277. Available at: doi:10.1016/j.ijinfomgt.2018.09.011.
- Eldesouky, B., Bakry, M., Maus, H. and Dengel, A. (2016) Seed, an End-User Text Composition Tool for the Semantic Web. In: Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F. and Flöck, F. *et al.* (Eds.) *The Semantic Web – ISWC 2016*. Springer International Publishing: Cham, pp. 218–233.
- European Food Safety Authority EFSA (2010) Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA Journal*, 8(6), 1637. Available at: doi:10.2903/j.efsa.2010.1637.
- European Food Safety Authority EFSA (2010) Scientific Opinion on a Quantitative Microbiological Risk Assessment of Salmonella in slaughter and breeder pigs. *EFSA Journal*, 8(4), 1547. Available at: doi:10.2903/j.efsa.2010.1547.
- European Food Safety Authority EFSA (2015) Principles and process for dealing with data and evidence in scientific assessments. *EFSA Journal*, 13(6). Available at: doi:10.2903/j.efsa.2015.4121.
- European Food Safety Authority EFSA (2015) Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs. *EFSA Journal*, 13(1), 3978. Available at: doi:10.2903/j.efsa.2015.3978.
- European Food Safety Authority EFSA (2018) EFSA Scientific Colloquium 23 – Joint European Food Safety Authority and Evidence-Based Toxicology Collaboration Colloquium Evidence integration in risk assessment: the science of combining apples and oranges 25–26 October 2017 Lisbon, Portugal. *EFSA Supporting Publications*, 15(3). Available at: doi:10.2903/sp.efsa.2018.EN-1396.

- Fagerholm, N., Torralba, M., Burgess, P.J. and Plieninger, T. (2016) A systematic map of ecosystem services assessments around European agroforestry. *Ecological Indicators*, 62, 47–65. Available at: doi:10.1016/j.ecolind.2015.11.016.
- Farrah, K., Young, K., Tunis, M.C. and Zhao, L. (2019) Risk of bias tools in systematic reviews of health interventions: an analysis of PROSPERO-registered protocols. *Systematic Reviews*, 8(1), 30. Available at: doi:10.1186/s13643-019-1172-8.
- Farrington, D.P. and Ttofi, M.M. (2009) School-Based Programs to Reduce Bullying and Victimization. *Campbell Systematic Reviews*, 5(1). Available at: doi:10.4073/csr.2009.6.
- Fragoso, G., Coronado, S. de, Haber, M., Hartel, F. and Wright, L. (2004) Overview and Utilization of the NCI Thesaurus. *Comparative and Functional Genomics*, 5(8), 648–654. Available at: doi:10.1002/cfg.445.
- Garg, A.X., Hackam, D. and Tonelli, M. (2008) Systematic Review and Meta-analysis: When One Study Is Just not Enough. *Clinical Journal of the American Society of Nephrology*, 3(1), 253–260. Available at: doi:10.2215/CJN.01430307.
- Gasparri, L. and Marconi, D. (2019) ‘Word Meaning’, in Edward, N. Z. (ed.) *The Stanford Encyclopedia of Philosophy*. Fall 2019. Metaphysics Research Lab, Stanford University.
- Gauderat, G., Picard-Hagen, N., Toutain, P.-L., Servien, R., Viguié, C. and Puel, S. *et al.* (2017) Prediction of human prenatal exposure to bisphenol A and bisphenol A glucuronide from an ovine semi-physiological toxicokinetic model. *Scientific Reports*, 7(1). Available at: doi:10.1038/s41598-017-15646-5.
- Golden, E. (2020) Evaluation of the global performance of eight in silico skin sensitization models using human data. *ALTEX*. Available at: doi:10.14573/altex.1911261.
- Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F. and Flöck, F. *et al.* (Eds.) (2016) *The Semantic Web – ISWC 2016*. Springer International Publishing: Cham.
- Guyatt, G.H., Oxman, A.D., Kunz, R., Atkins, D., Brozek, J. and Vist, G. *et al.* (2011) GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology*, 64(4), 395–400. Available at: doi:10.1016/j.jclinepi.2010.09.012.
- Guyatt, G.H., Oxman, A.D., Schünemann, H.J., Tugwell, P. and Knottnerus, A. (2011) GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology*, 64(4), 380–382. Available at: doi:10.1016/j.jclinepi.2010.09.011.
- Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y. and Alonso-Coello, P. *et al.* (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ (Clinical Research Ed.)*, 336(7650), 924–926. Available at: doi:10.1136/bmj.39489.470347.AD.

- Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y. and Alonso-Coello, P. *et al.* (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650), 924–926. Available at: doi:10.1136/bmj.39489.470347.AD.
- Guzelian, P.S., Victoroff, M.S., Halmes, N.C., James, R.C. and Guzelian, C.P. (2005) Evidence-based toxicology: a comprehensive framework for causation. *Human & Experimental Toxicology*, 24(4), 161–201. Available at: doi:10.1191/0960327105ht517oa.
- Haddaway, N.R. and Westgate, M.J. (2019) Predicting the time needed for environmental systematic reviews and systematic maps. *Conservation Biology : the Journal of the Society for Conservation Biology*, 33(2), 434–443. Available at: doi:10.1111/cobi.13231.
- Hansen, M.R.H., Jørs, E., Sandbæk, A., Kolstad, H.A., Schullehner, J. and Schlünssen, V. (2019) Exposure to neuroactive non-organochlorine insecticides, and diabetes mellitus and related metabolic disturbances: Protocol for a systematic review and meta-analysis. *Environment international*, 127, 664–670. Available at: doi:10.1016/j.envint.2019.02.074.
- Hardy, A., Benford, D., Halldorsson, T., Jeger, M.J., Knutsen, H.K. and More, S. *et al.* (2017) Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA Journal*, 15(8). Available at: doi:10.2903/j.efsa.2017.4971.
- Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (ed.) (2019) *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane. Available at: www.training.cochrane.org/handbook.
- Higgins, J.P.T. and Green, S. (2008) *Cochrane handbook for systematic reviews of interventions*. Wiley-Blackwell: Chichester, England, Hoboken, NJ.
- Higgins, J.P.T., Altman, D.G., Gøtzsche, P.C., Jüni, P., Moher, D., Oxman, A.D. and Savovic, J. *et al.* (2011) *The Cochrane Collaboration's tool for assessing risk of bias in randomised trials*.
- Higgins, JPT *et al.* (2019) *Methodological Expectations of Cochrane Intervention Reviews (MECIR)*. Cochrane. Available at: <https://community.cochrane.org/mecir-manual>.
- Hoffmann, S. and Hartung, T. (2006) Toward an evidence-based toxicology. *Human & Experimental Toxicology*, 25(9), 497–513. Available at: doi:10.1191/0960327106het648oa.
- Hooijmans, C.R. SYRCLE's risk of bias tool for animal studies. *BMC medical research methodology*.
- Hooijmans, C.R., Rovers, M., Vries, R.B. de, Leenaars, M. and Ritskes-Hoitinga, M. (2012) An initiative to facilitate well-informed decision-making in laboratory animal research: report of the First International Symposium on Systematic Reviews in Laboratory Animal Science. *Laboratory Animals*, 46(4), 356–357. Available at: doi:10.1258/la.2012.012052.

- Hultcrantz, M., Rind, D., Akl, E.A., Treweek, S., Mustafa, R.A. and Iorio, A. *et al.* (2017) The GRADE Working Group clarifies the construct of certainty of evidence. *Journal of Clinical Epidemiology*, 87, 4–13. Available at: doi:10.1016/j.jclinepi.2017.05.006.
- IARC, 2019. *IARC Monographs on the Identification of Carcinogenic Hazards to Humans: Preamble*.
- Ip, S., Hadar, N., Keefe, S., Parkin, C., Iovin, R. and Balk, E.M. *et al.* (2012) A Web-based archive of systematic review data. *Systematic Reviews*, 1, 15. Available at: doi:10.1186/2046-4053-1-15.
- Ives, C., Campia, I., Wang, R.-L., Wittwehr, C. and Edwards, S. (2017) Creating a Structured AOP Knowledgebase via Ontology-Based Annotations. *Applied in Vitro Toxicology*, 3(4), 298–311. Available at: doi:10.1089/aivt.2017.0017.
- James, K.L., Randall, N.P. and Haddaway, N.R. (2016) A methodology for systematic mapping in environmental sciences. *Environmental Evidence*, 5(1). Available at: doi:10.1186/s13750-016-0059-6.
- Janis, I.L. (1983, 1982) *Groupthink: Psychological studies of policy decisions and fiascoes*, 2nd edn. Houghton Mifflin: Boston.
- Jefferson, T., Jones, M.A., Doshi, P., Del Mar, C.B., Heneghan, C.J. and Hama, R. *et al.* (2012) Neuraminidase inhibitors for preventing and treating influenza in healthy adults and children. *The Cochrane Database of Systematic Reviews*, 1, CD008965. Available at: doi:10.1002/14651858.CD008965.pub3.
- Johnson, P.I., Koustas, E., Vesterinen, H.M., Sutton, P., Atchley, D.S. and Kim, A.N. *et al.* (2016) Application of the Navigation Guide systematic review methodology to the evidence for developmental and reproductive toxicity of triclosan. *Environment International*, 92-93, 716–728. Available at: doi:10.1016/j.envint.2016.03.009.
- Kaplan, A. (2017) *The conduct of inquiry: Methodology for behavioral science*. Routledge: Abingdon, Oxon, New York, NY.
- Kaplan, A. (Ed.) (2017) *The Conduct of Inquiry*. Routledge.
- Keshava, C., Davis, J.A., Stanek, J., Thayer, K.A., Galizia, A. and Keshava, N. *et al.* (2020) Application of systematic evidence mapping to assess the impact of new research when updating health reference values: A case example using acrolein. *Environment international*, 143, 105956. Available at: doi:10.1016/j.envint.2020.105956.
- Krauth, D., Woodruff, T.J. and Bero, L. (2013) Instruments for Assessing Risk of Bias and Other Methodological Criteria of Published Animal Studies: A Systematic Review. *Environmental Health Perspectives*, 121(9), 985–992. Available at: doi:10.1289/ehp.1206389.
- Kretsinger, R.H., Uversky, V.N. and Permyakov, E.A. (Eds.) (2013) *Encyclopedia of Metalloproteins*. Springer New York: New York, NY.

- Lagarde, F., Beausoleil, C., Belcher, S.M., Belzunces, L.P., Emond, C. and Guerbet, M. *et al.* (2015) Non-monotonic dose-response relationships and endocrine disruptors: a qualitative method of assessment. *Environmental Health*, 14(1). Available at: doi:10.1186/1476-069X-14-13.
- Land, M., Wit, C.A. de, Cousins, I.T., Herzke, D., Johansson, J. and Martin, J.W. (2015) What is the effect of phasing out long-chain per- and polyfluoroalkyl substances on the concentrations of perfluoroalkyl acids and their precursors in the environment? A systematic review protocol. *Environmental Evidence*, 4(1), 3. Available at: doi:10.1186/2047-2382-4-3.
- Lau, J., Ioannidis, J.P.A. and Schmid, C.H. (1998) Summing up evidence: one answer is not always enough. *The Lancet*, 351(9096), 123–127. Available at: doi:10.1016/S0140-6736(97)08468-7.
- Lau, J., Rothstein, H.R. and Stewart, G.B. (2013) History and Progress of Meta-analysis. In: *Handbook of Meta-analysis in Ecology and Evolution*. Princeton University Press.
- Lawlor, D.A., Tilling, K. and Davey Smith, G. (2016) Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, 45(6), 1866–1886. Available at: doi:10.1093/ije/dyw314.
- Li, J., Brisson, C., Clays, E., Ferrario, M.M., Ivanov, I.D. and Landsbergis, P. *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. *Environment international*, 119, 558–569. Available at: doi:10.1016/j.envint.2018.06.022.
- Lin, J. (2008) CSF as a Surrogate for Assessing CNS Exposure: An Industrial Perspective. *Current Drug Metabolism*, 9(1), 46–59. Available at: doi:10.2174/138920008783331077.
- Luderer, U., Eskenazi, B., Hauser, R., Korach, K.S., McHale, C.M. and Moran, F. *et al.* (2019) Proposed Key Characteristics of Female Reproductive Toxicants as an Approach for Organizing and Evaluating Mechanistic Data in Hazard Assessment. *Environmental Health Perspectives*, 127(7), 75001. Available at: doi:10.1289/EHP4971.
- Luechtefeld, T., Marsh, D., Rowlands, C. and Hartung, T. (2018) Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility. *Toxicological Sciences*, 165(1), 198–212. Available at: doi:10.1093/toxsci/kfy152.
- Macleod, M.R., Ebrahim, S. and Roberts, I. (2005) Surveying the literature from animal experiments: Systematic review and meta-analysis are important contributions. *BMJ*, 331(7508), 110.3. Available at: doi:10.1136/bmj.331.7508.110-b.
- Mandrioli, D., Schlünssen, V., Ádám, B., Cohen, R.A., Colosio, C. and Chen, W. *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. *Environment international*, 119, 174–185. Available at: doi:10.1016/j.envint.2018.06.005.
- Marshall, I.J. and Wallace, B.C. (2019) Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1). Available at: doi:10.1186/s13643-019-1074-9.
- Matta, K., Ploteau, S., Coumoul, X., Koual, M., Le Bizec, B. and Antignac, J.-P. *et al.* (2019) Associations between exposure to organochlorine chemicals and endometriosis in experimental studies: A systematic review protocol. *Environment international*, 124, 400–407. Available at: doi:10.1016/j.envint.2018.12.063.
- McGhee, D.J.M., Ritchie, C.W., Thompson, P.A., Wright, D.E., Zajicek, J.P. and Counsell, C.E. (2014) A Systematic Review of Biomarkers for Disease Progression in Alzheimer's Disease. *PLoS ONE*, 9(2), e88854. Available at: doi:10.1371/journal.pone.0088854.
- Miake-Lye, I.M., Hempel, S., Shanman, R. and Shekelle, P.G. (2016) What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. *Systematic Reviews*, 5(1). Available at: doi:10.1186/s13643-016-0204-x.
- Mignini, L.E. and Khan, K.S. (2006) Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. *BMC Medical Research Methodology*, 6, 10. Available at: doi:10.1186/1471-2288-6-10.
- Moher, D., Altman, D.G., Schulz, K.F., Simera, I. and Wager, E. (Eds.) (2014) *Guidelines for reporting health research: A user's manual*. John Wiley & Sons, Ltd: Chichester, West Sussex, Hoboken, NJ.
- Morgan, R.L., Beverly, B., Gherzi, D., Schünemann, H.J., Rooney, A.A. and Whaley, P. *et al.* (2019) GRADE guidelines for environmental and occupational health: A new series of articles in Environment International. *Environment international*, 128, 11–12. Available at: doi:10.1016/j.envint.2019.04.016.
- Morgan, R.L., Thayer, K.A., Bero, L., Bruce, N., Falck-Ytter, Y., Gherzi, D. and Guyatt, G. *et al.* (2016) *GRADE: Assessing the quality of evidence in environmental and occupational health*.
- Morgan, R.L., Whaley, P., Thayer, K.A. and Schünemann, H.J. (2018) Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environment international*, 121, 1027–1031. Available at: doi:10.1016/j.envint.2018.07.015.
- Mulrow, C.D. (1987) The medical review article: state of the science. *Annals of Internal Medicine*, 106(3), 485–488. Available at: doi:10.7326/0003-4819-106-3-485.
- National Library of Medicine, 2020. Citations Added to MEDLINE by Fiscal Year [WWW Document]. URL https://www.nlm.nih.gov/bsd/stats/cit_added.html (accessed 9.9.19).
- National Research Council (US) Committee on the Institutional Means for Assessment of Risks to Public Health, 2014. *Risk Assessment in the Federal*

- Government: Managing the Process*. National Academies Press (US), Washington (DC).
- Nosek, B.A., Ebersole, C.R., DeHaven, A.C. and Mellor, D.T. (2018) The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. Available at: doi:10.1073/pnas.1708274114.
- O'Connor, A.M., Glasziou, P., Taylor, M., Thomas, J., Spijker, R. and Wolfe, M.S. (2020) A focus on cross-purpose tools, automated recognition of study design in multiple disciplines, and evaluation of automation tools: a summary of significant discussions at the fourth meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Systematic Reviews*, 9(1). Available at: doi:10.1186/s13643-020-01351-4.
- OECD (2016) *Users' Handbook supplement to the Guidance Document for developing and assessing Adverse Outcome Pathways: OECD Series on Adverse Outcome Pathways*. Available at: doi:10.1787/5jlv1m9d1g32-en.
- Oldman, D. and Tanase, D. (2018) Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace. In: Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M. and Kaffee, L.-A. *et al.* (Eds.) *The Semantic Web – ISWC 2018*. Springer International Publishing: Cham, pp. 325–340.
- Oliveira, E.C., Ishikawa, E., Hironouchi, L.H., Granja, T.H., A. Nunes, M.V. de and Rodriguez, D. *et al.* (2017 - 2017) Ontology-Based CMS News Authoring Environment. In: *2017 IEEE 11th International Conference on Semantic Computing (ICSC), 2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, 30/01/2017 - 01/02/2017, San Diego, CA. IEEE, pp. 264–265.
- Orellano, P., Reynoso, J., Quaranta, N., Bardach, A. and Ciapponi, A. (2020) Short-term exposure to particulate matter (PM10 and PM2.5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. *Environment international*, 142, 105876. Available at: doi:10.1016/j.envint.2020.105876.
- Paez, A. (2017) Grey literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, 309(2), 597. Available at: doi:10.1111/jebm.12265.
- Page, M.J. and Moher, D. (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. *Systematic Reviews*, 6(1), 29. Available at: doi:10.1186/s13643-017-0663-8.
- Page, M.J., Shamseer, L., Altman, D.G., Tetzlaff, J., Sampson, M. and Tricco, A.C. *et al.* (2016) Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. *PLoS Medicine*, 13(5), e1002028. Available at: doi:10.1371/journal.pmed.1002028.
- Papameletiou, D.e.a. (2017) *SCOEL/REC/300 2-Nitropropane: Recommendation from the Scientific Committee on Occupational Exposure Limits*. Publications Office: Luxembourg.

- Patsopoulos, N.A., Analatos, A.A. and Ioannidis, J.P.A. (2005) Relative citation impact of various study designs in the health sciences. *JAMA*, 293(19), 2362–2366. Available at: doi:10.1001/jama.293.19.2362.
- Pelch, K.E., Li, Y., Perera, L., Thayer, K.A. and Korach, K.S. (2019) Characterization of Estrogenic and Androgenic Activities for Bisphenol A-like Chemicals (BPs): In Vitro Estrogen and Androgen Receptors Transcriptional Activation, Gene Regulation, and Binding Profiles. *Toxicological Sciences*, 172(1), 23–37. Available at: doi:10.1093/toxsci/kfz173.
- Pelch, K.E., Reade, A., Wolffe, T.A.M. and Kwiatkowski, C.F. (2019) PFAS health effects database: Protocol for a systematic evidence map. *Environment international*, 130, 104851. Available at: doi:10.1016/j.envint.2019.05.045.
- Phung, D., Des Connell and Chu, C. (2018) Cardiovascular Risk from Water Arsenic Exposure in Vietnam: Application of Systematic Review and Meta-Regression Analysis in Chemical Health Risk Assessment. *ISEE Conference Abstracts*, 2017(1), 48. Available at: doi:10.1289/isee.2017.2017-48.
- Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., Duffy, S., 2006. *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews. A Product from the ESRC Methods Programme* 211–219.
- Pope, C.A., Burnett, R.T., Turner, M.C., Cohen, A., Krewski, D. and Jerrett, M. *et al.* (2011) Lung Cancer and Cardiovascular Disease Mortality Associated with Ambient Air Pollution and Cigarette Smoke: Shape of the Exposure–Response Relationships. *Environmental Health Perspectives*, 119(11), 1616–1621. Available at: doi:10.1289/ehp.1103639.
- Porta, M. (2014) A Dictionary of Epidemiology. *BMJ*, 2(5402), 140. Available at: doi:10.1093/acref/9780199976720.001.0001.
- Poux, S. and Gaudet, P. (2017) Best Practices in Manual Annotation with the Gene Ontology. In: Dessimoz, C. and Škunca, N. (Eds.) *The Gene Ontology Handbook*. Springer New York: New York, NY, pp. 41–54.
- Proceedings of the Royal Society of Medicine* (Royal Society of Medicine, 1907-1977).
- Prozialeck, W.C. (2013) Biomarkers for Cadmium. In: Kretsinger, R.H., Uversky, V.N. and Permyakov, E.A. (Eds.) *Encyclopedia of Metalloproteins*. Springer New York: New York, NY, pp. 272–277.
- Quansah, R., Semple, S., Ochieng, C.A., Juvekar, S., Armah, F.A. and Luginaah, I. *et al.* (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. *Environment international*, 103, 73–90. Available at: doi:10.1016/j.envint.2017.03.010.
- Radke, E. G. *et al.* (accepted) ‘Application of US EPA IRIS systematic review methods to the health effects of phthalates: lessons learned and path forward’, *Environment international*.
- Radke, E.G., Braun, J.M., Meeker, J.D. and Cooper, G.S. (2018) Phthalate exposure and male reproductive outcomes: A systematic review of the human

- epidemiological evidence. *Environment international*, 121, 764–793. Available at: doi:10.1016/j.envint.2018.07.029.
- Radke, E.G., Galizia, A., Thayer, K.A. and Cooper, G.S. (2019) Phthalate exposure and metabolic effects: a systematic review of the human epidemiological evidence. *Environment international*, 132, 104768. Available at: doi:10.1016/j.envint.2019.04.040.
- Rethlefsen, M.L., Farrell, A.M., Osterhaus Trzasko, L.C. and Brigham, T.J. (2015) Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. *Journal of Clinical Epidemiology*, 68(6), 617–626. Available at: doi:10.1016/j.jclinepi.2014.11.025.
- Rhomberg, L. (2015) Hypothesis-Based Weight of Evidence: An Approach to Assessing Causation and its Application to Regulatory Toxicology. *Risk Analysis*, 35(6), 1114–1124. Available at: doi:10.1111/risa.12206.
- Rhomberg, L.R., Goodman, J.E., Bailey, L.A., Prueitt, R.L., Beck, N.B. and Bevan, C. *et al.* (2013) A survey of frameworks for best practices in weight-of-evidence analyses. *Critical Reviews in Toxicology*, 43(9), 753–784. Available at: doi:10.3109/10408444.2013.832727.
- Ritskes-Hoitinga, M., Leenaars, M., Avey, M., Rovers, M. and Scholten, R. (2014) Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. *The Cochrane Database of Systematic Reviews*, (3), ED000078. Available at: doi:10.1002/14651858.ED000078.
- Roberts, D., Brown, J., Medley, N. and Dalziel, S.R. (2017) Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. *Cochrane Database of Systematic Reviews*. Available at: doi:10.1002/14651858.CD004454.pub3.
- Robinson, I., Webber, J. and Eifrem, E. (2013) *Graph databases*. O'Reilly: Beijing, Sebastopol, CA.
- Rooney, A.A., Boyles, A.L., Wolfe, M.S., Bucher, J.R. and Thayer, K.A. (2014) Systematic review and evidence integration for literature-based environmental health science assessments. *Environmental Health Perspectives*, 122(7), 711–718. Available at: doi:10.1289/ehp.1307972.
- Rooney, A.A., Boyles, A.L., Wolfe, M.S., Bucher, J.R. and Thayer, K.A. (2014) Systematic review and evidence integration for literature-based environmental health science assessments. *Environmental Health Perspectives*, 122(7), 711–718. Available at: doi:10.1289/ehp.1307972.
- Rooney, A.A., Boyles, A.L., Wolfe, M.S., Bucher, J.R. and Thayer, K.A. (2014) Systematic Review and Evidence Integration for Literature-Based Environmental Health Science Assessments. *Environmental Health Perspectives*, 122(7), 711–718. Available at: doi:10.1289/ehp.1307972.
- Rooney, A.A., Cooper, G.S., Jahnke, G.D., Lam, J., Morgan, R.L. and Boyles, A.L. *et al.* (2016) How credible are the study results? Evaluating and applying internal

- validity tools to literature-based assessments of environmental health hazards. *Environment International*, 92-93, 617–629. Available at: doi:10.1016/j.envint.2016.01.005.
- Roundtable on Environmental Health Sciences, Research, and Medicine, Board on Population Health and Public Health Practice and Institute of Medicine (2014) *The Challenge: Chemicals in Today's Society*. National Academies Press (US).
- Rusyn, I. & Shapiro, A. 2013, "Health Assessment Workspace Collaborative (HAWC)"
- Sanchez, K.A., Foster, M., Nieuwenhuijsen, M.J., May, A.D., Ramani, T. and Zietsman, J. *et al.* (2020) Urban policy interventions to reduce traffic emissions and traffic-related air pollution: Protocol for a systematic evidence map. *Environment international*, 142, 105826. Available at: doi:10.1016/j.envint.2020.105826.
- Saran, A. and White, H. (2018) Evidence and gap maps: a comparison of different approaches. *Campbell Systematic Reviews*, 14(1), 1–38. Available at: doi:10.4073/cmdp.2018.2.
- Schaefer, H.R. and Myers, J.L. (2017) Guidelines for performing systematic reviews in the development of toxicity factors. *Regulatory Toxicology and Pharmacology*, 91, 124–141. Available at: doi:10.1016/j.yrtph.2017.10.008.
- Schunemann, H., Hill, S., Guyatt, G., Akl, E.A. and Ahmed, F. (2011) The GRADE approach and Bradford Hill's criteria for causation. *Journal of Epidemiology & Community Health*, 65(5), 392–395. Available at: doi:10.1136/jech.2010.119933.
- Schünemann, H.J. and Moja, L. (2015) Reviews: Rapid! Rapid! Rapid! ...and systematic. *Systematic Reviews*, 4(1), 389. Available at: doi:10.1186/2046-4053-4-4.
- Schwarzman, M.R., Ackerman, J.M., Dairkee, S.H., Fenton, S.E., Johnson, D. and Navarro, K.M. *et al.* (2015) Screening for Chemical Contributions to Breast Cancer Risk: A Case Study for Chemical Safety Evaluation. *Environmental Health Perspectives*, 123(12), 1255–1264. Available at: doi:10.1289/ehp.1408337.
- Sena, E.S., Currie, G.L., McCann, S.K., Macleod, M.R. and Howells, D.W. (2014) Systematic Reviews and Meta-Analysis of Preclinical Studies: Why Perform Them and How to Appraise Them Critically. *Journal of Cerebral Blood Flow & Metabolism*, 34(5), 737–742. Available at: doi:10.1038/jcbfm.2014.28.
- Sheehan, M.C. and Lam, J. (2015) Use of Systematic Review and Meta-Analysis in Environmental Health Epidemiology: a Systematic Review and Comparison with Guidelines. *Current Environmental Health Reports*, 2(3), 272–283. Available at: doi:10.1007/s40572-015-0062-z.
- Shepard, R.B. (Ed.) (2005) *Quantifying Environmental Impact Assessments Using Fuzzy Logic*. Springer New York.
- Silbergeld, E. (2013) Evidence-based toxicology: Strait is the gate, but the road is worth taking. *ALTEX*, 30(1), 67–73. Available at: doi:10.14573/altex.2013.1.067.

- Singla, V.I., Sutton, P.M. and Woodruff, T.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. *American Journal of Public Health*, 109(7), 982–984. Available at: doi:10.2105/AJPH.2019.305068.
- Smith, M.T., Guyton, K.Z., Gibbons, C.F., Fritz, J.M., Portier, C.J. and Rusyn, I. *et al.* (2016) Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. *Environmental Health Perspectives*, 124(6), 713–721. Available at: doi:10.1289/ehp.1509912.
- Sorensen, R. (2018) ‘Vagueness’, in Edward, N. Z. (ed.) *The Stanford Encyclopedia of Philosophy*. Summer 2018. Metaphysics Research Lab, Stanford University.
- Stearns, M.Q., Price, C., Spackman, K.A. and Wang, A.Y. (2001) SNOMED clinical terms: overview of the development process and project status. *Proceedings. AMIA Symposium*, 662–666.
- Stephens, M.L., Betts, K., Beck, N.B., Cogliano, V., Dickersin, K. and Fitzpatrick, S. *et al.* (2016) The Emergence of Systematic Review in Toxicology. *Toxicological Sciences : an Official Journal of the Society of Toxicology*, 152(1), 10–16. Available at: doi:10.1093/toxsci/kfw059.
- Sterne, J.A.C., Higgins, J.P.T. & Reeves, B.C. 2014, "A Cochrane risk of bias tool: for non randomised studies of interventions (ACROBAT-NRSI)"
- Stevens, A., Shamseer, L., Weinstein, E., Yazdi, F., Turner, L. and Thielman, J. *et al.* (2014) Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *BMJ (Clinical Research Ed.)*, 348, g3804. Available at: doi:10.1136/bmj.g3804.
- Stewart, G. (2010) Meta-analysis in applied ecology. *Biology Letters*, 6(1), 78–81. Available at: doi:10.1098/rsbl.2009.0546.
- Stewart, G.B. and Schmid, C.H. (2015) Lessons from meta-analysis in ecology and evolution: the need for trans-disciplinary evidence synthesis methodologies. *Research Synthesis Methods*, 6(2), 109–110. Available at: doi:10.1002/jrsm.1152.
- Sutton, A.J., Cooper, N.J. and Jones, D.R. (2009) Evidence synthesis as the key to more coherent and efficient research. *BMC Medical Research Methodology*, 9, 29. Available at: doi:10.1186/1471-2288-9-29.
- Taylor, J.A., Welshons, W.V. and vom Saal, F.S. (2008) No effect of route of exposure (oral; subcutaneous injection) on plasma bisphenol A throughout 24h after administration in neonatal female mice. *Reproductive Toxicology*, 25(2), 169–176. Available at: doi:10.1016/j.reprotox.2008.01.001.
- Thomas, P.D., Hill, D.P., Mi, H., Osumi-Sutherland, D., van Auken, K. and Carbon, S. *et al.* (2019) Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nature Genetics*, 51(10), 1429–1433. Available at: doi:10.1038/s41588-019-0500-1.

- Tošenovský, E. 2019, "Question for written answer to the Commission", Question for Written Answer to the Commission.
- UK Committee on Toxicity (COT) (2019) 'Statement on phosphate-based flame retardants and the potential for neurodevelopmental toxicity'. Available at: <https://cot.food.gov.uk/cotstatements/cotstatementsyrs/cot-statements-2019/cot-phosphate-based-flame-retardants-statement>.
- US Environmental Protection Agency (2005) *Guidelines for carcinogen risk assessment*. Environmental Protection Agency: Washington, DC.
- US National Research Council (2007) *Toxicity Testing in the 21st Century*. National Academies Press: Washington, D.C.
- US National Research Council (2009) *Science and Decisions: Advancing Risk Assessment*. Washington (DC).
- US National Research Council (2014) *A Framework to Guide Selection of Chemical Alternatives*. Washington (DC).
- US National Research Council (2014) *Review of EPA's Integrated Risk Information System (IRIS) Process*. National Academies Press: Washington, D.C.
- US National Research Council (2014) *Review of EPA's Integrated Risk Information System (IRIS) Process*. National Academies Press: Washington, D.C.
- US National Toxicology Program, 2015. *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration*.
- van Luijk, J., Bakker, B., Rovers, M.M., Ritskes-Hoitinga, M., Vries, R.B.M. de and Leenaars, M. (2014) Systematic reviews of animal studies; missing link in translational research? *PLoS ONE*, 9(3), e89981. Available at: doi:10.1371/journal.pone.0089981.
- Vandenberg, L.N., Ågerstrand, M., Beronius, A., Beausoleil, C., Bergman, Å. and Bero, L.A. *et al.* (2016) A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. *Environmental Health : a Global Access Science Source*, 15(1), 74. Available at: doi:10.1186/s12940-016-0156-6.
- Vandenberg, L.N., Ehrlich, S., Belcher, S.M., Ben-Jonathan, N., Dolinoy, D.C. and Hugo, E.R. *et al.* (2014) Low dose effects of bisphenol A. *Endocrine Disruptors*, 1(1), e26490. Available at: doi:10.4161/endo.26490.
- Villeneuve, D.L., Crump, D., Garcia-Reyero, N., Hecker, M., Hutchinson, T.H., LaLone, C.A. and Landesmann, B. *et al.* (2014) *Adverse outcome pathway (AOP) development I: strategies and principles*. Oxford University Press.
- Villeneuve, D.L., Crump, D., Garcia-Reyero, N., Hecker, M., Hutchinson, T.H., LaLone, C.A. and Landesmann, B. *et al.* (2014) *Adverse outcome pathway development II: best practices*. Oxford University Press.
- vom Saal, F.S., Nagel, S.C., Coe, B.L., Angle, B.M. and Taylor, J.A. (2012) The estrogenic endocrine disrupting chemical bisphenol A (BPA) and obesity.

- Molecular and Cellular Endocrinology*, 354(1-2), 74–84. Available at: doi:10.1016/j.mce.2012.01.001.
- Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M. and Kaffee, L.-A. *et al.* (Eds.) (2018) *The Semantic Web – ISWC 2018*. Springer International Publishing: Cham.
- Vries, R.B.M. de, Hooijmans, C.R., Langendam, M.W., van Luijk, J., Leenaars, M. and Ritskes-Hoitinga, M. *et al.* (2015) A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies. *Evidence-based Preclinical Medicine*, 2(1), e00007. Available at: doi:10.1002/ebm2.7.
- Vries, R.B.M. de, Wever, K.E., Avey, M.T., Stephens, M.L., Sena, E.S. and Leenaars, M. (2014) The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *ILAR Journal*, 55(3), 427–437. Available at: doi:10.1093/ilar/ilu043.
- Walker, V.R., Boyles, A.L., Pelch, K.E., Holmgren, S.D., Shapiro, A.J. and Blystone, C.R. *et al.* (2018) Human and animal evidence of potential transgenerational inheritance of health effects: An evidence map and state-of-the-science evaluation. *Environment international*, 115, 48–69. Available at: doi:10.1016/j.envint.2017.12.032.
- Wang, R.-L. (2020) Semantic characterization of adverse outcome pathways. *Aquatic Toxicology*, 222, 105478. Available at: doi:10.1016/j.aquatox.2020.105478.
- Wang, R.-L., Edwards, S. and Ives, C. (2019) Ontology-based semantic mapping of chemical toxicities. *Toxicology*, 412, 89–100. Available at: doi:10.1016/j.tox.2018.11.005.
- Watford, S., Edwards, S., Angrish, M., Judson, R.S. and Paul Friedman, K. (2019) Progress in data interoperability to support computational toxicology and chemical safety evaluation. *Toxicology and Applied Pharmacology*, 380, 114707. Available at: doi:10.1016/j.taap.2019.114707.
- Whaley, P. (2013) *Systematic Review and the Future of Evidence in Chemicals Policy*.
- Whaley, P., 2013. *Systematic review and the future of evidence in chemicals policy*.
- Whaley, P., Aiassa, E., Beausoleil, C., Beronius, A., Bilotta, G., Boobis, A. and Vries, R. de *et al.* (2019) *Recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER)*.
- Whaley, P., Aiassa, E., Beausoleil, C., Beronius, A., Bilotta, G. and Boobis, A. *et al.* (2020) Recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER). *Environment International*, 143, 105926. Available at: doi:10.1016/j.envint.2020.105926.
- Whaley, P., Aiassa, E., Beausoleil, C., Beronius, A., Bilotta, G., Boobis, A., Vries, R., Hanberg, A., Hoffmann, S., Hunt, N., Kwiatkowski, C., Lam, J., Lipworth, S., Martin, O., Randall, N., Rhomberg, L., Rooney, A.A., Schünemann, H.J., Wikoff, D., Wolffe, T., Halsall, C., 2019. A code of practice for the conduct of systematic reviews in toxicology and environmental health research (COSTER). [Preprint]

- Whaley, P., Edwards, S.W., Kraft, A., Nyhan, K., Shapiro, A., Watford, S., Wattam, S., Wolffe, T.A.M., Angrish, M., (submitted). Knowledge Organization Systems for Systematic Chemical Assessments.
- Whaley, P., Halsall, C., Ågerstrand, M., Aiassa, E., Benford, D. and Bilotta, G. *et al.* (2016) Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations. *Environment international*, 92-93, 556–564. Available at: doi:10.1016/j.envint.2015.11.002.
- Whaley, P., Piggott, T., Morgan, R.L., Wikoff, D., Hoffmann, S., Tsaïoun, K., Thayer, K., Schünemann, H.J., (in prep). “Biological plausibility” and the analysis of indirect evidence in environmental health systematic reviews: a GRADE concept paper.
- Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C. and Tudorache, T. *et al.* (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(suppl), W541-W545. Available at: doi:10.1093/nar/gkr469.
- Wikipedia contributors (2014) “Biological plausibility”. *Wikipedia, The Free Encyclopedia*. Available at: https://en.wikipedia.org/w/index.php?title=Biological_plausibility&oldid=614374435 (Accessed: 16 October 2019).
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M. and Baak, A. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1). Available at: doi:10.1038/sdata.2016.18.
- Williams, A.J., Grulke, C.M., Edwards, J., McEachran, A.D., Mansouri, K. and Baker, N.C. *et al.* (2017) The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics*, 9(1). Available at: doi:10.1186/s13321-017-0247-6.
- Wittwehr, C., Blomstedt, P., Gosling, J.P., Peltola, T., Raffael, B. and Richarz, A.-N. *et al.* (2020) Artificial Intelligence for chemical risk assessment. *Computational Toxicology*, 13, 100114. Available at: doi:10.1016/j.comtox.2019.100114.
- Wolffe, T.A.M., Vidler, J., Halsall, C., Hunt, N. and Whaley, P. (2020) A Survey of Systematic Evidence Mapping Practice and the Case for Knowledge Graphs in Environmental Health and Toxicology. *Toxicological Sciences*, 175(1), 35–49. Available at: doi:10.1093/toxsci/kfaa025.
- Wolffe, T.A.M., Whaley, P., Halsall, C., Rooney, A.A. and Walker, V.R. (2019) Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. *Environment international*, 130, 104871. Available at: doi:10.1016/j.envint.2019.05.065.
- Woodruff, T.J. and Sutton, P. (2010) Pulling back the curtain: improving reviews in environmental health. *Environmental Health Perspectives*, 118(8), a326-7. Available at: doi:10.1289/ehp.1002691.

- Woodruff, T.J. and Sutton, P. (2011) An Evidence-Based Medicine Methodology To Bridge The Gap Between Clinical And Environmental Health Sciences. *Health Affairs*, 30(5), 931–937. Available at: doi:10.1377/hlthaff.2010.1219.
- Woodruff, T.J. and Sutton, P. (2014) The Navigation Guide Systematic Review Methodology: A Rigorous and Transparent Method for Translating Environmental Health Science into Better Health Outcomes. *Environmental Health Perspectives*, 122(10), 1007–1014. Available at: doi:10.1289/ehp.1307175.
- World Health Organisation Chemical Risk Assessment Network, (in prep). A Framework for Conduct of Systematic Reviews in Chemical Risk Assessment. World Health Organisation.
- Yost, E.E., Euling, S.Y., Weaver, J.A., Beverly, B.E.J., Keshava, N. and Mudipalli, A. *et al.* (2019) Hazards of diisobutyl phthalate (DIBP) exposure: A systematic review of animal toxicology studies. *Environment international*, 125, 579–594. Available at: doi:10.1016/j.envint.2018.09.038.
- Zoeller, R.T., Bergman, Å., Becher, G., Bjerregaard, P., Bornman, R. and Brandt, I. *et al.* (2014) A path forward in the debate over health impacts of endocrine disrupting chemicals. *Environmental Health : a Global Access Science Source*, 13, 118. Available at: doi:10.1186/1476-069X-13-118.

Appendices

| | |
|--|-----|
| Appendix A: Five Lessons | 118 |
| Appendix B: Protocols.io | 121 |
| Appendix C: Systematic Evidence Maps | 133 |
| Appendix D: Knowledge Graphs | 143 |
| Appendix E: NASEM Presentation..... | 158 |

Appendix A: Five Lessons

- This document is online at: [10.1016/j.envint.2016.04.016](https://doi.org/10.1016/j.envint.2016.04.016)

Environment International 92–93 (2016) 553–555



Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Preface

Assuring high-quality evidence reviews for chemical risk assessment: Five lessons from guest editing the first environmental health journal special issue dedicated to systematic review



While systematic review (SR), the rigorous methodology for selecting, appraising and synthesising existing evidence in order to answer a research question, may not yet be mainstream among environmental scientists and toxicologists, interest in the methods and what they may bring to chemical risk research is growing rapidly and is evident in an exponential increase in publications over the last 20 years (Fig. 1).

Mirroring the rapid growth of a nascent literature is the proliferation of initiatives, many of which are collaborative, seeking to extend the conduct of systematic reviews to pre-clinical research and laboratory animal experimentation. These include the Systematic Review Centre for Laboratory animal Experimentation¹ (SYRCLE) and the Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies² (CAMARADES), while efforts to apply SR methods to the toxicological sciences are now coalescing in the form of networks such as the Navigation Guide³ and the Evidence Based Toxicology Collaboration⁴ (EBTC), among others. These initiatives are identifiable by a shared view that SR methods are a vital area of research in their own right, have the potential to greatly improve the scientific quality of reviews of existing evidence, and will facilitate the translation of pre-clinical and toxicological research into evidence-based medical, public health and environmental policy-making.

The purpose of this Special Issue is to contribute to this agenda by promoting interest in and discussion of how SR methods can advance the transparency and scientific rigour of chemical risk assessment (CRA). We have brought together assorted commentaries on the prospects and potential benefits of SR methods for CRA, methods papers explaining how SR methods can be adapted or refined for the CRA context, and a set of full-blown systematic reviews, each of which functions as a case study of how SR methods can apply in practice as well as being valuable pieces of environmental health research in their own right.

The increase in the number of toxicology journal papers with "systematic review" in the title is an encouraging indicator of the regard with which SRs are held in the scientific community. However, proven quality assurance procedures for SRs in environmental health research are limited. This risks a proliferation of publications of variable quality, potentially blunting the influence of SRs as powerful

tools for evidence-based decision-making and undermining the case for using SR methods to synthesise evidence in CRA. With the issue of quality assurance in mind we have drawn up a number of lessons which, while perhaps common knowledge in other fields, have been reinforced for us while editing this Special Issue. The lessons are aimed at SR authors, reviewers and, importantly, journal editors who are being faced with an increasing number of manuscripts that purport to be systematic reviews.

We believe this is the first Special Issue dedicated to systematic review published by an environmental health journal. In spite of the inevitable imperfections this entails, we hope the reader agrees this Special Issue has been a success. We would like to thank all the authors, peer reviewers and funders who contributed to this Special Issue and our initial workshop organised through the Royal Society of Chemistry, of which this Special Issue was one output (detailed in Whaley et al., 2015). We also hope the reader will share our enthusiasm for SR methods and recognise the potential for their uptake and effectiveness in shaping the future of chemical risk assessment.

Lesson 1: Submitting authors should be provided with detailed guidance about how to report systematic reviews and encouraged to describe how they fulfilled it

Uneven understanding from authors as to the precise requirements of conducting and reporting CRA-related SRs in a comprehensive and transparent fashion is unsurprising given the novelty of the methods. We received a number of SR submissions which, while of high potential scientific value, were obscured by poor write-up. In order to avoid rejecting good research for want of adequate reporting, the editors and peer-reviewers ended up with a substantial workload in providing the authors with guidance as to how their SRs should have been reported. The authors themselves had the burden of making substantial revisions to their manuscripts.

With hindsight, we believe we could have saved probably one revision round for several of the submitted SRs by insisting in advance that they conform at least to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), a relatively straightforward checklist of items to report in a systematic review or meta-analysis already widely endorsed by medical journals (Moher et al., 2009). One review (Joca et al., 2016), unprompted by us, even went so far as to explain in supplementary information how they had fulfilled each PRISMA requirement. This was extremely helpful in providing a clear

¹ Website: <https://www.radboudumc.nl/Research/Organisationofresearch/Departments/cdl/SYRCLE/Pages/default.aspx>.

² Website: <http://www.dcn.ed.ac.uk/camarades/>.

³ Website: <http://prhe.ucsf.edu/prhe/navigationguide.html>.

⁴ Website: <http://www.ebttox.com/>.

<http://dx.doi.org/10.1016/j.envint.2016.04.016>

0160-4120/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

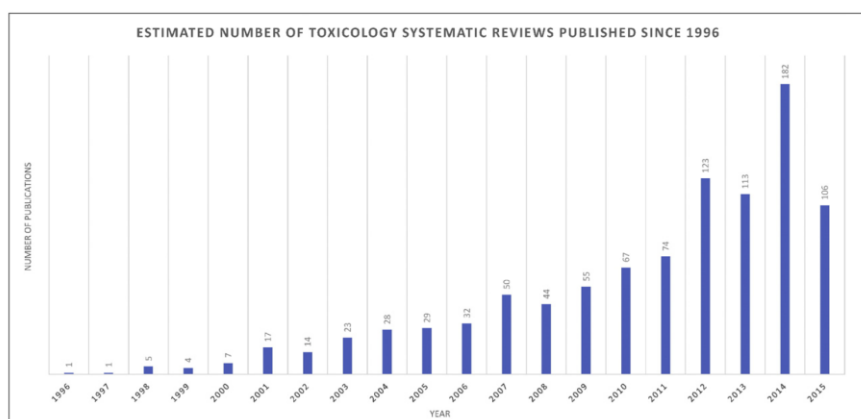


Fig. 1. Papers indexed in Web of Science (WoS) with the term "Systematic Review" in the publication title, filtered for "Toxicology" as topic, excluding topic of "Pharmacology Pharmacy". WoS database search excludes Biosis Citation Index (not subscribed). Date of search: 4 April 2016.

picture of the strengths and limitations of the SR methods employed and we would strongly encourage other SR authors to do the same.

The PRISMA checklist is not exhaustive and there may be room for developing detailed reporting guidance specifically for toxicology SRs. While editing the Special Issue we became aware of initiatives such as the Methodological Expectations of Cochrane Intervention Reviews (MECIR, 2012), which provide a lengthy and detailed checklist of "must-haves" and "should-haves" for conduct and reporting of SRs adapted for different medical disciplines. As editors, we would like to flag the potential for adapting MECIR standards to the current research context.

Lesson 2: Editors need to invest in developing a balanced peer-review group and cultivate a network of interdisciplinary expertise in the review pool

In principle, peer-review of an SR is straightforward: each submission should be attended by two content experts and a SR methods expert. The problem is, this is easier said than done. One SR submission spent 111 days between first reviewer accepting invitation to review the manuscript and the three required reviews finally being completed. Although as editors we bear full responsibility for this, it is indicative of several challenges we faced in securing peer-review for SRs, insofar as they are often lengthy, complex, and require a breadth of interdisciplinary expertise to be reviewed fairly. While content experts were relatively easy to find, experts in SR methods were much harder to secure and we ended up leaning heavily on a relatively small group of SR experts, to whom we are extremely grateful for their commitment and patience.

Of course, access to a comprehensive peer-review pool of interdisciplinary expertise is not something which can be secured overnight, but efforts need to be made by journals to help editors identify and keep track of reliable reviewers who can handle the specific demands of systematic reviews. Databases to help editors identify peer-reviewers do exist, and we used them in editing the Special Issue, but it was very difficult to filter appropriate reviewers from the long lists of those identified as potentially suitable. In particular, being able to quickly identify reviewers with specific SR experience (either as researchers or as reviewers) would have been very helpful.

Reviewers initially brought in as content experts will quickly acquire relevant SR experience in the course of reviewing SRs. With the right guidance and training (as we touch on in Lesson 3 below), we anticipate

that content experts can therefore be cultivated into a pool of competent SR reviewers. To be effective, editors need to treat this cultivation as an active process and should be supported by easy access to more detailed information about the review histories of individual peer-reviewers and, for example, relevant training they might have received.

Lesson 3: Peer-reviewers should be provided with detailed guidance and ideally training in how to critically appraise systematic reviews

There is a major challenge in ensuring that even an experienced SR researcher provides a sufficiently thorough critical appraisal of a submitted SR, such that all the important methodological features of the submission have been given due consideration. For less experienced SR reviewers, the challenge multiplies. For example, we found that reviewers without significant experience in SR were often bemused by the level of detail presented in the SRs they were reviewing and/or the value of an additional review in a field in which literature reviews might already be plentiful. We also found many reviewers were insufficiently alert to obvious flaws in conduct or reporting of a review. The best reviews came from experienced SR researchers with substantial field expertise; however, these researchers are currently limited in number and present an unsustainably small pool of reviewers from which to draw. As editors with experience in SR methods we were able to compensate for some of the shortcomings of the review process but such a hands-on approach, spending as much as eight hours on some submissions, is likely to be too time-consuming to become standard practice.

We believe that securing the balance of competence to assess both the scientific content of the systematic review, the limitations in design, conduct and reporting of the SR, and ensuring that the peer-review is sufficiently thorough, would have been significantly facilitated by provision for peer-reviewers of detailed guidance on how to critically appraise a SR (i.e. a structured approach to determining which methodological features need to be present in a SR, and how to distinguish when those features either have or have not been implemented validly). While uneven quality of peer-review comments is a fact of editorial life, editors and journals can do much more to educate and train peer-reviewers, to increase the likelihood that the review process will provide fair, valuable and comprehensive feedback to the submitting authors, and more consistently identify those SRs which should be published.

Lesson 4: Journals need to implement a formal-but-flexible standard for publishing SRs

As will quickly become evident to the reader of this Special Issue, we did not implement a standard approach for formatting and structuring reviews or handling supplementary material and appendices. Some papers present structured abstracts, some do not (the former is clearly preferable as it is standard practice in the field of medicine, for example). Similar to guidance for authors on what to report in a SR, publishers should have their own formal but flexible guidance on what they expect to present in a systematic review and how it should be structured. This should cover: basic SR structure; the provision of a structured abstract; the handling of appendices and supplemental material; and so forth.

Lesson 5: All systematic reviews should be preceded by formal publication of protocols

Pre-publication of protocols is already considered essential for systematic reviews in other fields (e.g. the Cochrane Collaboration in clinical medicine), in part to prevent methodological choices being influenced by what the reviewers might be learning in the course of conducting a SR. Editing the Special Issue reinforced another aspect of the value of protocols: they provide an opportunity for external appraisal and validation of planned methods before conducting the systematic review, which in turn allows SR authors to minimise effort before risking rejection of an inadequate, completed manuscript.

Rejection of a systematic review because of basic errors such as ineffective search strategies, ambiguously articulated or invalid eligibility criteria, or the use of statistical methods, is a poor return on the large time investment in conducting a review. This is potentially avoidable if a protocol is submitted for peer-review prior to the decision to proceed with conducting the full SR. While disappointing, the cost of rejection of a protocol is only the time spent planning a review, which is far preferable to rejection after completing a SR using flawed methods.

This first-stage peer-review of a submitted protocol may also provide valuable critical appraisal of methodological choices before the full SR process commences. Not only will the SR benefit from this, the preliminary but still substantial work done by the authors in developing a SR protocol can be recognised by citation in the literature. As editors of this Special Issue we were not in a position to insist on pre-publication of protocols, and while it is possible to pre-publish protocols through databases such as PROSPERO⁵, here we urge that consideration be given to the value

of formal publication of protocols in peer-reviewed journals as an important step in the quality assurance of SRs, in particular assuring the validity of methodological choices. This has already been implemented by the journal *Environmental Evidence* (*Collaboration for Environmental Evidence, 2016*) and is a practice which could be adopted elsewhere.

References

- Collaboration for Environmental Evidence, 2016. Guidelines for Authors. (accessed 10 April 2016) <http://www.environmentalevidence.org/information-for-authors>.
- Joca, L., Sacks, J.D., Moore, D., Lee, J.S., Sams 2nd, R., Cowden, J., 2016 Feb 18. Systematic review of differential inorganic arsenic exposure in minority, low-income, and indigenous populations in the United States. *Environ. Int.* <http://dx.doi.org/10.1016/j.envint.2016.01.011> (pii: S0160-4120(16)30011-3 [Epub ahead of print] PubMed PMID: 26896853).
- MECIR, 2012. Methodological expectations of cochrane intervention reviews (MECIR) standards for the conduct (version 2.2) and reporting (version 1.1) of new cochrane intervention reviews. (accessed 6 April 2016) <http://www.editorial-unit.cochrane.org/mecir>.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., P.R.I.S.M.A. Group, 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 6 (7), e1000097. <http://dx.doi.org/10.1371/journal.pmed.1000097>.
- Whaley, P., Halsall, C., Ågerstrand, M., Aiassa, E., Benford, D., Bilotta, G., Coggon, D., Collins, C., Dempsey, C., Duarte-Davidson, R., FitzGerald, R., Galay-Burgos, M., Gee, D., Hoffmann, S., Lam, J., Lassen, T., Levy, L., Lipworth, S., Ross, S.M., Martin, O., Meads, C., Meyer-Baron, M., Miller, J., Pease, C., Rooney, A., Sapiets, A., Stewart, G., Taylor, D., 2015 Dec 10. Implementing systematic review techniques in chemical risk assessment: challenges, opportunities and recommendations. *Environ. Int.* <http://dx.doi.org/10.1016/j.envint.2015.11.002> (pii: S0160-4120(15)30086-6 [Epub ahead of print] PubMed PMID: 26687863).

Paul Whaley*
Crispin Halsall

Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

11 April 2016

⁵ Website: <http://www.crd.york.ac.uk/PROSPERO/>.

Appendix B: Protocols.io

The interactive version of this document is online at:

<https://dx.doi.org/10.17504/protocols.io.biktckwn>

OPEN ACCESS

protocols.io



Sep 10, 2020

Generic Protocol for Environmental Health Systematic Reviews Based on COSTER Recommendations

Paul Whaley¹

¹Lancaster University

In Development dx.doi.org/10.17504/protocols.io.biktckwn

Systematic Reviews



Paul Whaley

ABSTRACT

A protocol template to help researchers follow the [COSTER recommendations](#) for conduct of systematic reviews. This instance covers the planning steps of a systematic review and will help with writing up the systematic review protocol.

The intent is to convert COSTER from a checklist of things which need to be done into a sequence of actions which can be followed by a research team.

When completing the protocol and either registering it or submitting it to a journal, please cite this instance of the protocol template and the parent manuscript, DOI [10.1016/j.envint.2020.105926](https://doi.org/10.1016/j.envint.2020.105926).

EXTERNAL LINK

<https://www.sciencedirect.com/science/article/pii/S016041202031881X>

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Whaley, Paul, Elisa Aiassa, Claire Beausoleil, Anna Beronius, Gary Bilotta, Alan Boobis, Rob de Vries, et al. 2020. "Recommendations for the Conduct of Systematic Reviews in Toxicology and Environmental Health Research (COSTER)." *Environment International* 143 (July): 105926.

DOI

dx.doi.org/10.17504/protocols.io.biktckwn

PROTOCOL CITATION

Paul Whaley 2020. Generic Protocol for Environmental Health Systematic Reviews Based on COSTER Recommendations. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.biktckwn>

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

Whaley, Paul, Elisa Aiassa, Claire Beausoleil, Anna Beronius, Gary Bilotta, Alan Boobis, Rob de Vries, et al. 2020. "Recommendations for the Conduct of Systematic Reviews in Toxicology and Environmental Health Research (COSTER)." *Environment International* 143 (July): 105926.

EXTERNAL LINK

<https://www.sciencedirect.com/science/article/pii/S016041202031881X>

KEYWORDS

systematic review, environmental health, toxicology, protocol

LICENSE

This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author

protocols.io

1

09/10/2020

Citation: Paul Whaley (09/10/2020). Generic Protocol for Environmental Health Systematic Reviews Based on COSTER Recommendations. <https://dx.doi.org/10.17504/protocols.io.biktckwn>

This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/) (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

and source are credited

IMAGE ATTRIBUTION

Image by Paul Whaley.

CREATED

Jul 15, 2020

LAST MODIFIED

Sep 10, 2020

PROTOCOL INTEGER ID

39283

GUIDELINES

Protocols.io has not yet been optimised as a means for reporting what was done in response to complex instructions such as those found in this protocol. Feedback on use of the protocol, and how to develop it to facilitate reporting of planned methods, would be very much appreciated.

Securing capacity, competencies, and tools

1 Assess the team's combined competence in conduct of a systematic review. *Recommendation 1.1.1.*

| Competency | Team member(s) (initials) |
|---|---------------------------|
| Information science (for e.g. search strategies) | |
| Evidence appraisal methods (i.e. risk of bias assessment) | |
| Statistical methods | |
| Domain or subject expertise | |
| Systematic review methods | |
| Team member competencies | |

2 Identify information management practices and tools for each stage of the review. *Recommendation 1.1.2.*

| Information management component | Tools or packages |
|--|-------------------|
| Reference manager | |
| Knowledge management tool | |
| Systematic review software | |
| Statistics software and packages | |
| Artificial Intelligence support tools (e.g. for screening) | |
| Information management tools and packages | |

3 List the potential conflicts of interest of the authors. *Recommendations 1.1.3 and 1.1.4.*

This should include both financial and non-financial interests which readers should be aware of in order to understand the motivations of the authors of the review.

By listing the interests as *potential*, you are confirming that they are not *apparent* conflicts of interest, i.e. they cannot reasonably be expected to compromise the integrity of the systematic review. People with apparent conflicts of interest should be excluded from decision-making roles in the review.

Interests should be declared using the ICMJE Conflict of Interest Disclosure Forms, as attached. The summary statements generated by the forms for each author can be copy-pasted into the table.

[ICMJE COI Disclosure Form.pdf](#)

| Author | ICMJE COI Summary |
|--------|-------------------|
| | |
| | |
| | |

Summary statements of authors declared conflicts of interest.

Setting the research question ("problem formulation")

4 Demonstrate the need for a new review. *Recommendation 1.2.1*

4.1 Describe the scientific value of the question(s), i.e. why it is important that it be investigated.


4.2 Describe the importance to stakeholders of the question(s) being asked.

4.3 Summarise relevant existing primary research and evidence syntheses to justify conducting a new systematic review.

5 Articulate the scientific rationale for each question via development of a theoretical framework. *Recommendation 1.2.2.* For example, this would describe how the exposure is related to the outcomes of interest if the systematic review is an investigation of an exposure-outcome relationship. The theoretical framework should include discussion of the biological plausibility of the relationship being investigated.

6 For each research question to be answered by the review, prospectively define a statement of the research objective in terms of Population, Exposure or Intervention, Comparator, Outcome, Study Design, and Target Condition, selected as appropriate. *Recommendation 1.2.3.*

- Authors may wish to refer to Morgan et al. 2018 for guidance on how to formulate research questions as PECO statements. Conceiving of an ideal study may also help characterise the PECO elements which define what type of study will be informative for your review findings.

 Morgan RL, Whaley P, Thayer KA, Schünemann HJ (2018). Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environment international*. <https://doi.org/10.1016/j.envint.2018.07.015>

- 6.1 **Define the target Population of interest.** These are the objects of investigation, i.e. the entities to which exposures or interventions happen.

| | |
|-----------------------------------|--|
| Species | |
| Sex | |
| Age | |
| Health status | |
| Additional characteristics | |

Characteristics of the population of interest. Add rows to cover other population characteristics relevant to the SR question.

- 6.2 **Define the target Exposure or Intervention of interest.** This concerns the administered or observed change in conditions of the objects of investigation. It should include timing, duration and dose.

| | |
|---------------------------------|--|
| Exposure or intervention | What is the exposure or intervention? |
| Timing | When does the exposure or intervention happen? |
| Duration | For how long does the exposure or intervention last? |
| Dose | What is the dose regimen (amount, frequency)? |

Timing, duration and dose of the exposure / intervention. Add rows to cover other exposure / intervention characteristics relevant to the SR question. Add a new table for each exposure or intervention of interest.

- 6.3 **Define the target Comparator of interest.** This concerns the characteristics of the exposure or intervention being used as the comparator to which the target exposure or intervention is being compared.

| | |
|-------------------|---|
| Comparator | What is the comparator exposure or intervention? |
| Timing | When does the comparator happen? |
| Duration | For how long is the comparator administered? |
| Dose | What is the dose of the comparator (amount, frequency)? |

Timing, duration and dose of the comparator. Add rows to cover other comparator characteristics relevant to the SR question.

- 6.4 **Define the target Outcome(s) of interest.** This concerns the change being measured in the exposure or intervention group. These should be the primary outcomes of interest to the systematic review which form the hypothesis or hypotheses being tested. Secondary outcomes can also be listed.

| | |
|--------------------------|--|
| Primary outcome 1 | |
| Primary outcome 2 | |
| | |

Primary outcomes of interest. Add new rows for each outcome of interest.

| | |
|----------------------------|--|
| Secondary outcome 1 | |
| Secondary outcome 2 | |
| | |

6.5 Define the Target Condition. This is the object of a test method for diagnosis or detection. It is only necessary for a systematic review of a diagnostic or detection test method.

| | |
|--|--|
| Target condition characteristic 1 | |
| Target condition characteristic 2 | |
| | |

Defining the eligibility criteria and designing the process for screening evidence for inclusion

7 Define and justify unambiguous and appropriate eligibility criteria for each component of the objective statement. *Recommendation 1.3.1, 1.3.3, 1.3.4, 1.3.5*

| PECO element | Description of eligibility criteria |
|------------------------------------|---|
| Eligible populations | Include e.g. age, sex, health status, socioeconomic status, occupation etc. |
| Eligible exposures | Include timing, methods for measurement exposure |
| Eligible comparators | The populations and exposures against which the exposed populations are being compared |
| Eligible primary outcomes | Specify the outcome, whether the outcome is apical (whole organism) or intermediate (is a marker of an apical outcome); the acceptable outcome measures (diagnostic criteria, scales, etc.) and timing of outcome measurement |
| Eligible secondary outcomes | Specify the outcome, whether the outcome is apical (whole organism) or intermediate (is a marker of an apical outcome); the acceptable outcome measures (diagnostic criteria, scales, etc.) and timing of outcome measurement |

| | |
|-------------------------------|---|
| Eligible study designs | Define eligible study designs by design features rather than design labels. |
|-------------------------------|---|

Describe the eligibility criteria for each PECO element. Add additional PECO elements as appropriate.

| PECO element | Description of exclusion criteria | Reasons for exclusion |
|-----------------------------|-----------------------------------|-----------------------|
| Excluded populations | | |
| Excluded exposures | | |
| Excluded comparators | | |
| Excluded outcomes | | |


Describe the criteria for exclusion of studies, according to each PECO element. Add additional PECO elements as appropriate.

- 8 **Define the points at which screening for eligibility will take place.** *Recommendation 1.3.2.* Will there be screening at title and abstract, full text, or both?

| |
|--|
| Points at which screening will take place |
| Describe whether there will be screening at title and abstract, full text, or both |
| Points at which screening will take place |

- 9 **Include all relevant, publicly-available evidence**, except for research for which there is insufficient methodological information to allow appraisal of internal validity. *Recommendation 1.3.6.* **Exclude evidence which is not publicly available.** *Recommendation 1.3.9*

| |
|---|
| Policy on eligibility of grey literature and unpublished evidence |
| Describe how grey literature will be handled in the systematic review. If some or all grey literature is to be excluded, explain why and anticipate its implications as a limitation of review methods. |
| Policy on grey literature for the systematic review. |

 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.

Only publicly available information about a study should be eligible for inclusion. If the planned SR will bring into the public domain evidence which was previously inaccessible, this makes the evidence eligible for inclusion.

Studies for which there is insufficient information for risk of bias to be evaluated should be excluded from a SR, to prevent the inclusion in a SR of evidence that is potentially misleading but cannot be identified as such by the reviewers.

- 10 **Include evidence which is relevant to review objectives irrespective of whether its results are in a usable form.** *Recommendation 1.3.7*

| |
|--|
| Policy on eligibility of studies with unusable data |
|--|

Describe how studies which report their results in a manner incompatible with planned analyses will be handled in the systematic review.

Policy on usability of study data



COSTER recommends that documents 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; however, they may be excluded from specific synthesis steps such as meta-analysis.

11 Include relevant evidence irrespective of language. *Recommendation 1.3.8.*

Policy on eligibility of studies based on language

State the language/s in which the systematic review will be written, and how studies not written in that language will be handled.

Policy on language

Languages to be included in the systematic review

| |
|--|
| |
| |
| |
| |

List of included languages in the systematic review

12 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. *Recommendation 3.4*

Multiple publications policy

Describe how multiple publications derived from the same study will be aggregated.

Policy on handling of multiple publications from same study

13 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. *Recommendation 3.1*

Team members who will conduct screening

Method for resolving disputes

| | |
|--|--|
| | |
|--|--|

Planned approach to duplicate screening and dispute resolution

14 Design the PRISMA flow chart for presentation of the results of the screening process.

Recommendation 3.2

15 Pilot test the screening process. *Recommendation 1.4.7*

A generic protocol for piloting the screening stage of a systematic review is available here:
<https://www.protocols.io/view/a-general-protocol-for-pilot-testing-the-screening-bkc9ksz6>

Defining the strategy for searching for evidence relevant to the review objectives

16 Design sufficiently sensitive search criteria, so that studies which meet the eligibility criteria of the review are not inadvertently excluded. Document the search methods in sufficient detail to render them transparent and reproducible. *Recommendations 1.4.1, 2.6*16.1 Search all the key scientific databases for the topic, including national, regional and subject-specific databases. *Recommendation 2.1*

| List of databases |
|-------------------|
| Database 1 |
| Database 2 |
| Database 3 |

List of databases searched in the systematic review

16.2 Structure search strategies for each database, electronic and other source, using appropriate^{2d} controlled vocabulary, free-text terms and logical operators in a manner which prioritises sensitivity. Document the search methods and results in sufficient detail to render them transparent and reproducible. *Recommendations 2.3, 2.6*

| Database | Search strategy |
|----------|-----------------|
| | |
| | |
| | |

Search strategy for each database in the systematic review



Atkinson KM, Koenka AC, Sanchez CE, Moshontz H, Cooper H (2015). Reporting standards for literature searches and report inclusion criteria: making research syntheses more transparent and easy to replicate. Research synthesis methods.
<https://doi.org/10.1002/jrsm.1127>

16.3 Define reproducible strategies for identifying and searching sources of grey literature^{2d} (databases, websites etc.). Document the search methods and results in sufficient detail to render them transparent and reproducible. *Recommendations 2.2 and 2.6*

| Grey literature source | Search strategy | Date of search | No. of results |
|------------------------|-----------------|----------------|----------------|
| | | | |
| | | | |

Search strategy for each source of grey literature in the review

16.4 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. **Search by contacting relevant individuals and organisations.**

Recommendations 2.4 and 2.5

| Supplementary search strategies | Indicate if will be used |
|---|--------------------------|
| Hand search references of included studies | |
| Hand search references of relevant reviews | |
| Hand search references of studies cited by included studies | |
| By contacting individuals and organisations | |
| Other | |
| Supplementary search strategies | |

17 Plan for re-running 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). *Recommendation 2.7*

| | |
|---------------------------|---|
| Timing | When will the searches be updated prior to publication of the review? |
| Sources | Which sources will be searched again? |
| Level of screening | What level of screening will be conducted? |
| Updating findings | How will review findings be updated in context of new studies? |


Policy for updating searches

Methods for synthesising and evaluating the evidence

18 Design the "characteristics of included studies" table. *Recommendation 1.4.2*

19 Design and pilot the data extraction forms. *Recommendation 1.4.7*

20 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. *Recommendation 1.4.3*

 Review teams may find the **FEAT** (Focus-Extent-Application-Transparency) mnemonic to be useful in defining their risk of bias assessment methods.

- **Focus:** The focus of the tool should be exclusively the internal validity of a study. If other quality constructs are of interest, each should be assessed in a separate process.
- **Extent:** All the important threats to internal validity should be covered by the tool. If observational studies are being appraised, the threats should include all important confounders.
- **Application:** The appraisal process should produce consistent, accurate descriptions of the extent to which a study is vulnerable to each identified threat to internal validity. The judgements should be in a form which can be logically incorporated into the evidence synthesis.
- **Transparency:** The reason for each judgement should be documented, quoting as justification relevant text from the study documentation.

Refer to Section 5 of the COSTER recommendations for detail on how the risk of bias assessment process should be conducted.

20.1 Define the tool selection and modification process (how will a suitable tool be identified, and what process will be followed to identify and validate any necessary modifications?) *Recommendation 1.4.3*

| Tool selected | Studies to which it is applied | Modifications made | Method for validating modifications |
|---------------|--------------------------------|--------------------|-------------------------------------|
| | | | |
| | | | |

Selection of tools to be used in systematic review

20.2 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. *Recommendation 5.3*

| Team members conducting risk of bias assessment | Method for identifying and settling disputes |
|---|--|
| | |

Approach to conducting risk of bias assessment

20.3 Define the training and piloting process for the risk of bias assessment (how will the review team be trained in use of the tool, and what are the conditions under which the piloting process will be determined satisfactory?) *Recommendation 1.4.7*

- 21 **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. *Recommendations 1.4.4, 6.1*

| Synthesis Component | Planned Methods |
|--|-----------------|
| Qualitative or narrative methods | |
| Quantitative methods | |
| Conditions for combining studies in overall and subgroup analyses | |
| Choice of effect measure | |
| Assessment of heterogeneity (6.3) and consequences of developing summary results (6.4) | |
| Effect modifiers for subgroup analysis | |
| Transformation of scales into common measures (6.2) | |
| Assessment of publication bias (6.5) | |
| Impact of the risk of bias assessment on the synthesis (6.6) | |
| Sensitivity analyses (6.7) | |
| Other methods | |

Methods for synthesising the included evidence



Refer to section 6 of COSTER for detailed recommendations for how evidence should be synthesised in systematic reviews. Popay et al. (2006) attached provides very useful guidance on how to approach the non-quantitative components of the synthesis.

[Popay et al. 2006 - Guidance on the Conduct of Narrative Synthesis in Systematic Reviews.](#)

- 22 **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.) *Recommendation 1.4.5*



The components of assessment of confidence or certainty in the evidence are described in section 7 of COSTER.

- 22.1 **Pilot the process for the assessment of confidence in the results of the synthesis of the evidence.** How will the review team be trained in use of the tool, and what are the conditions under which the piloting process will be determined satisfactory? *Recommendation 1.4.7*

- 23 **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.** *Recommendation 1.4.6*

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.

Registering and publishing the protocol

- 24 **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. *Recommendation 1.5.1*
- 25 **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. *Recommendation 1.5.2*
- 26 **Publish the final version of the protocol in a public archive**, prior to screening studies for inclusion in the review. *Recommendation 1.5.3*



Publication of the protocol in a journal is equivalent to publication in a public archive.

Appendix C: Systematic Evidence Maps

This manuscript is available at this DOI: [10.1016/j.envint.2019.05.065](https://doi.org/10.1016/j.envint.2019.05.065)

Environment International 130 (2019) 104871



Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint



Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management



Taylor A.M. Wolffe^{a,b,*}, Paul Whaley^{a,d}, Crispin Halsall^a, Andrew A. Rooney^c, Vickie R. Walker^c

^a Lancaster Environment Centre, Lancaster University, Lancaster, UK

^b Yorlas Group, Lancaster Environment Centre, Lancaster University, Lancaster, UK

^c Division of the National Toxicology Program, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA

^d Evidence-Based Toxicology Collaboration, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

ARTICLE INFO

Handling Editor: Haima Boogaard

Keywords:
Systematic review
Evidence mapping

ABSTRACT

Background: While systematic review (SR) methods are gaining traction as a method for providing a reliable summary of existing evidence for health risks posed by exposure to chemical substances, it is becoming clear that their value is restricted to a specific range of risk management scenarios - in particular, those which can be addressed with tightly focused questions and can accommodate the time and resource requirements of a systematic evidence synthesis.

Methods: The concept of a systematic evidence map (SEM) is defined and contrasted to the function and limitations of systematic review (SR) in the context of risk management decision-making. The potential for SEMs to facilitate evidence-based decision-making are explored using a hypothetical example in risk management priority-setting. The potential role of SEMs in reference to broader risk management workflows is characterised.

Results: SEMs are databases of systematically gathered research which characterise broad features of the evidence base. Although not intended to substitute for the evidence synthesis element of systematic reviews, SEMs provide a comprehensive, queryable summary of a large body of policy relevant research. They provide an evidence-based approach to characterising the extent of available evidence and support forward looking predictions or trendspotting in the chemical risk sciences. In particular, SEMs facilitate the identification of related bodies of decision critical chemical risk information which could be further analysed using SR methods, and highlight gaps in the evidence which could be addressed with additional primary studies to reduce uncertainties in decision-making.

Conclusions: SEMs have strong and growing potential as a high value tool in resource efficient use of existing research in chemical risk management. They can be used as a critical precursor to efficient deployment of high quality SR methods for characterising chemical health risks. Furthermore, SEMs have potential, at a large scale, to support the sort of evidence summarisation and surveillance methods which would greatly increase the resource efficiency, transparency and effectiveness of regulatory initiatives such as EU REACH and US TSCA.

1. Introduction

Systematic review is the epitome of the evidence-based approaches that have revolutionized clinical decision-making. The methodology was developed in response to medical practitioners' need to distill clear and reliable conclusions about the efficacy of clinical interventions from an evidence base seemingly full of contradiction, heterogeneity and bias (Chalmers et al., 2002; Garg et al., 2008; Higgins and Green, 2011). This need parallels that of chemicals policy; where conclusions regarding the safety of exposure to a chemical substance must be synthesised from a significantly more disparate evidence base (Whaley et al., 2016).

Consequently, interest in the application of systematic review to regulatory decision-making contexts within chemicals policy and wider environmental health is growing. This is evidenced by the increasing number of systematic reviews published in the field (Whaley and Halsall, 2016), the establishment of collaborations and workgroups dedicated to development and dissemination of environmental health systematic review methodology (Morgan et al., 2016; NTP, 2015; Woodruff and Sutton, 2014), and the adoption and use of systematic review by regulatory bodies such as the United States Environmental Protection Agency (US EPA) (EPA, 2018; The National Academies of Sciences, 2017) and World Health Organization (Mandrioli et al., 2018).

* Corresponding author at: Lancaster Environment Centre, Lancaster University, Lancaster, UK.
E-mail address: twolffe@lancaster.ac.uk (T.A.M. Wolffe).

<https://doi.org/10.1016/j.envint.2019.05.065>

Received 28 March 2019; Received in revised form 10 May 2019; Accepted 24 May 2019

Available online 26 June 2019

0160-4120/© 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Growing interest in systematic review approaches is indicative of the evolutionary journey chemicals regulation follows as it attempts to reconcile past oversights with present day knowledge and mounting future challenges. A number of legacy chemicals released to market under past regulatory workflows persist on the market without risk assessment. Meanwhile, an overwhelming number of new chemicals are presented for assessment each year while awaiting release to market under modern regulatory workflows (European Commission, 2007; Pool and Rusch, 2014). This amounts to increasing strain on regulatory processes, which must operate without a proportionate increase in resource availability. While providing and/or gathering relevant data for new chemicals now forms a vital part of risk assessment, advances in analytical techniques and scientific understanding continue to broaden the scope of this data beyond the realms of traditional *in vivo* toxicity testing. Although vital for compiling a more complete understanding of a chemical's toxicity, the broad scope and increasing availability of such data presents challenges for decision-makers tasked with handling, appraising and interpreting this data for risk assessment. Failure to have a transparent structure for considering all relevant data appropriate to risk assessment (e.g. a stepwise approach for addressing *in vitro* data following evidence from *in vivo* studies or comprehensive assessment of all *in vitro* data) reduces stakeholder confidence and has the potential to bias regulatory decisions. Studies reporting results amenable to the observer bias of independent assessors, or to the vested interests of non-independent assessors, may be cherry picked from the wider evidence base. Even where all relevant studies are considered, the role that scientific judgement plays in the process of appraisal and interpretation of data can lead to conflicting conclusions between different regulatory bodies (Whaley et al., 2016). Transparency in identifying both the evidence and scientific judgement are critical to establishing trust in decision-making.

Systematic review offers a framework for piecing together this varied data in a transparent and resource efficient manner, such that a more complete picture of toxicity can inform regulatory decision-making. It details methodology for ensuring all such data is identified, gathered and considered – preventing cherry picking of studies that only provide part of the complete toxicity profile for a chemical, or that present biased or unrepresentative results. As well as reducing bias, all steps of the methodology are designed to maximise transparency. A well conducted and reported systematic review effectively outlines the research question, the approach taken to address the question, the evidence considered, and the scientific judgement applied to reaching conclusions. Thus, differences across reviews or regulatory bodies can be effectively identified and explained. Considering the results of all relevant studies makes maximum use of existing data and increases the precision of a systematic review's conclusions. This allows reliable decisions to be made without the commissioning of redundant and repetitive primary research, or conversely identifies specific knowledge gaps at which smart testing strategies can be focused.

Although the aim of systematic review (i.e. to transparently and robustly synthesise all available data in answer to a research question) aligns well with the needs of chemicals policy, conflicts between the practicalities associated with the methodology and those associated with regulatory frameworks hinder their wider uptake, and/or the production of reviews that are of sufficient quality to produce trustworthy results (Kelly et al., 2016; Marshall et al., 2018; Reynen et al., 2018). Key areas of conflict include the time and resource intensity of the systematic review process, the scope of the research questions addressed by the methodology, and the ease with which the output of a systematic review can be accessed, interpreted and updated. Further, the fluid and rapidly expanding nature of scientific research and the chemicals industry creates a constant and pressing need for evidence surveillance, such that regulators can keep pace of the growing body of scientific literature and update regulation accordingly. This challenge demands a responsive and living solution beyond the reach of current systematic review practice.

In this manuscript, we briefly outline systematic review methodology to illustrate its strengths and highlight the transferable barriers which have been suggested as preventing its wider uptake in other fields (Oliver and Dickson, 2016). We discuss how these difficulties may be addressed through the novel implementation of systematic evidence mapping in environmental health. Systematic evidence maps (SEMs) provide a broad and comprehensive overview of an evidence base (Haddaway, Bernes, Jonsson, & Hedlund, 2016; James et al., 2016). They facilitate the identification of trends which can be used to inform more efficient systematic review, or more targeted primary research. The methodology behind SEMs, and how this might be adapted to suit the demands and limitations of regulatory decision-making in chemicals policy is discussed, along with the advantages and future potential of SEMs as a fundamental tool for evidence-informed risk management and decision-making.

2. The application of systematic review methods in chemical risk management

The utility and advantages of systematic review methods for advancing chemical risk assessment have been extensively documented elsewhere (Aiassa et al., 2015; Hoffmann et al., 2017; Hooijmans et al., 2012; Rooney et al., 2014; Vandenberg et al., 2016; Whaley et al., 2016; Woodruff and Sutton, 2014). Systematic review provides a transparent and reproducible approach to summarising and critically assessing existing evidence on potential health risks associated with exposure to a chemical substance. These transparent methods serve to document the basis of scientific judgments, minimising the potential for bias and error presented by more traditional narrative approaches in which opinion is not clearly distinguished from evidence.

The key features of a systematic review (Table 1) are:

- a clearly specified research objective - usually captured in a Population-Exposure-Comparator-Outcome (PECO) statement
- a comprehensive search strategy
- screening of the search results - for evidence relevant to addressing the research objective
- extraction of data from included studies - using a prespecified data extraction framework
- critical appraisal of included studies - according to a prespecified set of quality criteria, usually targeting risk of bias
- synthesis of findings from the included studies - using suitable quantitative statistical methods and otherwise qualitative methods as appropriate
- characterisation of confidence in the evidence for the results of the synthesis - according to a prespecified set of criteria
- statement of conclusions - including an assessment of limitations in design and conduct of the review itself.

Specific methodological decisions concerning each of these key features, from definition of the PECO statement to the chosen synthesis approach, are specified in a pre-published protocol.

However, with the methodology's pursuit of rigor and comprehensiveness comes a significant demand for time and resources. Evidence from medical systematic reviews indicates it takes on average approximately 70 weeks to progress a systematic review from protocol registration in the PROSPERO registry (National Institute for Health Research, 2018) to publication of the final systematic review (Borah et al., 2017). Variance around this average is wide (from 6 to 186 weeks), but the significance of person-hours and planning time prior to protocol registration is not considered in these estimates. More recent analysis of environmental science systematic reviews estimates an average of 164 (full time equivalent) person-days required for completion of systematic reviews (Haddaway and Westgate, 2018). However, in the absence of comparable evidence in the field of chemical risk assessment, these figures agree with anecdotal reports of the

Table 1
The key features of systematic reviews and their primary advantages. PECO = Population-Exposure-Comparator-Outcome.

| Systematic review step | Primary advantages |
|--|--|
| Pre-published protocol | Reduces risk that expectation bias will influence reviewers' choice of methods and approaches for analysis mid-review; if formally published, external peer review can reduce risk of limitations in planned methods from compromising final results. |
| Statement of objectives | Provides a structured framework for the aims of the review (including specific statement of the research question and PECO criteria) against which appropriate review methods can be defined. |
| Comprehensive search | Reduces risk of only partial retrieval of the overall body of evidence that is relevant to answering the research question. |
| Screening against eligibility criteria (study inclusion) | Reduces risk of only partial retrieval of the overall body of evidence that is relevant to answering the research question, in particular the risk of selection bias when reviewers are deciding which evidence to include in the review. |
| Data extraction using appropriate extraction tools | Reduces risk of inconsistent or partial retrieval of data from studies included in the review, reducing risk of selective use of data from studies deemed relevant to answering the research question. |
| Critical appraisal of included studies | Encourages consistent assessment of validity of included studies according to factors internal to study design, reducing risk of expectation bias or other factors causing studies to be inappropriately weighted, and helping ensure that bias in the findings of the included studies is not transmitted through to the findings of the review. |
| Synthesis of included studies | Pooling or integration of sufficiently comparable studies increases the power of an analysis, whether quantitative or qualitative, allowing overall trends in results to be more reliably identified. |
| Characterisation of confidence in the evidence | Encourages consistent assessment of the validity of the results of the synthesis according to features which manifest at the level of body of evidence as a whole rather than the individual study. Outlining the scientific judgement applied in rating confidence is key to the transparency of subsequent conclusions. |
| Drawing conclusions/key review output | Qualitative and/or quantitative summary effect estimates help direct policy decisions based on permissible exposure levels and related controls; assessment of limitations in the review methods helps ensure that any residual potential biases in the review are made clear to the reader and can additionally be accounted for in uncertainty assessment and consequent risk management action. |

average systematic review taking around 12 to 18 months to progress from inception to publication. A significant factor which contributes to the length of the systematic review process is the manual way in which each step of the methodology is conducted. All studies returned by a systematic search strategy are generally screened by human reviewers, in duplicate, one-by-one, before included studies undergo a similarly manual data extraction and critical appraisal step.

Systematic review management software has been developed (e.g. "HAWC: Health Assessment Workplace Collaborative," 2013; Covidence, 2019; Evidence Partners, 2019; Science for Nature and People Partnership Evidence-Based Conservation working group, Conservation International, Datakind, 2018; Sciome, 2018; Thomas et al., 2010; CAMARADES-NC3Rs, 2019) to assist human reviewers with maintaining transparency in SRs and with organising the review process. Acknowledging the impedance caused by a review's manual workload, review management software is beginning to incorporate machine learning as a means of automating labour-intensive tasks (e.g. Evidence Partners, 2019; Science for Nature and People Partnership Evidence-Based Conservation working group, Conservation International, Datakind, 2018; Sciome, 2018; CAMARADES-NC3Rs, 2019). Automation has the potential to result in significantly reduced workloads and subsequent demands for time and resources (Mara-eves et al., 2015). Pending further advances, the time and resource demands of systematic review are at conflict with the intense time/resource pressure under which regulatory processes must operate (Innvaer et al., 2002; Oliver and Dickson, 2016).

Also at conflict with the demands of regulatory decision-making is the narrow scope of systematic reviews, which are designed to address a specific and clearly defined objective or research question. To ensure a manageable, relevant and focused review, suitable research questions are typically closed framed, such that the review can synthesise a single, coherent answer. These closed-framed questions are well suited to the decision-making contexts of medicine (the field from which systematic reviews originate), but may be difficult to apply to chemical risk assessment. The web of interlinked endpoints, potential variation in sensitive populations, uncharacterised low dose effects, and unknown behaviour of a chemical in the environment or in contact with other chemicals can mean that the decision-critical information which can be supplied by a tightly focused research question is often not readily apparent in chemical risk assessment contexts. Even where such a question can be devised, and the answer reached through systematic review, the specificity of the research problem and its resolution are

likely to comprise only part of the much broader range of unaddressed decisions and information requirements faced by risk managers.

3. Systematic evidence maps for chemical risk management

In light of the time and resource intensity of current systematic review practice, identifying the most informative research questions is important for maximising the value and efficiency of systematic reviews in regulatory decision-making. Investing resources in systematic review as a means of addressing specific research questions is inefficient if there is a lack of data available for answering those questions. Devising specific research questions therefore becomes a reactive process, rather than a proactive one. This is at odds with the goals of chemicals policy, which aims to predict and prevent harm as a result of exposure to chemical substances.

Decision-makers therefore need to monitor and understand the evidence base as a whole – such that emerging trends or issues of potential concern can be identified and investigated in a timely manner. Identifying trends in the evidence base, including evidence clusters and evidence gaps, facilitates the formulation of proactive research questions by relevant stakeholders. Reviewers need not rely on environmental health outcomes becoming infamous or epidemic as an indicator of sufficient evidence for an efficient and valuable synthesis. Instead, trends in the availability of evidence ensure prevention of synthesis attempts for which there is insufficient data (or for which syntheses already exist) and promote the targeting of primary research efforts at evidence gaps. This kind of evidence surveillance has traditionally been the domain of scoping reviews. These reviews are often narrowly focused precursors to systematic reviews. Thus a specific systematic review question has already begun to be framed, and the literature scoped for sufficient data to address/focus it – rather than vice versa (e.g. Bolden et al., 2017). Scoping reviews also typically present their findings in tabular format. This compromises the accessibility of the evidence they scope, and makes them ill-suited for applications beyond determining whether there is sufficient literature to merit a systematic review (Grant and Booth, 2009).

Instead, the introduction of systematic evidence mapping, a methodology recently adapted from the social sciences (Clapton et al., 2009) for environmental management (James et al., 2016), has the potential to facilitate evidence surveillance in a transparent and reproducible manner, providing a broader understanding of the extant evidence base through interactive outputs.

Table 2
A comparison of systematic review and systematic evidence mapping methodology and their respective roles in risk management decision-making (adapted from James et al., 2016). SR = systematic review, SEM = systematic evidence map, RM = risk management, TDI = tolerable daily intake.

| Step | Conduct of step in SRs related to assessing chemical health risks | Conduct of step in SEMs related to assessing chemical health risks | SR vs SEM for responding to risk management needs |
|--|---|--|---|
| Pre-published protocol | Define all methods in advance of conduct of review | Same | Provides transparency; reduces bias; opportunity for peer review and stakeholder engagement. Applies to both SRs and SEMs. |
| Statement of objectives | Question concerns the effect of an exposure on health; or the effect of intervening to reduce exposure in terms of health benefit. Usually targets a single or few exposures and outcomes. | Question concerns the state of the evidence base for a topic. Usually open-ended and encompassing a range of multiple related exposures and outcomes. | SR: Focused, closed questions of SRs best service specific RM decisions such as characterising specific health risks/TDIs. SEM: Open questions of SEMs best service scenarios in which evidence should be surveyed and scoped, such as problem identification and priority-setting. |
| Comprehensive search | Search terms highly resolved and specified for most key elements of the objective statement, returning a moderate volume of evidence. | Wide ranging search strings of lower specificity based on topic rather than defining all key elements of the objective in the search. | SR: Narrow searches efficiently identify evidence related to exposure-outcome pairs. Maximum feasible number of sources searched to ensure collation of all relevant evidence for synthesis. SEM: Broader, topic-based SEM search allows evidence supportive of multiple decision scenarios to be identified. Flexible number of sources searched, or sources searched in a step-wise manner as appropriate to broader research objectives. |
| Screening against eligibility criteria (study inclusion) | Inclusion criteria specified in detail for all key elements of the objective. | Inclusion criteria defined in terms of topic rather than key elements of the objective. | SR: As for search, specific inclusion criteria ensure SRs efficiently service a specific research question. SEM: Broad objectives ensure inclusion of evidence relating to multiple decision scenarios. |
| Data extraction using tested extraction sheets | Complete extraction of meta-data and study findings. | Extraction of meta-data; optional extraction of study findings and other study characteristics depending on SEM objectives. | SR: Data extraction determined by objectives. SEM: Data extraction more flexible and can respond to needs of risk management process to develop fit-for-purpose maps of varying degrees of comprehensiveness. |
| Coding of extracted data using controlled vocabularies | Coding facilitates grouping of included studies for synthesis/integration according to review objectives. Coding is closely related to review objectives and data extraction process, whereby narrow research question and PECO statement inherently define specific code applicable to raw extracted data. | Coding facilitates broad comparison of heterogeneous data across an evidence base. Broad map objectives necessitate extensive coding process, whereby specific code must be defined in a step distinct from the formulation of end-users' specific research questions. | SR: Tight review objectives pre-specify applied code (e.g. considering ages 0-18 as 'Child' for reviews focusing on a population of 'Children'). Narrower range, or greater specificity of controlled vocabulary terms applicable per item of extracted data. SEM: Code pre-specified where possible, but addition of new terms (which could not be accounted for <i>a priori</i>) considered flexible. Any one item of extracted data may be coded by multiple and variably resolved terms. Openly accessible ontologies may be used for coding to promote consistency and interoperability. |
| Critical appraisal of included studies | Assessment of internal validity (risk of bias) conducted for all included studies. | Study validity assessment is optional and to some extent restricted if outcome is not a defined aspect of the SEM; study characteristics relevant to risk of bias assessment can be extracted. | SR: Describe the internal validity of the evidence base, which is an essential step of characterising confidence in the evidence. SEM: Flexible, critical appraisal step can be omitted; study methods are mapped or methodological quality assessed to goals, can be part of stepwise approach where quality only assessed for studies addressing key outcomes etc. |
| Synthesis of included studies | Quantitative synthesis where possible to produce characterisation of hazard from exposure; qualitative synthesis where pooling studies is not possible. | Reports of systematic maps can provide narrative synthesis of characteristics of the evidence key to a given decision-making context. | SR: Synthesis supports a specific type of decision context. SEM: Primary output is a more context-agnostic database which can be used by risk managers to support multiple decisions in the RM workflow; or to aid in a stepwise approach. |
| Characterisation of confidence in the evidence | Assessment of confidence or certainty in the results of the synthesis, according to characteristics of the evidence base taken as a whole. | SEMs do not synthesise included studies. SEMs help identify regions of evidence with characteristics indicative of being worth further, detailed analysis in support of a prospective decision. | SR: Provide detailed conclusions on certainty of evidence in hazard characterisation or to support risk assessments. SEM: Support a range of decisions, particularly decisions to focus research and review, e.g. indicating clusters where evidence may be strong enough to warrant SR (e.g. have a reasonable likelihood of changing a TDI), fill in gaps to reduce uncertainty and for surveillance. |
| Drawing conclusions/key review outputs | SRs primarily provide a summary effect estimate and surrounding uncertainty based on strength of the evidence and review methods. | SEMs primarily provide a searchable database of the characteristics of the evidence base, making the knowledge base locked away in manuscripts accessible to decision-makers. | SR: provide a qualitative and/or quantitative summary effect estimate in answer to a narrow and specific decision-making question. SEM: identify evidence gluts for synthesis. When combined with an understanding of RM needs, transparent criteria for prioritization of gluts for synthesis and gaps for commissioning primary research can be presented. |

The methodological steps involved in constructing a systematic evidence map are similar to those involved in the initial stages of producing a systematic review (see Table 2, adapted from James et al., 2016) whereby a systematic search strategy is employed to collate evidence, which is subsequently screened for relevance before undergoing data extraction. The key difference between the methodologies comes in the form of their aims and subsequent outputs. Systematic reviews collate a relatively narrow subset of the evidence base to answer a specific research question. Conversely, SEMs do not attempt to answer a specific, closed-framed research question, and are instead guided by much broader research objectives. SEMs collate a sufficiently broad subset of evidence such that many different specific research questions might be formulated from, and addressed with, a single systematic evidence map. SEMs are concerned with characterising the evidence base within a given research area, such that the availability, type and features of the evidence can be clearly mapped and explored through data visualization.

To facilitate this exploration, the output of a SEM takes the form of a queryable database (Clapton et al., 2009; James et al., 2016) as opposed to the lengthy and technical documents which form the main output of a systematic review. The database format allows users to query the evidence base according to their research interests, providing functionality which is void from systematic review documents and their associated static data tables. This format addresses the inability of systematic evidence mappers to predict what the specific research interests of users might be by providing the option to search for, and select, the specific subsets of data relevant to a particular use case.

Whereas systematic reviews present users with select information from included studies (i.e. data relevant to addressing the research question), SEMs aim to extract a broader range of data from included studies and aim to maintain the native format of these data. In this sense, the search and screening process are the steps of SEM methodology most affected by its research objective or context, as the focus of data extraction remains broad regardless. This is in contrast to systematic review, where all steps are heavily influenced by its research question. The data extracted for inclusion in a SEM database can then be flexibly categorised, or “coded” to facilitate comparison of an otherwise heterogeneous evidence base.

Resolution of coding can be adapted to suit the needs of regulators. For example, coding the species under investigation in a study might use categories such as “Sprague-Dawley”, “Rat”, “Rodent” or “Mammal”; or may use all of these categories such that the data can be interrogated in successively deeper levels of detail. As well as facilitating variably resolved interrogation of the evidence base, coding plays a significant role in systematic mapping’s amenability to updating. Use of universal, standardised ontologies for coding, such as the Unified Medical Language System (UMLS) (U.S. National Library of Medicine, 2016), offers a degree of consistency that future users can readily exploit when updating a map (Baker et al., 2018). These ontologies also offer interoperability between SEMs, creating the potential to expand and merge evidence maps – a feature likely to become increasingly attractive as the scope of evidence relevant to assessing toxicity grows along with our understanding of its interconnectedness.

In current practice it is common to present users with SEMs that house only coded information for simplicity and ease of access (e.g. Papathanasopoulou et al., 2016). However, this conflates data extraction with coding. Maintaining the native format of extracted data and applying coding on top of this therefore ensures maximum transparency in SEMs. This additionally promotes the ease with which a map can be updated as advancing scientific understanding calls for coding categories to be redefined. As with systematic reviews, the data extraction and coding steps of a SEM represent a manual workload. Presenting only coded data may offer a saving in the resource intensity of the process. However, in maintaining a transparent link between raw extracted data and the code used to categorise it, SEMs offer a gateway to automation – whereby controlled vocabulary ontologies can be used to

train machine learning algorithms to automatically identify, extract and code data from the literature.

Pending such advances, the time required to conduct a fit for purpose systematic map in environmental health is uncharacterised. Evidence from the wider environmental sciences (Haddaway and Westgate, 2018) suggests that (on average) systematic maps take longer to complete than systematic reviews. This is due to the generally larger number of studies they manually collate, screen and extract data from. While maps might present a larger upfront cost in terms of time, their multipurpose nature has the potential to offer more long-term resource savings compared to exclusively conducting systematic reviews. This is because a single systematic evidence map may continue to be useful to several different aspects of the regulatory workflow (see Sections 4 and 5 below).

As the purpose of a SEM is to characterise the evidence base, there is no risk of allocating resources to the production of an inconclusive output, as is the case for “empty” systematic reviews (systematic reviews which ask research questions for which there is too little included evidence for them to reach a conclusion or be supportive of a decision). In fact, systematic evidence maps may reduce the resource strain associated with systematic reviews. A SEM’s broad overview of the evidence base allows fast identification of topics for which there is sufficient data to warrant a full systematic review. The SEM itself, if conducted to sufficiently rigorous standards, can even replace the literature search and screening process of a systematic review. As SEMs present all available relevant evidence on a broader topic such as the “health effects of bisphenol-A” (obtained through a systematic but less specific search strategy), filtering this information according to the PECO statement of a systematic review may act in an equivalent manner to approaching the literature with a more focused search strategy in the first instance. The pre-screened nature of this subset is likely to reduce the number of false positive results, facilitating faster syntheses.

As advances in machine learning facilitate more highly resolved data extraction processes, future SEMs may even store enough detail for them to form the basis of meta-analytical syntheses. If all data contained within study reports is extracted and indexed within a SEM, there would be no data required specifically for syntheses which could not be found in the SEM. This would allow SEMs to form the dataset on which meta-analytical and predictive toxicological models are based, the results of which may additionally be incorporated into the SEM itself – facilitating more transparent, resource-efficient and easily updated syntheses.

4. Exploring the evidence base with SEMs

Systematic evidence mapping facilitates identification of trends which are informative for many risk management scenarios. To illustrate the flexibility and potential utility of SEMs’ trendspotting capacity, this section highlights the type of data visualization and exploration possible through querying subsets of information in a SEM database. Specifically, “priority setting” (National Academy of Sciences, 1983; Pool and Rusch, 2014), the process by which regulators identify the most pressing chemical substances for assessment and regulation (e.g. from a pool of unassessed legacy chemicals) is presented as context for the exploration of a hypothetical SEM.

Several factors are relevant to prioritizing individual chemicals for assessment, broadly ranging from recorded levels of exposure to evidence for toxicity. Underlying these broad considerations are several more specific factors such as the bio-accessibility of the chemical, the relevance of its toxicity evidence for predicting health risks in human populations etc. In order to make the most efficient use of resources and the systematic review process, decision-makers require access to a means of comparing these features to justify prioritization of a particular chemical for review/risk assessment.

This is the role of a SEM, which may be constructed with the aim of

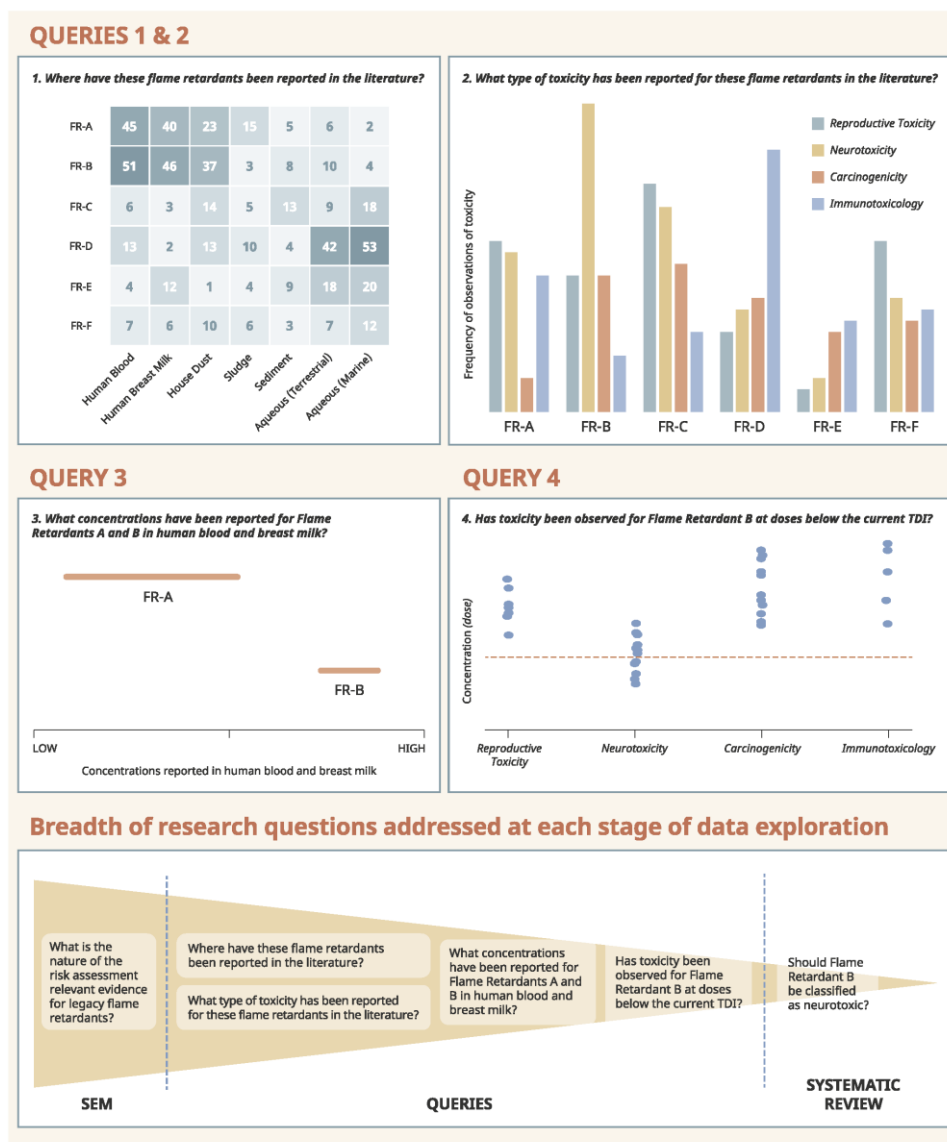


Fig. 1. The process of identifying trends and exploring the evidence landscape involves querying the SEM database and visualizing the results of the query. Queries may start by asking broader questions which consider a wider range and volume of data (e.g. Queries 1 and 2). Users may then further explore any trends of interest discovered in the results of these broad queries by running narrower queries which consider a more specific subset of data (e.g. Queries 3 and 4). Data displayed in this Figure have been artificially generated to illustrate a hypothetical use case for SEMs. FR = flame retardant, TDI = tolerable daily intake, SEM = systematic evidence map.

identifying and characterising the risk assessment relevant evidence for a broader group of legacy chemicals, e.g. flame retardants. Once data has been extracted and coded from the literature, the SEM can be explored with a succession of queries of increasingly narrow focus, each considering a narrower subset of the evidence base than the last, such that a research question appropriate for more detailed synthesis is resolved at the end of a process which begins with a very broad research objective. This is illustrated in Fig. 1 using the hypothetical context of priority setting with a group of arbitrary chemicals, in this case flame retardants (FRs) A–F.

Queries 1 and 2 depicted in Fig. 1 explore the frequency with which the literature observes a flame retardant in a coded location category (e.g. human blood, human breast milk, house dust, etc.) and the frequency with which the literature observes an association between a flame retardant and a coded toxicity category (e.g. reproductive toxicity, neurotoxicity etc.). The heatmap visualizing the results of Query 1 shows a comparatively large number of observations of FRs A and B in location categories directly relevant to human populations (i.e. human blood and breast milk). Query 2 clarifies whether these observations require further attention by indicating what kind of toxicity information is available for each flame retardant. The bar chart visualization indicates comparable numbers of observations for most of the flame retardants and types of toxicity but a comparatively large number of observations that associate FR B with neurotoxicity.

Based on (hypothetical) existing evidence, Queries 1 and 2 indicate flame retardants A and B as potential candidates for full assessment. Resolving which to prioritize involves accessing more study-specific information through a series of queries which consider a successively narrow subset of the evidence base. Despite availability of toxicity data, observing flame retardants in human relevant locations might not be concerning if the concentrations observed are negligible. Thus Query 3 examines the range of concentrations reported in the literature for FRs A and B in human blood and breast milk. Visualization of Query 3 indicates a wider range of lower concentrations reported for FR A, compared to a narrower range of higher concentrations for FR B. Query 4 then examines the relevance of these concentrations against the current estimated tolerable daily intake (TDI) for FR B, indicating several observations of toxicity below the current TDI and supporting prioritization of FR B for assessment. Further, the relatively large volume of observations of neurotoxicity may indicate sufficient data available to conduct a systematic review on FR B's relationship with neurotoxicity.

However, it is important to distinguish the results of SEM queries from synthesis. SEMs only present what has been studied in the literature – they cannot present what has not been studied, and do not always assess the risk of bias of the findings they report. Thus, while a high number of observations of flame retardants A and B in human relevant locations is a valid trend to explore further, it does not necessarily mean that there are fewer of the other flame retardants present in human relevant locations, but rather that there may simply be fewer of these flame retardants studied at all. Identification of such evidence gaps is equally valid for focusing primary research. For example, the relatively high number of observations of reproductive toxicity for FR F, but comparatively low number of observations of this flame retardant in any exposure locations might warrant re-analysis of samples or new exposure studies to verify whether exposure to this substance is of concern.

The SEM is also sufficiently flexible that different trends can be investigated, and different research questions formulated, based on the priorities of regulators. For example, the number of observations in the literature which found FR D in aquatic environments might spur further investigation into the ecotoxicity of this compound. A single SEM exercise therefore makes efficient use of resources in its potential to meet the varied needs of several end users.

5. The role of SEMs in wider risk management workflows

In addition to priority setting, SEMs have the potential to fill several roles within wider workflows.

5.1. Data gathering

Although evidence synthesis methodology can be considered costly in terms of time and resources, this cost can be dwarfed by the equivalent resource demands associated with conducting primary research relevant to assessing the hazards associated with exposure to a chemical, as illustrated with more established examples in the field of medicine (Glasziou et al., 2006). In an effort to manage these demands, reduce the production of research waste, and comply with principles such as the three Rs (European Chemicals Agency, 2018a, 2018b; National Centre for the Replacement Refinement and Reduction of Animals in Research, 2018), a key first step in many regulatory workflows is the identification and gathering of all pre-existing evidence relevant to a specific risk management decision. This can be illustrated in regulatory frameworks such as the European Union's REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) initiative, which requires registrants to make an attempt to identify all available, pre-existing evidence on the hazards associated with the chemical substance under registration (European Chemicals Agency, 2018a, 2018b). Similarly, REACH imposes a “one substance, one registration” policy, whereby all parties with an interest in registration of a substance must share data, minimising repeat testing. Although promoted in guidance documents (European Chemicals Agency, 2016), a lack of a sufficiently robust methodology for finding, collating, housing and reporting these data leads to poor transparency, and therefore does not remove the potential for cherry picking of key studies which may not be representative of the evidence base as a whole.

SEMs have the potential to provide this much needed transparency. The nature of a SEM's output being a collection of relevant search results, and specific information coded from those results, introduces a greater level of accountability for registrants. Studies are identified by registrants as “key”, “supporting” etc. based on the perceived relevance, adequacy and reliability of the evidence they provide for a specific endpoint, assessed using “sound scientific judgement” (European Chemicals Agency, 2011). These assignments are aided by application of the Klimisch criteria (Klimisch et al., 1997) – a rating methodology criticised for its lack of transparency and failure to consider non-industry sources of evidence (Ingre-Khans et al., 2019). This poor transparency hinders the appraisal of registrants' choices (e.g. of key study), and the degree to which those choices can be considered representative of the wider evidence base. Using SEM methodology alleviates this issue by requiring registrants to clearly document the efforts of their search and screening process, constructing a database of the pool of evidence considered in their evaluations. Additionally, applying code to the specific extracted study features which influence a decision to assign a study as “key”, “supporting”, “weight-of-evidence” etc. serves to document the basis for these decisions in a structured and queryable way. As registrants submit SEMs at the level of single substances, these efforts can be merged to build a SEM that spans all registered substances. This facilitates appraisal of registrants' choices of key study in the context of the wider evidence base. The ability to explore trends in the features influencing assignment of key studies may even assist in refining and improving the registration process – as emerging issues or shortcomings can be quickly evidenced.

5.2. Problem formulation

Beyond offering improvements in transparency during the data gathering phase, SEMs may be of particular value to the problem formulation stage of regulatory decision-making. Problem formulation is a prerequisite to conducting a chemical risk assessment, identifying an

issue of regulatory relevance around which the assessment will be focused (Solomon et al., 2016). These issues can be subtle and difficult to identify at a sufficiently early stage in the field of environmental health, putting the problem formulation process at risk of focusing on issues of lower severity or significance. In implementing a SEM with a broad (lower resolution) coding process, but with a key focus on the hierarchy of coded data and the nature in which this data is related, trends in the evidence base can be effectively and efficiently identified. This allows risk assessors to use these broad, coded parameters to reliably identify problems in need of further assessment, either through secondary syntheses (if the SEM presents a sufficiently large evidence cluster) or primary research (if the SEM indicates an evidence gap).

5.3. Read-across

Identifying trends in the evidence base may also play a significant role in read-across applications. Read-across allows the toxicologically relevant properties of a chemical to be inferred by comparison with a structurally similar chemical of known toxicological behaviour (European Chemicals Agency, 2017a). Read-across aligns well with the need to make best use of existing evidence (van Leeuwen et al., 2009), and the storage of data in a related manner within a SEM could allow the identification of appropriate read-across scenarios. In filtering an evidence map by outcome features, exposures which behave in a similar manner can be identified and investigated further for chemical similarity and/or shared modes of action. This information can be used to group substances, such that data-rich members of the group can be used to make predictions about data-poor members, without pursuing further primary research (Vink et al., 2010). Conversely, filtering an evidence map by chemical group or structural similarity may allow identification of shared outcomes, of similar relevance to read-across applications.

5.4. Evidence surveillance

Once regulation is in place, it is vital that it is kept up to date. Such is the role of the ongoing, evidence surveillance phase of regulatory decision-making. Within REACH, registrants are required to update their registration dossiers "whenever new information is available" (European Chemicals Agency, 2017b), such that dossiers are living products. However, a report commissioned by the European Chemicals Agency (ECHA) found that 64% of REACH registration dossiers submitted to ECHA since 2008 have never been updated (Amec Foster Wheeler Environment and Infrastructure UK Limited, 2017). The report details several obstacles experienced by registrants faced with updating dossiers, including technical difficulties, issues of ownership or responsibility for updates among co- and lead registrants, the potentially labour-intensive nature of updating dossiers and a perception of REACH registration being the "end of a process".

Openly accessible and easily updated SEMs may serve to address such obstacles. As the population of a SEM database does not require detailed analysis or complex interpretation of the raw data, SEMs could be amenable to automation. Technological advances in text-mining and artificial intelligence might assist the automatic screening, extraction and coding of new information as it is published, based on the data fields and coding ontologies used to populate the original SEM. Although some years away from implementation, application of SEM methodology in the interim will promote fast uptake of such technological advances.

6. Conclusion

Systematic evidence mapping presents a transparent and robust methodological framework with which to assess the evidence landscape at the level of individual chemical risk management and innovation, to regulatory decision-making in chemicals policy. The broad scope of

SEMs lowers the barrier to evidence synthesis in chemical risk assessment through more efficient use of resources. Future developments in text mining and machine learning are likely to further reduce the resource intensity of the methodology, and of chemical risk assessment in general. These advances will enable the automatic production of highly resolved SEMs capable of synthesising evidence or feeding predictive models.

In the interim pursuit of a more evidence-based approach to chemicals policy, the resource strain associated with producing a SEM can be managed through adaptation of the methodology to present day limitations. Depending on the needs of the user and the constraints of their use case, SEM methodology is sufficiently flexible that it may be adapted (e.g. by searching fewer databases, extracting data based on only title/abstract etc.) without compromising the utility of the end product in the same way as the results of a synthesis might be adversely affected by modification of systematic review methodology. By working closely with stakeholders to define objectives, the scope of the SEM (i.e. bibliographic databases covered, types of studies included, etc.) can be adjusted as appropriate to objectives. For example, critical appraisal of studies may not be imperative to the aim of the SEM and may therefore be omitted or might be planned as part of a stepwise approach after the SEM identifies pockets of evidence of interest to stakeholders. Although designed to reduce the resource strain of SEM exercises, such flexible adaptation of the methodology does not compromise the fitness-for-purpose of SEMs as a means of identifying and comparing trends in the availability of evidence in a vast and heterogeneous information landscape.

Consequently, examples of research activities producing fit-for-purpose SEM outputs and/or developing aspects of SEM methodology specific to chemicals policy contexts are beginning to emerge (Beverly, 2019), with research institutes such as NTP-OHAT and The Endocrine Disruption Exchange (TEDX) conducting evidence mapping activities (NTP-OHAT, 2019; The Endocrine Disruption Exchange, 2019). A key consideration for these emerging efforts is the accessibility of SEMs' queryable output for non-technical audiences. To this end, researchers have made use of a variety of readily available and user-friendly tools (e.g. Datawrapper GmbH, 2019; IBM, 2019; QlikTech International AB, 2019; Tableau Software, 2019 etc.) to facilitate visualization of, and promote interaction with, the data collated in evidence surveillance exercises (e.g. Pelch et al., 2019; Walker et al., 2018). These tools may similarly serve to lower the barrier to accessing (as well as producing) SEMs, provided the underlying database is made available for more specialist users. Although future technological advances will have significant implications for the production and use of SEMs, these efforts indicate how SEM methodology can be effectively applied in present day, highlighting how SEMs can be adapted for engaging with a variety of stakeholders. More immediate establishment of (adapted) SEM infrastructure in current regulatory workflows will therefore not only lower resource barriers to evidence-based decision-making, but will ensure that technological advances in automation, and in SEM methodology itself, can be readily exploited by regulatory decision-makers in chemicals risk management.

Funding sources

TW's PhD is financially supported by the Centre for Global Eco-Innovation (a programme funded by the European Regional Development Fund) and Yordas Group, a global consultancy in the area of chemical safety, regulations and sustainability. PW's contribution to the manuscript was funded by the Evidence-Based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health. AR and VW were supported by the National Institute of Environmental Health Sciences, Division of the National Toxicology Program. The authors declare they have no actual or potential competing financial interests.

Author contributions

TW, PW, CH, AR and VW established the principal ideas for the manuscript and developed an outline. TW wrote the first draft of the manuscript following further discussion of the concept with PW and CH. All authors reviewed and edited the manuscript and contributed to its development.

Acknowledgements

The authors would like to acknowledge the input of the two internal reviewers at the National Toxicology Program, Kristen Ryan and Scott Masten, who provided suggestions for improving the manuscript prior to submission. The authors would also like to thank Miriam Sturdee for help with re-formatting Fig. 1.

References

- Aiassa, E., Higgins, J.P.T., Frampton, G.K., Greiner, M., Afonso, A., Anzal, B., Verloo, D., 2015. Applicability and feasibility of systematic review for performing evidence-based risk assessment in food and feed safety. *Crit. Rev. Food Sci. Nutr.* 55 (7), 1026–1034. <https://doi.org/10.1080/10408398.2013.769933>.
- Amezc Foster Wheeler Environment & Infrastructure UK Limited, 2017. A Study to Gather Insights on the Drivers, Barriers, Costs and Benefits for Updating REACH Registration and CLP Notification Dossiers.
- Baker, N., Boobis, A., Burgoon, L., Carney, E., Currie, R., Fritsche, E., Daston, G., 2018. Building a developmental toxicity ontology. *Birth Defects Res.* 110 (6), 502–518. <https://doi.org/10.1002/bdr2.1189>.
- Beverly, B., 2019. Abstract 3267: Potential Alternatives to Systematic Review: Evidence Maps and Scoping Reviews. Retrieved from <https://www.toxicology.org/events/am/AM2019/program-details.asp>.
- Bolden, A.L., Rochester, J.R., Schultz, K., Kwiatkowski, C.F., 2017. Polycyclic aromatic hydrocarbons and female reproductive health: a scoping review. *Reprod. Toxicol.* 73, 61–74. <https://doi.org/10.1016/j.reprotox.2017.07.012>.
- Borah, R., Brown, A.W., Capers, P.L., Katsir, K.A., 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 7 (2), 1–7. <https://doi.org/10.1136/bmjopen-2016-012545>.
- CAMARADES-NC3RS, 2019. Preclinical Systematic Review & Meta-analysis Facility (SYRF). Retrieved February 6, 2019, from <http://syrf.org.uk/>.
- Chalmers, L., Hedges, L., Cooper, H., 2002. A brief history of research synthesis. *Eval. Health Prof.* 25 (1), 12–37. <https://doi.org/10.1177/0163278702025001003>.
- Clapton, J., Rutter, D., Sharif, N., 2009. SCIE Systematic Mapping Guidance April 2009 MG. Retrieved from <https://www.scie.org.uk/publications/researchresources/rro3.pdf>.
- Covidence, 2019. Covidence. Retrieved January 14, 2019, from <https://www.covidence.org/home>.
- Datawrapper GmbH, 2019. Datawrapper. Retrieved from <https://www.datawrapper.de/>.
- EPA, 2018. Application of Systematic Review in TscA Risk Evaluations, 1–247. Retrieved from https://www.epa.gov/sites/production/files/2018-06/documents/final_application_of_sr_in_tsc_a_05-31-18.pdf.
- European Chemicals Agency, 2011. Guidance on Information Requirements and Chemical Safety Assessment Chapter R. 4: Evaluation of Available Information. Retrieved from https://echa.europa.eu/documents/10162/13643/information_requirements_r4_en.pdf.
- European Chemicals Agency, 2016. Guidance on registration. <https://doi.org/ECHA-12-G-07-EN>.
- European Chemicals Agency, 2017a. Guidance on information requirements and chemical safety assessment: QSARs and grouping of chemicals. In: *Guidance for the Implementation of REACH*, pp. 134. <https://doi.org/10.2823/43472>.
- European Chemicals Agency, 2017b. Study finds companies lack incentives for updating their REACH registrations. Retrieved November 8, 2018, from <https://echa.europa.eu/-/study-finds-companies-lack-incentives-for-updating-their-reach-registrations>.
- European Chemicals Agency, 2018a. Information requirements. Retrieved November 9, 2018, from <https://echa.europa.eu/regulations/reach/registration/information-requirements>.
- European Chemicals Agency, 2018b. Strategy for gathering your data. Retrieved November 9, 2018, from <https://echa.europa.eu/support/registration/strategy-for-gathering-your-data>.
- European Commission, 2007. REACH in Brief. <https://doi.org/10.1067/j.cpsup.2007.09.005>.
- Evidence Partners, 2019. DistillerSR. Retrieved January 14, 2019, from <https://www.evidencepartners.com/products/distillers-systematic-review-software/>.
- Garg, A.K., Hackam, D., Tonelli, M., 2008. Systematic review and meta-analysis: when one study is just not enough. *Clin. J. Am. Soc. Nephrol.* 3 (1), 253–260. <https://doi.org/10.2215/CJN.01430307>.
- Glasziou, P., Djibegovic, B., Burls, A., 2006. Are systematic reviews more cost-effective than randomised trials? *Lancet* 367 (9528), 2057–2058. [https://doi.org/10.1016/S0140-6736\(06\)68919-8](https://doi.org/10.1016/S0140-6736(06)68919-8).

- Grant, M.J., Booth, A., 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Inf. Libr. J.* 26 (2), 91–108. <https://doi.org/10.1111/j.1471-1842.2009.00848.x>.
- Haddaway, N.R., Westgate, M.J., 2018. Predicting the time needed for environmental systematic reviews and systematic maps. *Conserv. Biol.* 0 (0), 1–10. <https://doi.org/10.1111/cobi.13231>.
- Haddaway, N.R., Bernes, C., Jonsson, B.-G., Hedlund, K., 2016. The benefits of systematic mapping to evidence-based environmental management. *Ambio* 45 (5), 613–620. <https://doi.org/10.1007/s13280-016-0773-x>.
- HAWC Project, 2013. HAWC: Health Assessment Workplace Collaborative. Retrieved from <https://hawcproject.org/>.
- Higgins, J.P.T., Green, S., 2011. *Cochrane Handbook for Systematic Reviews of Interventions* (editors) Retrieved from <http://handbook-5-1.cochrane.org/>.
- Hoffmann, S., de Vries, R.B.M., Stephens, M.L., Beck, N.B., Dirven, H.A.A.M., Fowle, J.R., Goodman, J.E., Hartung, T., Kimber, L., Lallu, M.M., Thayer, K., Whaley, P., Wikoff, D., Tsalou, K., 2017. A primer on systematic reviews in toxicology. *Arch. Toxicol.* 91 (7), 2551–2575. <https://doi.org/10.1007/s00204-017-1980-3>.
- Hooijmans, C.R., Rovers, M., de Vries, R.R., Leenaars, M., Ritskes-Hoitinga, M., 2012. An initiative to facilitate well-informed decision-making in laboratory animal research: report of the First International Symposium on Systematic Reviews in Laboratory Animal Science. *Lab. Anim.* 46 (4), 356–357. <https://doi.org/10.1258/la.2012.012052>.
- IBM, (2019). Cognos Analytics. Retrieved from <https://www.ibm.com/analytics/cognos-analytics>.
- Ingre-Khans, E., Ågerstrand, M., Beronius, A., Rudén, C., 2019. Reliability and relevance evaluations of REACH data. *Toxicol. Res.* 8 (1), 46–56. <https://doi.org/10.1039/c8tx000216a>.
- Invvaer, S., Vist, G., Trommald, M., Oxman, A., 2002. Review article health policy-makers' perceptions of their use of evidence: a systematic review. *J. Health Serv. Res.* 7 (4), 239–244. <https://doi.org/10.1258/1355819020437778>.
- James, K.L., Randall, N.P., Haddaway, N.R., 2016. A methodology for systematic mapping in environmental sciences. *Environ. Evid.* 5 (1), 7. <https://doi.org/10.1186/s13750-016-0059-6>.
- Kelly, S.E., Moher, D., Clifford, T.J., 2016. Quality of conduct and reporting in rapid reviews: an exploration of compliance with PRISMA and AMSTAR guidelines. *Syst. Rev.* 5 (1), 1–19. <https://doi.org/10.1186/s13643-016-0258-9>.
- Klimisch, H.-J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul. Toxicol. Pharmacol.* 25 (1), 1–5. <https://doi.org/10.1006/rtp.1996.1076>.
- Mandrioli, D., Schülssens, V., Ádám, B., Cohen, R. A., Colosio, C., Chen, W., ... Scheepers, P. T. J. (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. *Environ. Int.* 119(June), 174–185. <https://doi.org/10.1016/j.envint.2018.06.005>.
- Mara-eyes, A.O., Thomas, J., McNaught, J., Miwa, M., Ananiadou, S., 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst. Rev.* 1–22. <https://doi.org/10.1186/2046-4053-4-5>.
- Marshall, I., Marshall, R., Wallace, B., Brassey, J., Thomas, J., 2018. Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. *J. Clin. Epidemiol.* <https://doi.org/10.1016/j.jclinepi.2018.12.015>.
- Morgan, R.L., Thayer, K.A., Bero, L., Bruce, N., Falck-Ytter, Y., Ghersi, D., Schünemann, H.J., 2016. GRADE: assessing the quality of evidence in environmental and occupational health. *Environ. Int.* 92–93, 611–616. <https://doi.org/10.1016/j.envint.2016.01.004>.
- National Academy of Sciences, 1983. *Risk Assessment in the Federal Government: Managing the Process*. National Academy Press, Washington, DC. <https://doi.org/10.17226/366>.
- National Centre for the Replacement Refinement & Reduction of Animals in Research, 2018. The 3Rs. Retrieved November 8, 2018, from <https://www.nc3rs.org.uk/the-3rs>.
- National Institute for Health Research, 2018. PROSPERO - International Prospective Register of Systematic Reviews. Retrieved December 31, 2018, from <https://www.crd.york.ac.uk/prospero/>.
- NTP, 2015. Handbook for Conducting a Literature-based Health Assessment Using OHAT Approach for Systemic Review and Evidence Integration. pp. 1–98. Retrieved from https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf.
- NTP-OHAT, 2019. About the Office of Health Assessment and Translation. Retrieved from <http://ntp.niehs.nih.gov/>.
- Oliver, S., Dickson, K., 2016. Policy-relevant systematic reviews to strengthen health systems: models and mechanisms to support their production. *Evid. Policy* 12 (2), 235–259. <https://doi.org/10.1332/174426415X1439993605641>.
- Papathanasopoulou, E., Queirós, A.M., Beaumont, N., Hooper, T., Nunes, J., 2016. What evidence exists on the local impacts of energy systems on marine ecosystem services: a systematic map. *Environ. Evid.* 5 (1), 1–12. <https://doi.org/10.1186/s13750-016-0075-6>.
- Pelch, K.E., Bolden, A.L., Kwiatkowski, C.F., 2019. Environmental chemicals and autism: a scoping review of the human and animal research. *Environ. Health Perspect.* 127 (4), 46001. <https://doi.org/10.1289/EHP4386>.
- Pool, R., Rusch, E., 2014. Identifying and Reducing Environmental Health Risks of Chemicals in Our Society: Workshop Summary. The National Academies Press, Washington, DC.
- QlikTech International AB, 2019. Qlik Sense. Retrieved from <https://www.qlik.com/us/products/qlik-sense>.
- Reynen, E., Robson, B., Ivory, J., Hwee, J., Straus, S.E., Pham, B., Tricco, A.C., 2018. A retrospective comparison of systematic reviews with same-topic rapid reviews. *J. Clin. Epidemiol.* 96, 23–34. <https://doi.org/10.1016/j.jclinepi.2017.12.001>.

- Rooney, A.A., Boyles, A.L., Wolfe, M.S., Bucher, J.R., Thayer, K.A., 2014. Systematic review and evidence integration for literature-based environmental health science assessments. *Environ. Health Perspect.* 122 (7), 711–718. <https://doi.org/10.1289/ehp.1307972>.
- Science for Nature and People Partnership Evidence-Based Conservation working group, Conservation International, DataKind, 2018. Colandr. Retrieved January 14, 2019, from. <https://www.colandrapp.com/signin>.
- Sciome, 2018. SWIFT-Review. Retrieved January 14, 2019, from. <https://www.sciome.com/swift-review/>.
- Solomon, K.R., Wilks, M.F., Bachman, A., Boobis, A., Moretto, A., Pastoor, T.P., Embry, M.R., 2016. Problem formulation for risk assessment of combined exposures to chemicals and other stressors in humans. *Crit. Rev. Toxicol.* 46 (10), 835–844. <https://doi.org/10.1080/10408444.2016.1211617>.
- Tableau Software (2019). Tableau. Retrieved from <https://www.tableau.com/>
- The Endocrine Disruption Exchange. (2019). TEDX Publications. Retrieved from <https://endocrinedisruption.org/interactive-tools/publications/>
- The National Academies of Sciences, 2017. Application of Systematic Review Methods in an Overall Strategy for Evaluating Low-dose Toxicity From Endocrine Active Chemicals. Washington, DC: The National Academies Press. <https://doi.org/10.17226/24758>.
- Thomas, J., Brunton, J., Graziosi, S., 2010. EPPI-Reviewer 4.0: Software for Research Synthesis. London: Social Science Research Unit. Institute of Education, University of London. <http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3299#Research>.
- U.S. National Library of Medicine, 2016. Unified Medical Language System (UMLS). Retrieved December 31, 2018. In: from, . <https://www.nlm.nih.gov/research/umls/>.
- van Leeuwen, K., Schultz, T.W., Henry, T., Diderich, B., Veith, G.D., 2009. Using chemical categories to fill data gaps in hazard assessment. *SAR QSAR Environ. Res.* 20 (3–4), 207–220. <https://doi.org/10.1080/10629360902949179>.
- Vandenberg, L.N., Ågerstrand, M., Beronius, A., Beausoleil, C., Bergman, Å., Bero, L.A., ... Rüdén, C., 2016. A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. *Environ. Health* 15 (1), 1–19. <https://doi.org/10.1186/s12940-016-0156-6>.
- Vink, S.R., Mikkers, J., Bouwman, T., Marquart, H., Kroese, E.D., 2010. Use of read-across and tiered exposure assessment in risk assessment under REACH – a case study on a phase-in substance. *Regul. Toxicol. Pharmacol.* 58, 64–71. <https://doi.org/10.1016/j.yrtph.2010.04.004>.
- Walker, V.R., Boyles, A.L., Pelch, K.E., Holmgren, S.D., Shapiro, A.J., Blystone, C.R., Rooney, A.A., 2018. Human and animal evidence of potential transgenerational inheritance of health effects: an evidence map and state-of-the-science evaluation. *Environ. Int.* 115 (December 2017), 48–69. <https://doi.org/10.1016/j.envint.2017.12.032>.
- Whaley, P., Halsall, C., 2016. Assuring high-quality evidence reviews for chemical risk assessment: five lessons from guest editing the first environmental health journal special issue dedicated to systematic review. *Environ. Int.* 92–93, 553–555. <https://doi.org/10.1016/j.envint.2016.04.016>.
- Whaley, P., Halsall, C., Ågerstrand, M., Aiassa, E., Benford, D., Bilotta, G., ... Taylor, D., 2016. Implementing systematic review techniques in chemical risk assessment: challenges, opportunities and recommendations. *Environ. Int.* 92–93, 556–564. <https://doi.org/10.1016/j.envint.2015.11.002>.
- Woodruff, T.J., Sutton, P., 2014. The navigation guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. *Environ. Health Perspect.* 122 (10), 1007–1014. <https://doi.org/10.1289/ehp.1307175>.

Appendix D: Knowledge Graphs

This manuscript is available at this DOI: [10.1093/toxsci/kfaa025](https://doi.org/10.1093/toxsci/kfaa025)



SOT | Society of
Toxicology
academic.oup.com/toxsci

TOXICOLOGICAL SCIENCES, 2020, 35–49

doi: 10.1093/toxsci/kfaa025
Advance Access Publication Date: February 25, 2020
Research Article

A Survey of Systematic Evidence Mapping Practice and the Case for Knowledge Graphs in Environmental Health and Toxicology

Taylor A.M. Wolfe^{*,†,1}, John Vidler[‡], Crispin Halsall^{*}, Neil Hunt[†] and Paul Whaley^{*,§}

^{*}Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK; [†]Yordas Group, Lancaster University, Lancaster LA1 4YQ, UK; [‡]School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, UK; and [§]Evidence-Based Toxicology Collaboration, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205

¹To whom correspondence should be addressed at E-mail: t.wolfe@lancaster.ac.uk

ABSTRACT

Systematic evidence mapping offers a robust and transparent methodology for facilitating evidence-based approaches to decision-making in chemicals policy and wider environmental health (EH). Interest in the methodology is growing; however, its application in EH is still novel. To facilitate the production of effective systematic evidence maps for EH use cases, we survey the successful application of evidence mapping in other fields where the methodology is more established. Focusing on issues of “data storage technology,” “data integrity,” “data accessibility,” and “transparency,” we characterize current evidence mapping practice and critically review its potential value for EH contexts. We note that rigid, flat data tables and schema-first approaches dominate current mapping methods and highlight how this practice is ill-suited to the highly connected, heterogeneous, and complex nature of EH data. We propose this challenge is overcome by storing and structuring data as “knowledge graphs.” Knowledge graphs offer a flexible, schemaless, and scalable model for systematically mapping the EH literature. Associated technologies, such as ontologies, are well-suited to the long-term goals of systematic mapping methodology in promoting resource-efficient access to the wider EH evidence base. Several graph storage implementations are readily available, with a variety of proven use cases in other fields. Thus, developing and adapting systematic evidence mapping for EH should utilize these graph-based resources to ensure the production of scalable, interoperable, and robust maps to aid decision-making processes in chemicals policy and wider EH.

Key words: systematic evidence map; knowledge graph; evidence synthesis.

Data relevant to assessing the human and ecological health risks associated with exposure to chemical substances are increasingly available to stakeholders (Barra Caracciolo et al., 2013; Lewis et al., 2016). This trend is owed to a variety of factors, including the advent of the Internet and increasingly sensitive analytical techniques (Lewis et al., 2016), regulatory and economic changes (Lyndon, 1989; Pool and Rusch, 2014), demands for increased

transparency (Ingre-Khans et al., 2016), stricter regulatory data requirements (Commission of the European Communities, 2001; United States Environmental Protection Agency, 2016), reform of regulatory reliance on *in vivo* toxicity testing (ECHA, 2016), and a continually growing chemicals industry. The growing pool of available evidence has significant potential for informing regulatory and risk management decision making.

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society of Toxicology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

35

Downloaded from <https://academic.oup.com/toxsci/article/175/1/35/5756220> by guest on 09 September 2020

Box 1 Glossary of Terms

| | |
|-------------------------------|--|
| Database | An organized and structured collection of information (data) stored electronically within a computer system, which allows data to be accessed, manipulated, and updated. |
| Systematic evidence map (SEM) | A queryable database of systematically gathered evidence (eg, academic literature and industry reports). SEMs extract and structure data and/or metadata for exploration following a rigorous methodology which aims to minimize bias and maximize transparency. |
| Coding | The process of assigning controlled vocabulary labels or categories (referred to as "code") to data, which allows comparisons to be drawn despite the heterogeneity of the underlying dataset. For example, extracted data such as "mouse," "rat," and "guinea pig" might all be coded as "rodent" for broad comparison. |
| Query | A request for data from a database. By requesting data that meets a particular set of conditions, users can query a database for a subset of information of relevance to their specific research interests. |
| Schema | The organizational plan ("blueprint") for the structure of a database, detailing the entities stored in the database, the attributes associated with those entities, how those entities are related, what data-types can be stored in the database, etc. |
| Schemaless | Refers to databases which do not have a fixed and predefined schema. |
| Schema, on-write | Refers to the application of a schema before data is stored (written) to the database. |
| Schema, on-read | Refers to the application of a schema after data has been written to the database, at the time the data is accessed (read). |
| Ontology | A shared and reusable conceptualization of a domain which applies a logically related controlled vocabulary to describe the domain concepts, their properties and relations. |

Evidence-based approaches aim to minimize the bias associated with cherry-picking an unrepresentative subset of evidence for consideration in the decision-making process. They advocate for robust, transparent consideration of *all* relevant, available data and are the core of the evidence-based toxicology movement (Hoffmann and Hartung, 2006; Hoffmann et al., 2017). However, locating, organizing, and evaluating all relevant data is challenging when the quantity of that data is very large and growing exponentially.

Systematic evidence mapping is 1 such evidence-based approach to drawing into consideration all data which are relevant to chemicals policy and risk management workflows (see Wolfe et al., 2019). Systematic evidence maps (SEMs) are queryable databases of systematically gathered research (Box 1). They provide users with the computational access needed to organize, compare, analyze, and explore trends across a broader evidence base (Clapton et al., 2009; James et al., 2016) by:

- Collating data from different sources and storing it in a single location, such that users need only query a single database to satisfy their information requirements;
- Extracting unstructured data and storing it in a structured format, such that data can be programmatically accessed and analyzed;
- Categorizing extracted data using controlled vocabulary code, such that evidence can be broadly and meaningfully compared despite its inherent heterogeneity.

SEMs organize and characterize an evidence base such that it can be explored by a variety of end-users with varied specific research interests. The methodology was developed to address some of the limitations of systematic review and has found application in fields where formulating a single, narrowly focused review question is difficult or uninformative (Haddaway et al., 2016; James et al., 2016; Oliver and Dickson, 2016; Wolfe et al., 2019). Similarly faced with this challenge is chemicals policy and the fields which it encompasses, ie, environmental health (EH) and toxicology. It is difficult to frame a single research question with a scope which is simultaneously narrow enough to elicit the synthesis of a coherent conclusion through systematic review, and also broad enough to address the varied information requirements of chemicals policy workflows. This means that potentially several syntheses over multiple systematic reviews are required to facilitate a single decision-making process in chemicals policy. However, the significant demand for time and resources associated with systematic reviews, and the unmatched resource availability of chemicals policy, necessitates a priority setting, or problem formulation process to ensure the most efficient use of systematic review. Thus, systematic evidence mapping provides a valuable first step in this prioritization process, where the identification of emerging trends across the wider evidence base ensures resources can be targeted most efficiently (see Wolfe et al. [2019] for further discussion of the applications of SEMs in chemicals policy).

These issues are likely to become increasingly pressing as the chemicals policy paradigm shifts toward more evidence-based approaches and methods such as systematic review gain prominence. For example, agencies such as the U.S. EPA (EPA, 2018), EFSA (European Food Safety Authority, 2010), and WHO (Mandrioli et al., 2018; World Health Organization, 2019) have already begun to incorporate systematic review in their chemical risk assessment frameworks. Thus, ensuring that evidence synthesis efforts are targeting the most appropriate issues, and that the data collated for synthesis can be accessed for alternative applications, potentially across agencies, is increasingly important.

Interest in the application of SEM methodology for this context is beginning to emerge in the form of SEM exercises targeting chemicals policy issues (Martin et al., 2018; Pelch et al., 2019), various working groups expanding their evidence synthesis activities to include broader scoping and surveillance exercises (NTP-OHAT, 2019; Pelch et al., 2019; The Endocrine Disruption Exchange, 2019; Walker et al., 2018), and conference sessions discussing the potential benefits of SEMs for EH (Beverly, 2019). This emerging interest in SEM methodology, and its ability to facilitate evidence-based approaches, necessitates study of the factors key to its successful adaptation to EH contexts.

Therefore, we seek to understand how SEM databases are built and presented to end-users in fields where the practice is more mature. We hope that contextualizing this understanding within the needs of chemicals policy, risk management, and wider EH research will expedite the development of effective evidence mapping methods in this domain.

To achieve this, we examine the current state-of-the-art and common practices associated with constructing and presenting a SEM database in environmental management, a field with a strong history of systematic mapping publications and method development (Collaboration for Environmental Evidence, 2019c; Haddaway et al., 2016, 2018a; James et al., 2016). We discuss the implications of current practices for EH and highlight the challenges associated with using rigid data structures for storing the highly connected and heterogeneous data associated with the field. We outline the need for more flexible data structures in

Table 1. The Concepts Used to Guide Data Extraction and Subsequent Assessment and Discussion of the Outputs of CEE Systematic Mapping Exercises

| Concept | Definition | Metadata Extracted |
|-------------------------|---|---|
| Data storage technology | How data extracted and collated during the systematic mapping exercise were stored for future exploration | Format in which the systematic map database is presented to users (eg, spreadsheet, relational database, in-text data table, and in-text figure). |
| Data integrity | How accurately the systematic map is able to represent the raw study data on which it is based | How the relationships between entities (or study attributes) which underpin the raw data are maintained in the systematic map. |
| Data accessibility | How easy it is for end-users to access the data relevant to their research interests, or the ability of the systematic map to return data relevant to an end-user's queries | The querying mechanisms recommended in the systematic map's study report (eg, filtering table columns and navigating interactive dashboards). |
| Transparency | The ability of end-users to verify how the systematic map represents the raw study data on which it is based, ie, whether the map maintains a link between raw extracted data and eg, controlled vocabulary code. | Whether the map maintains a link between raw extracted data and controlled vocabulary code (eg, map presents code-only, map presents raw data and code), and how this link is maintained. |

EH SEMs and introduce the concept of "knowledge graphs" as an effective and intuitive model for the storage and querying of highly connected EH data. Finally, we discuss graph-based SEMs in the context of current, complementary efforts in the development of toxicological ontologies, outlining the future of systematic evidence mapping for regulatory decision making.

MATERIALS AND METHODS

Survey of published Collaboration for Environmental Evidence SEMs. We identified a dataset of exemplar SEMs for analysis: the complete set of SEMs of the Collaboration for Environmental Evidence (CEE). These maps were chosen because of CEE's role in pioneering the adaptation of systematic mapping methodology from the social sciences (Clapton et al., 2009; James et al., 2016). Through example (Collaboration for Environmental Evidence, 2019b), communication (Collaboration for Environmental Evidence, 2019a), published guidance (James et al., 2016), and reporting standards (Haddaway et al., 2018b), CEE advocate for systematic mapping and represent an ongoing case study for how the methodology can be developed as a policy and decision-making tool. Understanding how systematic map outputs serve this function, and what methodological adaptation is required to produce these outputs, is vital for successfully applying the methodology in EH. Thus, the outputs (ie, the queryable databases) of CEE's more firmly established systematic mapping practice were surveyed.

All CEE systematic maps completed before July 2019 were identified in the CEE Library (<http://www.environmentalevidence.org/completed-reviews>, last accessed July 2019). The study reports and the Supplementary information for these maps were downloaded and key metadata extracted, including title, authors, publication date, and map objectives (Supplementary Table 1). Metadata regarding the output of the systematic mapping exercises were then gathered and assessed in duplicate by T.A.M.W. and P.W. using a data extraction sheet which asked open-ended questions relating to 4 key themes of analysis: data storage technology; data integrity; data accessibility; and transparency (Table 1). These themes were developed in discussion among J.V., T.A.M.W., and P.W.

"Data storage technology" concerns the software used to construct the systematic map databases and their associated data storage formats.

"Data integrity" concerns the structures of the CEE maps. Although an important aspect of data integrity, appraising the

data extraction efforts of mappers (ie, confirming that the data extracted, coded, and stored in the database are an accurate representation of their raw counterparts in the primary literature) was beyond the scope of this exercise. Rather than verifying the data, *how* that data are represented (regardless of *what* is represented) by the systematic map database output was assessed by focusing on the ability of the systematic map to maintain the relationships which underpin these data. For example, a mapper may have extracted data from a study which investigates outcomes in a population. Although the mapper may have extracted data such as "outcome x" and "population y"—the manner in which the database structures and organizes these data will determine whether end-users can decipher that "outcome x" is somehow related to "population y."

"Data accessibility" concerns the capacity for CEE's systematic maps to facilitate data exploration by end-users. Systematic maps are research products in their own right (Haddaway et al., 2016). They should therefore present end-users with a means of programmatically accessing and querying the data they store, such that trends in potentially large datasets can be quickly identified with minimal manual effort. Accessibility is an important consideration when producing maps for an audience of varied technical skill, where ensuring that the map is accessible for nonspecialist users should not compromise the ability of more technical users to run complex queries. Therefore, the extent to which CEE systematic mapping exercises consider accessibility from the perspective of users was surveyed by extracting eg, details on the level of guidance provided to end-users wishing to query the systematic map database, and recording P.W. and T.A.M.W.'s experience of interacting with and querying the maps.

Finally, "transparency" concerns how systematic maps facilitated an end-user's ability to validate the extent to which the data presented in a map represents the data in the primary research. This was achieved by determining whether the map preserved a link between raw data and assigned controlled vocabulary labels/categories ("code" - see Box 1).

T.A.M.W. and P.W. independently noted answers to the data extraction questions before discussing and agreeing on an aggregate, consensus view. This was to contribute to comprehensive coverage of potential discussion points in relation to each theme. These aggregate assessments are presented in Supplementary Tables 1-6 and are used to evidence the state-of-the-art in terms of producing queryable systematic map databases for exploration of the environmental management

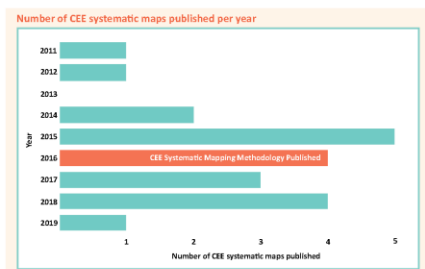


Figure 1. Publication history of CEE SEMs indicating the number of maps published per year. The year in which the CEE guidance on systematic mapping methods was published (2016) is marked on the corresponding bar (James et al., 2016).

literature. Their contents are referenced throughout the Results and Discussion sections of this survey.

RESULTS

Twenty-one systematic maps covering a variety of topics were identified in the CEE library, published between October 2011 and January 2019 (Figure 1).

The aggregated, narrative assessments of each CEE systematic map can be found in Supplementary Tables 1–6. The extracted data and aggregated assessments for each CEE systematic map are organized as follows:

- Supplementary Table 1—Bibliographic information
- Supplementary Table 2—Data storage technology
- Supplementary Table 3—Data integrity
- Supplementary Table 4—Data accessibility
- Supplementary Table 5—Transparency
- Supplementary Table 6—Additional notes

Excluded Maps

Two systematic maps (Johnson et al., 2011; Mcintosh et al., 2018) are assessed in the Supplementary information but are excluded from further analysis, as neither provided a database output which could be analyzed using our framework. Mcintosh et al. (2018) yielded a null result and therefore provided no database; Johnson et al. (2011) predated CEE's Environmental Evidence journal and its definition of systematic mapping and, although it is included in the CEE library, presented only in-text tables without an accompanying database.

Data Storage Technology

Two different data storage technologies are used in the outputs of CEE systematic mapping projects: spreadsheets constructed in Microsoft Excel ($n = 14$); and relational databases constructed with the Microsoft Access relational database management system ($n = 5$). One mapping exercise used both of these technologies to present its outputs in 2 different formats (Haddaway et al., 2014). The 2 versions of Haddaway et al. (2014) appear to be identical except that the spreadsheet version includes the results of a critical appraisal process where the relational database version does not. As the spreadsheet version presents the more complete dataset, Haddaway et al. (2014) has been coded as a spreadsheet-based systematic map for the purposes of this survey (see Supplementary Table 2, discussed in the "Data

Integrity" section). A brief description of each identified storage technology can be found in Table 2.

Data Integrity

A single, flat data table (2-dimensional array of rows and columns) was the output for the majority (84%) of CEE systematic maps (16 of 19 maps surveyed, ignoring any look-up tables housing controlled vocabulary code). 80% (4 out of 5) of the maps using the relational database storage technology were also structured as a single, flat data table.

Three maps presented more than 1 table. Two presented at least 2 tables in separate files which were not formally related to each other (Haddaway et al., 2018a; Sola et al., 2017), and 1 presented multiple tables which were related to each other in a 1:1 manner within a relational database. Systematic maps were considered to be stored in more than 1 table if there was limited overlap of the data fields housed in each table ie, if querying the map required accessing information from more than 1 table. Sola et al. (2017) is an example of this, providing the results of its quality appraisal process separately to the data it extracted and coded from the literature—thus any queries investigating critical appraisal in conjunction with another variable require the user to access information from both tables. This distinction was required because some maps, Haddaway et al. (2014) and Randall et al. (2015), presented their outputs in multiple tables, but the additional tables were simply subsets of the most complete table (ie, there was no data in the smaller tables not already present in the largest table).

Several studies included in the systematic maps contained multiple potential values for a particular attribute eg, if a single study had multiple populations and/or multiple outcomes.

Common strategies for maintaining relationships between such data in the tables of CEE maps included "expanding rows" ($n = 6$), "expanding columns" ($n = 2$), or a combination of both ($n = 5$) (see Figure 2). The remaining 6 maps either did not present/extract studies with multiple potential values per attribute ($n = 1$) or opted to house multiple values within a single cell of the table ($n = 5$, discussed further below).

"Expanding rows" refers to the practice of structuring a data table in long form: recording an entity over multiple rows. In long-form tables, a study investigating eg, 3 different outcomes might be recorded over 3 different rows. Although the data entered under the "outcome" data field might be unique in each of these 3 rows, the data for all other attributes will be repeated (Figure 3A).

In contrast, "expanding columns" describes the practice of structuring a data table in wide form; expanding what would be considered a single data field in long-form tables across several columns. Thus, all unique values associated with the data field can be recorded across a single row, eg, a study reporting 3 different outcomes might be recorded across a single row if the "outcome" attribute is split into 3 unique columns (eg, "outcome 1," "outcome 2," and "outcome 3") (Figure 3B).

The other strategy for presenting related data in a table was to record multiple values within a single cell for multiple data fields ($n = 1$), whereas 1 map presented multiple values per cell for only a single data field within the database (this distinction matters for reasons we discuss below). The practice of presenting multiple values in a single cell of the database was observed for most (5 of 6) of the maps which avoided expanding row/column structure, and similarly for most (5 of 6) of the maps adopting a long form, expanded row structure.

Table 2. Description of the Storage Technologies Used by CEE Systematic Maps

| Storage Technology | Description |
|----------------------|--|
| Spreadsheets | Spreadsheets are stand-alone applications which offer functionality for end-users wishing to explore and/or manipulate data (Zynda, 2013). A spreadsheet stores data in the cells of 2-dimensional arrays made up of rows and columns. By referencing the coordinates of cells in mathematical formulae, spreadsheet applications such as Microsoft Excel facilitate analysis, transformation, and visualization of tabular data. Although designed and optimized for quantitative data and accounting applications, spreadsheets are commonly used for storing and organizing data in a variety of research contexts, including systematic mapping exercises. |
| Relational databases | A relational database uses several formally described tables to organize data. Each table stores instances of an entity (across rows), described by a series of attributes (columns). In contrast to storing data in a single, flat data table, relational databases are able to preserve the connection between related entities. These connections are predefined and created through a system of referencing unique identifiers (primary/foreign keys) in corresponding tables. This allows users to enrich their queries with connected information, such that more complex questions can be asked of the evidence base (Elmasri and Navathe, 2013). |

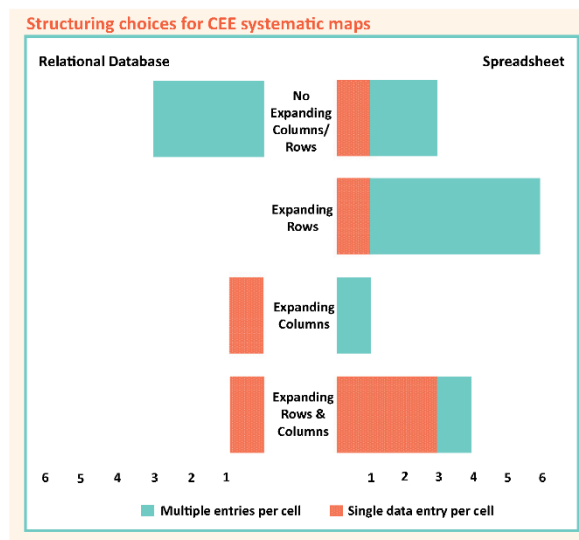


Figure 2. The number of CEE systematic maps that are structured with expanding rows and/or expanding columns as a means of preserving data relationships. Maps using the relational database storage technology are presented on the left, while maps using the spreadsheet storage technology are presented on the right. In addition, the numbers of systematic maps which store multiple values within a single cell of their data table/s are indicated by solid shading, whereas those that do not are indicated by patterned shading.

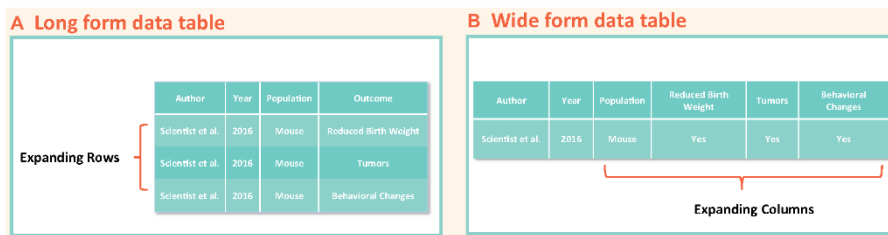


Figure 3. Illustrative example of how "expanding-rows" (A) and "expanding columns" (B) are used in long-form (A) and wide-form (B) tabular data structures, respectively.

Downloaded from https://academic.oup.com/toxsci/article/175/1/35/5756220 by guest on 09 September 2020

Data Accessibility

Eighteen of 19 surveyed systematic maps presented users with static data visualizations within their study reports (eg, bar charts, tables, and heat maps) as a means of accessing trends within the evidence. Six systematic maps additionally provided users with an open-access interactive data visualization dashboard, such that users could choose trends for exploration within the map. Four of the 6 maps supplied comprehensive guidance and/or instruction for users wishing to interact with the visualization dashboard.

Far fewer mapping exercises provided any such comprehensive guidance for querying their database output, with only 2 of 19 maps providing a detailed help file for users wishing to query the database (Haddaway et al., 2014; Randall and James, 2012). This was also seen in mapping exercises presenting guidance on interacting with their data visualization dashboards, none of which provided equivalent detailed guidance for querying the underlying database. Instead, 6 CEE systematic maps dedicated only brief discussion to querying within the text of their study reports, leaving 11 maps which offered no discernible guidance.

Where provided, the querying practices identified in user guidance/instruction were “filtering,” “sorting/”ordering,” “searching,” or some combination thereof (see Supplementary Table 4). Specific examples of queries which could be run against the database were rarely provided in such guidance, with only 2 of 19 maps providing an illustrative example of how a user’s plain-text question is translated into querying the database (Haddaway et al., 2014; Randall and James, 2012), and a further 2 of 19 making only brief mention of how a specific data field might be filtered (Cresswell et al., 2018; Randall et al., 2015). None of the maps reported the queries or querying processes used to generate visualizations or analyses. Two maps (Cheng et al., 2019; McKinnon et al., 2016) indicated that an additional data processing step had been conducted eg, using the statistical programming language R. Cheng et al. (2019) provided a link to the code used for this analysis, however the link was broken at the time this survey was conducted.

Transparency

Thirteen of 19 surveyed CEE systematic maps presented only the controlled vocabulary code which was used to classify the data of interest, not recording the raw data itself in the map. Six of 19 maps maintained a link between this code and the raw data/the coders’ interpretation of the raw data. Approaches to this included using data fields which contain free-form text alongside the controlled vocabulary terms applied to categorize this free text (5 of 6 maps, Macura et al., 2015), and providing the location of the raw data within the original study report represented as code in the systematic map (1 of 6 maps, Haddaway et al., 2015).

Seventeen of 19 CEE mapping exercises provided a codebook. Codebooks were generally supplied separate to the systematic map database, in a different file and/or format ($n = 14$), although some incorporated codebooks into the database as either look-up tables ($n = 1$, Leisher et al., 2016), or separate spreadsheets within the same workbook as the systematic map ($n = 2$, Bernes et al., 2015, 2017).

Codebooks largely presented the controlled vocabulary terms used to code study attributes (12 of 17) but did not always provide this detail (5 of 17). For codebooks which did provide controlled vocabulary terms, a narrative description or discussion of the potential types of data which might be assigned certain codes was presented in only 2 of the codebooks.

Relationships between controlled vocabulary terms were generally omitted from codebooks and/or the systematic map databases themselves, except for 1 map which structured its code as a hierarchy of nested terms (Haddaway et al., 2015).

DISCUSSION

CEE has been a driving force for the introduction of systematic mapping to the environmental sciences. Their maps act as case studies for adapting evidence-based methodologies to other fields. CEE’s involvement of stakeholders in their systematic mapping approach has undoubtedly resulted in outputs of value to those stakeholders and their specific research contexts (Haddaway and Crowe, 2018). The following discussion does not critique the use of CEE’s systematic maps for their intended purposes, but instead takes the perspective of EH applications to identify transferable aspects of current practice and remaining challenges.

Systematically Mapping the EH Evidence Base: General Considerations

EH data are complex, heterogeneous, and highly interconnected (Vinken et al., 2014). Chemical risk assessment and risk management seek to understand the outcomes which result from these complex connections—synthesizing evidence of varied resolution and origin eg, considering in combination evidence from bio- and/or environmental monitoring, *in vitro*, *in vivo*, *in silico*, and/or epidemiological studies (Martin et al., 2018; Rhomborg et al., 2013; Vandenberg et al., 2016).

The relationships which hold the disparate EH evidence base together are vital for building a more complete understanding of toxicity. These relationships underpin adverse outcome pathways (ie, how molecular initiating events lead to apical outcomes through a causal pathway of connected key events [Edwards et al., 2015]), quantitative structure-activity models (ie, how the chemical structure of a substance can be quantitatively related to its physicochemical properties and biological activity [Schultz et al., 2003]), read-across applications (ie, where predictions for data-poor substances are based on structurally related data-rich substances) and other key components of chemicals policy workflows. Such relationships are also vital for understanding the impacts of real-world exposures to mixtures of chemical substances (Sexton and Hattis, 2007).

The interconnectedness of the EH evidence base means that even if SEM methodology is used to explore just a subset of EH research, or to facilitate just 1 component of chemicals policy workflows—the data collated, extracted, and coded are likely to be of relevance to a myriad of alternative EH research interests and chemicals policy applications. Thus producing “multi-purpose,” interoperable EH SEMs that can be queried according to a variety of specific use cases is the most resource-efficient means of implementing the methodology.

However, many of the complex relationships constituting the EH evidence base are unknown to individual users, who will only have cognitive access to part of the total knowledge space in a given domain. Thus, in addition to facilitating the identification of trends which are based on relationships already known to users, EH SEMs should also facilitate the identification of relationships which are unknown to users. This would enable a more highly resolved and customizable querying process which extends beyond the user’s personal understanding of the domain, adding valuable connected contextual information with which to explore and interpret trends. It is this value, gained through accessing as well as exploring relationships—

A Relationships in a flat data table

| | A | B | C | D |
|---|----------|---|----------|---|
| 1 | | | | |
| 2 | | | | |
| 3 | α | | | |
| 4 | | | | |
| 5 | β | | γ | |
| 6 | | | | |

B Illustrative example

| | Species | Age | Sex | Outcome |
|---|---------|-----------|--------|---------|
| 1 | Mouse | 1 year | Male | Tumors |
| 2 | Rat | 2 years | Female | Tumors |
| 3 | Human | 15 years | Female | Tumors |
| 4 | Human | 30 years | Female | Tumors |
| 5 | Mouse | 0.5 years | Male | Tumors |
| 6 | Human | 35 years | Female | Tumors |

Figure 4. A, The relationship between attribute A and entity 3 is explicit in the formal structure of the array. However, the relationship between attribute A and attribute C is implicit and has to be inferred by the user from features external to the table eg, conventions around interpreting tabular data. The external conventions are not part of, or known to, the table and may not be known to the user. B, For example, a user may (in this case, correctly) infer that "sex" is a property of "species" and not "outcome," but this inference is made using external conventions and contextual understanding—the relationship is not in fact known to the table. All the table can assert is that each entity 1 through 6 has a relationship to properties of sex, age, species, and outcome, respectively.

along with the inherent complexity of those relationships—which makes the flat and rigid tabular data structures currently characterizing CEE systematic maps ill-suited to the task of systematically mapping EH data.

Limitations of Current Evidence Mapping Practice: Data Storage and Structure

Data storage is the fundamental component required for creating a systematic map database, underpinning many of the themes assessed in this survey. This discussion focuses on issues of data storage technology and its close relationship with data integrity.

Use of spreadsheets (and other flat data tables). The majority of CEE systematic maps are stored and structured as flat data tables, mostly as spreadsheets. Tables are a simple, familiar, and robust means of structuring data. However, maintaining relationships within a 2-dimensional array of rows and columns can be challenging. This is because the only explicit relationships in a 2-dimensional array (single table), are between the attributes (columns) and the entities (rows). Any relationships which exist between columns/attributes in a table can only be inferred by the user (Figure 4). We found making such inferences a challenge when surveying systematic maps of research outside of our own fields of expertise (see Supplementary Table 3). The prior knowledge required to successfully navigate data relationships within tabular maps limits their accessibility for less specialized users.

A variety of techniques were employed by CEE maps for maintaining the relationships between attributes, and for

allowing attributes to record multiple values. Of particular note were the practices of expanding columns to produce wide-form tables, and of housing multiple values within a single cell. Although expanding columns and/or housing multiple data entries in single cells do not threaten data integrity when applied to only 1 single attribute (see Thorn et al., 2016, Supplementary Table 3), a loss of referential integrity was noted for maps implementing this practice for multiple attributes.

Such loss is illustrated in Figure 5, whereby column expansion (Figure 5A), and similarly multivalued cells (Figure 5B), falsely assert data relationships unrepresentative of the raw extracted data. Loss of referential integrity is acknowledged by Neaves et al. (2015), where the authors highlight falsely asserted interattribute relationships as a limitation of their mapping exercise.

The alternative strategy used by CEE systematic mappers when structuring data as a flat table was row expansion. Although advantageous for maintaining referential integrity, these long-form data structures can be challenging to process. They can create confusion for end-users interpreting what the study "unit" (entity) which constitutes a new row in the data table is (see Supplementary Table 3). Users must also be cautious of duplicates when querying specific data fields within the table. Duplicating data can also increase the risk of data-entry errors for systematic mappers tasked with manually populating a long-form table, resulting in inconsistencies.

In summary, the spreadsheet storage technology is an unsuitable long-term solution for EH SEMs, with wide-form tables potentially compromising data integrity, and long-form tables being impractical and/or error-prone.

A Expanding columns leading to loss of referential integrity

| Author | Year | Mouse | Rat | Reduced Birth Weight | Tumors | Behavioral Changes |
|------------------|------|-------|-----|----------------------|--------|--------------------|
| Scientist et al. | 2016 | Yes | Yes | Yes | Yes | Yes |

Population
Outcome

B Multi-valued cells leading to loss of referential integrity

| Author | Year | Population | Outcome |
|------------------|------|------------|--|
| Scientist et al. | 2016 | Mouse, Rat | Reduced Birth Weight, Tumors, Behavioral Changes |

Figure 5. A, Loss of referential integrity resulting from the column expansion of more than 1 study attribute (data field). The recording of multiple populations and multiple outcomes on a single row compromises the ability of users to decipher which population was affected by which outcome. The table asserts that both populations (mice and rats) were affected by all 3 outcomes (reduced birth weight, tumors, and behavioral changes), respectively—which may not be truly representative of the raw data, compromising data integrity. B, This is similarly observed when multivalued cells are used for more than 1 study attribute.

Use of relational databases. Many of the discussed challenges associated with implementing systematic maps as flat data tables or spreadsheets are addressed by relational databases—the alternative storage technology identified in current systematic mapping practice (see Table 2). Relational databases divide entities into their own, referenceable tables—allowing links between related entities to be created and maintained. These links are coded into the database itself, and therefore do not rely on an end-user's implicit understanding of external conventions to correctly interpret.

The structure of a relational database is organized in an on-write schema, which is effectively a "blueprint" for the database (Karp, 1996); ie, the schema defines what constitutes an entity and therefore a data table, which attributes describe an entity, how an entity is related to other entities and therefore how data tables must reference others, all before data are stored. This necessitates a sound understanding of both the data to be stored in the database, and also the potential applications of the database. In fact, the optimization of end-users' capacity to query the database for a particular application is a key driver of schema design (Blaha et al., 1988).

The "schema first, data later" (Liu and Gawlick, 2015) approach of relational databases requires a more detailed level of prior knowledge regarding the structure of the evidence and/or the applications of the database. This is problematic for EH SEMs for several reasons.

First, the potential applications of an EH SEM are varied. Even where a specific use case is known, an EH SEM should at least avoid restricting access to the evidence base for alternative uses. Second, SEM methodology advises against making decisions which are based on post hoc assessment of included

studies (James et al., 2016). However, without this assessment it is difficult to design a schema capable of housing all the entities and relationships likely to arise from the varied study designs and/or evidence streams collated through an EH SEM exercise. Even if this prior assessment were advocated by SEM methodology and did not lead to the introduction of bias or inconsistencies, there would likely be far too much data for mappers to feasibly consider in the design of an EH SEM's schema.

Third, SEMs are currently constructed by human mappers, who screen, assess, and extract data from 1 included study at a time. In this manner, mappers' understandings of the relationships between entities are limited to the level of the individual study. Thus, it can be difficult to design a schema able to appropriately account for relationships which occur at an interstudy level, compromising end-users' ability to query these relationships. For example, a one-to-many relationship between population and outcome entities may be appropriate at the level of the individual study, where a single population can be investigated for many outcomes. However, at the evidence-base level, a particular outcome may in fact have been reported by many studies, and therefore investigated in many different populations—making a many-to-many relationship between population and outcome, and a schema capable of representing this relationship, more appropriate. Alternatively consider the relationships between adverse outcomes along a causal pathway. Although a relationship between eg, Outcome A and Outcome C might become apparent at the evidence base level, mappers may only have access to relationships between eg, Outcome A and Outcome B, or Outcome B and Outcome C—which occur at the individual study level.

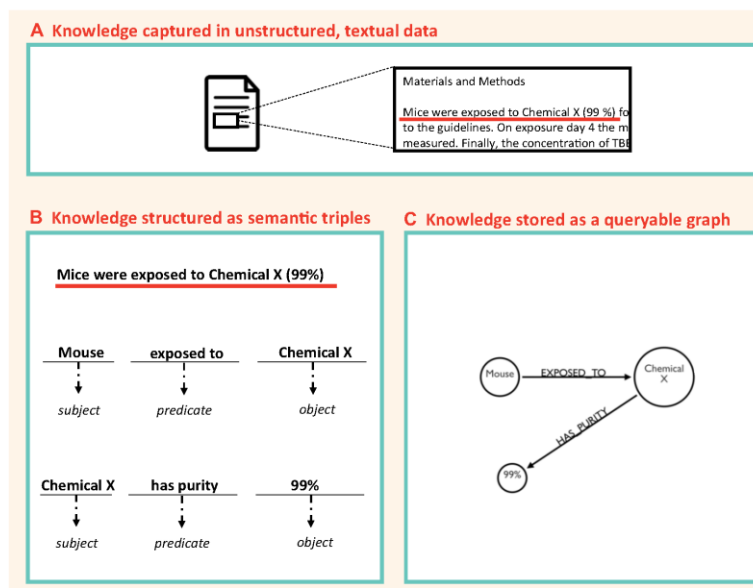


Figure 6. (A) Knowledge captured in unstructured, textual formats e.g. scientific articles, is distributed and programmatically inaccessible. (B) This knowledge can be structured in an intuitive and machine-readable way as a series of semantic subject-predicate-object triples – where entities are the subjects and/or objects and the relationships between entities are the predicates. (C) Entities can be stored as the nodes of a graph. The semantic value of the relationships between entities are preserved and stored as edges. The graph can continue to grow to produce a queryable representation of all knowledge on a topic (see Figure 7).

Finally, the growing volume and scope of EH data means that even if it were possible to devise a schema capable of accounting for all study designs that exist at present, new, and emerging study designs would soon out-date the schema, necessitating laborious, and potentially error-prone schema migration (Segaran et al., 2009).

Avoiding these issues and attempting to balance the rigidity of a schema with the fluidity or heterogeneity of the data it organizes forces mappers to implement work-arounds (eg, compromising the resolution of SEMs), the likes of which might compromise the utility of SEMs for chemicals policy applications (see Supplementary File 1).

Overcoming the Limitations of Spreadsheets and RDBs: Knowledge Graphs for Mapping EH Evidence

Expanding and enriching the application of SEMs to varied EH research problems requires moving away from the rigidity of tabular data structures and their predefined relationships. Instead, SEMs in EH should utilize more flexible, *schemaless* data models and storage technologies. We believe this flexibility is offered by knowledge graphs and associated graph-based data storage technologies.

Knowledge graphs. The scientific knowledge codified in a study report can be readily formalized as a set of subject-predicate-object “triples.” These triples can be stored as mathematical “graphs” (nodes and edges) where the nodes are the entities (subjects and objects) and the edges are the predicates, or relationships, between

the subjects and the objects (see Figure 6). Because the graph is a direct representation of the semantic content of the studies being stored, it can be said to represent the knowledge captured in the study—hence “knowledge graph” (Ontotext, 2019b).

In graph database implementations, data are stored as nodes and relationships are stored as edges. Unlike the relational model, the graph model regards relationships as first-class entities, and keeps them alongside the values they connect. Rather than “artificially” creating relationships through cross referencing primary and foreign keys in data tables, graph databases natively store relationships, preserving their semantic value, and making them accessible to queries (Figure 6 and 7) (Robinson et al., 2015). This is particularly valuable when the relationships underpinning data cannot be directly characterized *a priori*, or when the relationship between 2 pieces of information (nodes) can only be inferred through traversal of relationships which indirectly connect those nodes (Ontotext, 2019c) (eg, the inferred causal relationship between “Chemical X” and “Tumours” in Figure 7).

The graph model’s flexibility and emphasis on relationships allows it to accommodate new developments in EH research. Data produced by studies of novel design can be incorporated among, and related to, preexisting data in the database without needing to update schema and subsequently migrate data (Robinson et al., 2015). This is illustrated in Figure 7 which expands the amount of data populating the graph in Figure 6.

Knowledge graphs are already being exploited in other fields centered around the analysis of highly connected data (Ghrab

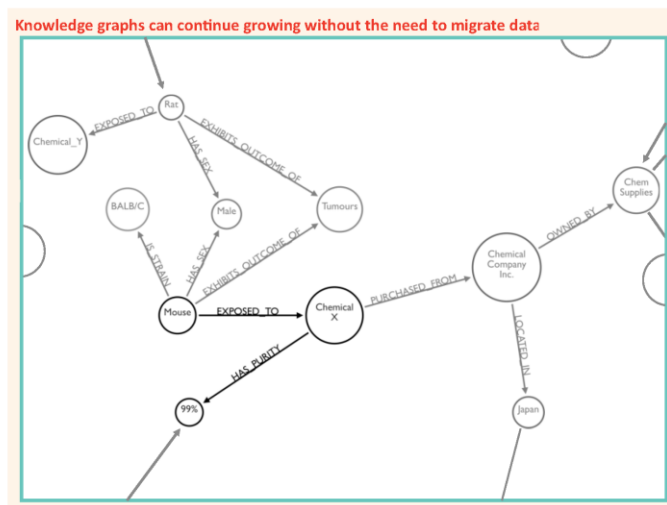


Figure 7. Storing relationships as first class entities allows knowledge graphs to continue to grow and expand without needing to revise schema and migrate data. This flexibility is particularly useful when relationships between entities cannot be characterised *a priori*.

et al., 2016). Notable use cases for graphs include: mapping complex networks of biological interaction (Aggarwal and Wang, 2010; Have and Jensen, 2013; Pavlopoulos et al., 2011); representing chemical structures (Aggarwal and Wang, 2010); tracking communication and transaction chains for fraud detection (Castellort and Laurent, 2016; Sadowski and Rathle, 2015); feeding recommendation engines for online retailers (Webber, 2018); facilitating highly customized outputs for social media platforms (Gupta et al., 2013; Weaver and Tarjan, 2013); promoting a more proactive service from search engines (Singhal, 2012); and many more. The key commonality between these applications is the identification of trends or patterns of information that facilitate the generation of new knowledge that is actionable or of value to decision-making.

Schemaless data storage and data exploration. As relationships are stored as queryable, first-class entities—the schema which implicitly structures data begins to emerge naturally and can be discovered and exploited by knowledge finding applications on-read (Janković et al., 2018; Kleppmann, 2017).

In CEE's current systematic mapping practice, trend exploration is predominantly reliant on filtering columns of a data table for specific values of interest. This requires that users are familiar with the structure of the database i.e., they know which columns house values of interest, what those values of interest are, and that their interests align with the data model imposed by the tabular map. By comparison, graphs are amenable to some ambiguity in a user's query. Beyond the potential existence of an entity of interest, users do not require prior knowledge of the graph's structure, or the relationships connecting the entity of interest to others, to successfully gain an understanding of the graph space around that entity. This facilitates the building of data models which contextualize this understanding within a particular application.

In current systematic mapping practice, data models are closely tied to the data storage mechanism and its structure. Knowledge graphs do not fix data models on-write, separating data models from data storage—thus it is possible to apply multiple models to the same graph, optimizing access to the evidence base for a variety of interests and queries. Changes can also be readily incorporated into these data models without migrating the underlying data they access.

Ontologies. A key component of wider data modeling activities is the development of domain-specific ontologies. An ontology is an agreed upon and shared “conceptualization” of a domain (Dillon et al., 2008), comprising a formal specification of terms used for describing knowledge and concepts within a domain and their relationships to each other, expressed through a standardized controlled vocabulary (Ashburner et al., 2000; National Center for Biomedical Ontology, 2019). Developing domain-specific ontologies closely mirrors the coding step of systematic evidence mapping, which is designed to conceptualize the evidence base through organizing extracted data using a controlled vocabulary of terms.

In knowledge graph applications, ontologies are stored as data themselves (Noy and Klein, 2004)—forming an additional “layer” within the graph. Raw extracted data stored in the graph can be viewed as instances of an ontology's classes. By using data models to bind nodes of raw data to the nodes of a suitable ontology, users can navigate the evidence base through this ontology—but do not lose the ability to access the underlying raw data relevant to more highly resolved queries. Furthermore, maintaining a link between raw data and the controlled vocabulary code of a shared toxicological ontology serves to promote transparency, interoperability (Hardy et al., 2012), and the development of training sets for machine-learning classifiers.

However, these concepts are underexplored in current evidence mapping practice where the majority of maps presented code in lieu of raw extracted data. This compromises transparency and limits users' ability to query data at variable resolution. In addition, coding vocabularies were rarely descriptive of the relationships that linked 1 term to another, with only 1 map organizing its code as a hierarchy of nested terms (Haddaway et al., 2015). Where relationships between code were implied, this was generally stored in separate codebooks (ie, not as data within the database)—requiring users to consult a separate document for interpretation.

Other Lessons From Current Systematic Evidence Mapping Practice

Studying the key features of a systematic map database, ie, storage technology and the data structuring choices available for those technologies, highlights the need to pursue more flexible, schemaless approaches when adapting the methodology for EH. We have identified knowledge graphs as the technology capable of providing this flexibility. Although briefly covered in the above discussion, this survey identified additional aspects of current evidence mapping practice which are worthy of discussion.

Data accessibility, user-interfaces, and map documentation. A queryable database is the main, but not sole, output of mapping exercises. All CEE maps are accompanied by a study report which details methodology, presents key trends through data visualization, and/or describes further research needs. These accompanying reports can be thought of as documentation for their database products. In the context of software development, documentation is a formal written account of each stage of development and the effective use of the software for its intended application. It is an asynchronous means of communication between all involved stakeholders, including end-users and future developers, which transforms the tacit knowledge of developers into an explicit, exchangeable format (Ding et al., 2014; Rus and Lindvall, 2002).

We found that, in general, the documentation of the maps was insufficient to make explicit the tacit knowledge of the map developers. This presented a barrier to successfully and efficiently querying the SEMs assessed in our survey. We observed that mappers' knowledge of their data model, database structure and intended uses for their database were generally underreported in accompanying SEM study reports. Discussion dedicated to instructing end-users on how they could or should interact with the database was particularly limited. This might compromise the ability of nonspecialist users to query SEMs for their own research interests. Similarly, trends visualized and analyzed in SEM study reports, which might serve as illustrative examples of how to interact with the SEM, were not accompanied by any documentation of the queries used to obtain the analyzed subset of evidence from the database—apart from 1 instance where the authors referred to code in GitHub, but the link was broken (Cheng et al., 2019).

A more common practice for facilitating end-user access to trends in the evidence base was the development of interactive data visualization dashboards (Bernes et al., 2015). Unlike their underlying databases, these dashboards were generally accompanied by documentation detailing how users could interact with the dashboard. This interaction was intuitive and required minimal technical expertise—with many dashboards adopting "point-and-click" functionality. However, interactive visualization dashboards should not be conflated with the systematic map database itself. These dashboards represent the visualized

outputs of a set of predefined queries, where users can select which of the set to display. They can be thought of as user-interfaces which have been optimized for particular queries. However, users cannot devise and visualize customized queries through such dashboards. For this, access to the underlying database is required—reinforcing the need for its documentation.

Thus the role of high-quality software documentation in promoting transparency, growth, development and maintenance of SEMs as living evidence products should not be underestimated when adapting the methodology for EH.

Including database software capacity in evidence mapping teams. A final point of interest from this survey of current systematic mapping practice is that the multidimensionality of the relational database storage technology was not utilized in the CEE maps which employed the technology. This was evidenced by systematic maps which used a flat data structure even within a relational database software environment. Such maps included Neaves et al. (2015)—which presented a single, flat data table with expanded columns despite the authors' acknowledgment of the limitations of this structure and the capacity of the chosen storage technology to overcome them.

Reasons for implementing flat relational databases were unclear or unreported. However, facilitating the access of nonspecialist users to SEM outputs may have been a potential driver of this practice. Flat tables are associated with simple querying processes such as filtering columns, whereas relational databases require a more technically demanding process of constructing queries in structured query language (SQL). However, these concerns can readily be addressed by developing user-interfaces such as the visualization dashboards discussed above, and do not explain why inherently flat storage technologies, such as spreadsheets, were not used preferentially in such cases.

Thus, an alternative motivation for implementing flat relational databases might be a lack of familiarity with database storage technologies. This highlights a key challenge for adapting SEM methodology to EH, where subject specialists interested in mapping EH evidence may not have the necessary training to successfully implement graph-based storage. This underscores the value of comprehensive documentation—where the technical construction and querying of emerging maps might serve as training opportunities for others interested in the methodology. It also indicates the importance of developing these skills within mapping teams—where recruiting databasing specialists to SEM teams might be considered as important as recruiting statisticians to systematic review teams.

CONCLUSION

Systematic evidence mapping is an emerging methodology in EH. It offers a resource-efficient means of gaining valuable insights from a vast and rapidly growing evidence base. Its overarching aims, of organizing data and providing computational access to research, should facilitate evidence-based approaches to chemical risk assessment and risk management decision-making.

The methodology has been applied in the wider environmental sciences by the CEE. Characterizing the state-of-the-art of CEE systematic mapping practices offers valuable lessons for adapting the methodology for EH.

In particular, the rigid data structures which dominate current practice are ill-suited to the complex, heterogeneous and

highly connected data constituting the EH, and toxicology evidence bases. Flat data structures and those which are closely linked to predefined, on-write schema are optimized for a narrow range of specific use cases, which fits poorly with the much broader range of uses associated with chemicals policy workflows.

Successful adaptation of SEM methodology for EH would be accelerated by adopting flexible, schemaless database technologies in place of rigid, schema-first approaches. We have argued that knowledge graphs are 1 technological solution, which potentially provide an intuitive and scalable means of representing all of the connected, complex knowledge on a topic. Converse to the flat or relational databases favored by current practice, knowledge graphs store relationships between data as first-class entities, preserving their semantic value and making them accessible to queries. This ability to explore data through relationships or "patterns of information" does not require that users are familiar with a predefined data model or schema. This vastly expands the exploratory use cases of SEMs and even facilitates the discovery of new, previously uncharacterized relationships.

There are several readily available commercial and open-source graph database implementations (ArangoDB, 2019; Neo4j, 2019; Ontotext, 2019a; Stardog, 2019), and a variety of knowledge graph applications which demonstrate the power and utility of the graph data model and its inferencing capacity. Such resources are valuable for investigating the storage and exploration of SEMs as knowledge graphs and help to lower the entry barrier associated with familiarizing and training mappers in the use of a technology novel to the field.

SUPPLEMENTARY DATA

Supplementary data are available at *Toxicological Sciences* online.

ACKNOWLEDGMENTS

The authors would like to thank Mike Wolfe at Yordas Group for input helpful to the revision of this article.

FUNDING

T.A.M.W.'s PhD is financially supported by the Centre for Global Eco-innovation (a programme funded by the European Regional Development Fund) and Yordas Group, a global consultancy in the area of chemical safety, regulations, and sustainability. P.W.'s contribution to the manuscript was funded by the Evidence-Based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health.

DECLARATION OF CONFLICTING INTERESTS

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

AUTHOR CONTRIBUTIONS

J.V. introduced the concept of graph databases, after which T.A.M.W., J.V., P.W., and C.H. established the principal ideas for the manuscript and developed an outline. T.A.M.W. wrote the first draft of the manuscript. T.A.M.W. and P.W.

conducted the survey of CEE systematic maps. J.V. offered technical expertise and edited the discussion accordingly. N.H. contributed to the revision of the manuscript, offering regulatory and chemical risk assessment expertise. All authors reviewed and edited the manuscript and contributed to its development.

REFERENCES

- Aggarwal, C. C., and Wang, H. (2010). *Managing and Mining Graph Data*, Vol. 40. Springer, New York. <https://doi.org/10.1007/978-1-4419-6045-0>.
- ArangoDB. (2019). *Graphs and ArangoDB*. Available at: <https://www.arangodb.com/arangodb-training-center/graphs/>. Accessed October 2019.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). The Gene Ontology Consortium, Michael Ashburner1, Catherine A. Ball3, Judith A. Blake4, David Botstein3, Heather Butler1, J. Michael Cherry3, Allan P. Davis4, Kara Dolinski3, Selina S. Dwight3, Janan T. Eppig4, Midori A. Harris3, David P. Hill4, Laurie Is. Nat. Genet. 25, 25–29.
- Barra Caracciolo, A., de Donato, G., Finizio, A., Grenni, P., Santoro, S., and Petrangeli, A. B. (2013). A new online database on chemicals in accordance with REACH regulation. *Hum. Ecol. Risk Assess.* 19, 1682–1699.
- Bernes, C., Bullock, J. M., Jakobsson, S., Rundlöf, M., Verheyen, K., and Lindborg, R. (2017). How are biodiversity and dispersal of species affected by the management of roadsides? A systematic map. *Environ. Evid.* 6, 1–16.
- Bernes, C., Jonsson, B. U., Junninen, K., Löhmus, A., Macdonald, E., Müller, J., and Sandström, J. (2015). What is the impact of active management on biodiversity in boreal and temperate forests set aside for conservation or restoration? A systematic map. *Environ. Evid.* 4, 1–22.
- Beverly, B. (2019). *Abstract 3267: Potential Alternatives to Systematic Review: Evidence Maps and Scoping Reviews*. Available at: <https://www.toxicology.org/events/am/AM2019/program-details.asp>. Accessed October 2019.
- Blaha, M. R., Premerlani, W. J., and Rumbaligh, J. E. (1988). Relational database design using an object-oriented methodology. *Comput. Pract.* 31, 414–427.
- Castellort, A., and Laurent, A. (2016). Rogue behavior detection in NoSQL graph databases. *J. Innov. Digit. Ecosyst.* 3, 70–82.
- Cheng, S. H., Macleod, K., Ahlroth, S., Onder, S., Perge, E., Shyamsundar, P., Rana, P., Garside, R., Kristjanson, P., McKinnon, M. C., et al. (2019). A systematic map of evidence on the contribution of forests to poverty alleviation. *Environ. Evid.* 8, 1–22.
- Clapton, J., Rutter, D., and Sharif, N. (2009). *SCIE Systematic Mapping Guidance April 2009*. Available at: <https://www.scie.org.uk/publications/researchresources/r03.pdf>. Accessed October 2019.
- Collaboration for Environmental Evidence. (2019a). *CEE Meetings*. Available at: <https://www.environmentalevidence.org/meetings>. Accessed October 2019.
- Collaboration for Environmental Evidence. (2019b). *Completed Reviews*. Available at: <http://www.environmentalevidence.org/completed-reviews>. Accessed July 2019.
- Collaboration for Environmental Evidence. (2019c). *Environmental Evidence: Systematic Map Submission Guidelines*. Available at: <https://environmentalevidencejournal.biomedcentral.com/>

- submission-guidelines/preparing-your-manuscript/systematic-map. Accessed October 2019.
- Commission of the European Communities. (2001). *White Paper: Strategy for a Future Chemicals Policy*, Vol. 13. doi: 10.1007/BF03038641.
- Cresswell, C. J., Wilcox, A., Randall, N. P., and Cunningham, H. M. (2018). What specific plant traits support ecosystem services such as pollination, bio-control and water quality protection in temperate climates? A systematic map. *Environ. Evid.* 7, 1–13.
- Dillon, T., Chang, E., Hadzic, M., and Wongthongtham, P. (2008). Differentiating conceptual modelling from data modelling, knowledge modelling and ontology modelling and a notation for ontology modelling. In *Conferences in Research and Practice in Information Technology Series*, 79.
- Ding, W., Liang, P., Tang, A., and Van Vliet, H. (2014). Knowledge-based approaches in software documentation: A systematic literature review. *Inf. Softw. Technol.* 56, 545–567.
- ECHA. (2016). *Practical Guide How to Use Alternatives to Animal Testing to Fulfil Your Information Requirements for REACH Registration*.
- Edwards, S. W., Tan, Y.-M., Villeneuve, D. L., Meek, M. E., and McQueen, C. A. (2015). Adverse outcome pathways—Organizing toxicological information to improve decision making. *J. Pharmacol. Exp. Ther.* 356, 170–181.
- Elmasri, R., and Navathe, S. B. (2013). *The Relational Data Model and Relational Database Constraints. Fundamentals of Database Systems*. Pearson Education, UK.
- EPA. (2018). *Application of Systematic Review in TSCA Risk Evaluations*, pp. 1–247. Available at: https://www.epa.gov/sites/production/files/2018-06/documents/final_application_of_sr_in_tasca_05-31-18.pdf.
- European Food Safety Authority. (2010). Application of systematic review methodology to food and feed safety assessments to support decision making¹. EFSA Guidance for those carrying out systematic reviews. *EFSA J.* 8, 1637.
- Ghrab, A., Romero, O., Skhiri, S., Vaisman, A., and Zimányi, E. (2016). GRAD: On Graph Database Modeling. Available at: <http://arxiv.org/abs/1602.00503>. Accessed October 2019.
- Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., and Zadeh, R. (2013). WTF: The Who to Follow Service at Twitter, WWW '13: Proceedings of the 22nd international conference on World Wide Web, pp. 505–514. 10.1145/2488388.2488433.
- Haddaway, N. R., Bernes, C., Jonsson, B.-G., and Hedlund, K. (2016). The benefits of systematic mapping to evidence-based environmental management. *AMBIO* 45, 613–620.
- Haddaway, N. R., Brown, C., Eales, J., Eggers, S., Josefsson, J., Kronvang, B., Randall, N. P., and Uusi-Kämppe, J. (2018a). The multifunctional roles of vegetated strips around and within agricultural fields. *Environ. Evid.* 7, 1–43.
- Haddaway, N. R., and Crowe, S. (eds) (2018). *Stakeholder Engagement in Environmental Evidence Synthesis*. Available at: <http://www.eviem.se/Documents/projekt/2018/SRbookAll.pdf>. Accessed October 2019.
- Haddaway, N. R., Hedlund, K., Jackson, L. E., Kätterer, T., Lugato, E., Thomsen, I. K., Jørgensen, H. B., and Söderström, B. (2015). What are the effects of agricultural management on soil organic carbon in boreo-temperate systems? *Environ. Evid.* 4, 1–29.
- Haddaway, N. R., Macura, B., Whaley, P., and Pullin, A. S. (2018b). ROSES RepOrting standards for Systematic Evidence Syntheses: Pro forma, flow—Diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environ. Evid.* 7, 4–11.
- Haddaway, N. R., Styles, D., and Pullin, A. S. (2014). Evidence on the environmental impacts of farm land abandonment in high altitude/mountain regions: A systematic map. *Environ. Evid.* 3, 17–19.
- Hardy, B., Apic, G., Carthew, P., Clark, D., Cook, D., Dix, I., Escher, S., Hastings, J., Heard, D. J., Jeliaskova, N., et al. (2012). Toxicology ontology perspectives. *ALTEX* 29, 139–156.
- Have, C.T., and Jensen, L. J. (2013). Databases and ontologies are graph databases ready for bioinformatics? *Bioinformatics* 29, 3107–3108.
- Weaver, J and Tarjan, P. (2013). Facebook linked data via the Graph API. *Semant. Web.* 4, 245–250.
- Hoffmann, S., de Vries, R. B. M., Stephens, M. L., Beck, N. B., Dirven, H. A. A. M., Fowle, J. R., Goodman, J. E., Hartung, T., Kimber, I., Lalu, M. M., et al. (2017). A primer on systematic reviews in toxicology. *Arch. Toxicol.* 91, 2551–2575.
- Hoffmann, S., and Hartung, T. (2006). Toward an evidence-based toxicology. *Hum. Exp. Toxicol.* 25, 497–513.
- Ingre-Khans, E., Ågerstrand, M., Beronius, A., and Rudén, C. (2016). Transparency of chemical risk assessment data under REACH. *Environ. Sci.: Process. Impacts* 18, 1508–1518.
- James, K. L., Randall, N. P., and Haddaway, N. R. (2016). A methodology for systematic mapping in environmental sciences. *Environ. Evid.* 5, 7.
- Janković, S., Mladenović, S., Mladenović, D., Vesković, S., and Glavić, D. (2018). Schema on read modeling approach as a basis of big data analytics integration in EIS. *Enterp. Inf. Syst.* 12, 1180–1201.
- Johnson, V., Fitzpatrick, I., Floyd, R., Simms, A. (2011). What is the evidence that scarcity and shocks in freshwater resources cause conflict instead of promoting collaboration? CEE review 10-010. Collaboration for Environmental Evidence, Bangor, UK.
- Karp, P. D. (1996). Database links are a foundation for interoperability. *Trends Biotechnol.* 14, 273–279.
- Kleppmann, M. (2017). *Designing Data-intensive Applications*. O'Reilly Media, Inc, UK. Available at: <https://www.oreilly.com/library/view/designing-data-intensive-applications/9781491903063/>.
- Leisher, C., Tamsah, G., Booker, F., Day, M., Samberg, L., Prosnitz, D., Agarwal, B., Matthews, E., Roe, D., Russell, D., et al. (2016). Does the gender composition of forest and fishery management groups affect resource governance and conservation outcomes? A systematic map. *Environ. Evid.* 5, 1–10.
- Lewis, K. A., Tziliavakis, J., Warner, D. J., and Green, A. (2016). An international database for pesticide risk assessments and management. *Hum. Ecol. Risk Assess.* 22, 1050–1064.
- Liu, Z. H., and Gawlick, D. (2015). *Management of Flexible Schema Data in RDBMSs—Opportunities and Limitations for NoSQL in CIDR 2015*.
- Lyndon, M. (1989). Information economics and chemical toxicity: Designing laws to produce and use data. *Mich. Law Rev.* 87, 1795–1861.
- Macura, B., Secco, L., and Pullin, A. S. (2015). What evidence exists on the impact of governance type on the conservation effectiveness of forest protected areas? Knowledge base and evidence gaps. *Environ. Evid.* 4, 24.
- Mandrioli, D., Schlünssen, V., Ádám, B., Cohen, R. A., Colosio, C., Chen, W., Fischer, A., Godderis, L., Göen, T., Ivanov, I. D., 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. *Environ. Int.* **119**, 174–185.
- Martin, O. V., Geueke, B., Groh, K. J., Chevrier, J., Fini, J.-B., Houlihan, J., Kassotis, C., Myers, P., Nagel, S. C., Pelch, K. E., et al. (2018). Protocol for a systematic map of the evidence of migrating and extractable chemicals from food contact articles. 10.5281/zenodo.2525277.
- Martin, P., Bladier, C., Meek, B., Bruyere, O., Feinblatt, E., Touvier, M., Watier, L., and Makowski, D. (2018). Weight of evidence for hazard identification: A critical review of the literature. *Environ. Health Perspect.* **126**, 076001–076015.
- McIntosh, E. J., Chapman, S., Kearney, S. G., Williams, B., Althor, G., Thorn, J. P. R., Pressey, R. L., McKinnon, M. C., and Grenyer, R. (2018). Absence of evidence for the conservation outcomes of systematic conservation planning around the globe: a systematic map. *Environ. Evid.* **7**, 22.
- McKinnon, M. C., Cheng, S. H., Dupre, S., Edmond, J., Garside, R., Glew, L., Holland, M. B., Levine, E., Masuda, Y. J., Miller, D. C., et al. (2016). What are the effects of nature conservation on human well-being? A systematic map of empirical evidence from developing countries. *Environ. Evid.* **5**, 1–25.
- National Center for Biomedical Ontology. (2019). BioPortal. Available at: <https://bioportal.bioontology.org/>. Accessed October 2019.
- Neaves, L. E., Eales, J., Whitlock, R., Hollingsworth, P. M., Burke, T., and Pullin, A. S. (2015). The fitness consequences of inbreeding in natural populations and their implications for species conservation—A systematic map. *Environ. Evid.* **4**, 1–17.
- Neo4j. (2019). Neo4j. Available at: <https://neo4j.com/>. Accessed October 2019.
- Noy, N. F., and Klein, M. (2004). Ontology evolution: Not the same as schema evolution. *Knowl. Inf. Syst.* **6**, 428–440.
- NTP-OHAT. (2019). About the Office of Health Assessment and Translation. Available at: <http://ntp.niehs.nih.gov/>. Accessed October 2019.
- Oliver, S., and Dickson, K. (2016). Policy-relevant systematic reviews to strengthen health systems: Models and mechanisms to support their production. *Evid. Policy* **12**, 235–259.
- Ontotext. (2019a). Ontotext GraphDB. Available at: <https://www.ontotext.com/products/graphdb/>. Accessed October 2019.
- Ontotext. (2019b). What is a Knowledge Graph? Available at: <https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/>. Accessed October 2019.
- Ontotext. (2019c). What is Inference? Available at: <https://www.ontotext.com/knowledgehub/fundamentals/what-is-inference/>. Accessed October 2019.
- Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Aerts, J., Schneider, R., and Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Min.* **4**, 1–27.
- Pelch, K. E., Bolden, A. L., and Kwiatkowski, C. F. (2019). Environmental chemicals and autism: A scoping review of the human and animal research. *Environ. Health Perspect.* **127**, 046001.
- Pelch, K. E., Reade, A., Wolffe, T. A. M., and Kwiatkowski, C. F. (2019). PFAS health effects database: Protocol for a systematic evidence map. *Environ. Int.* **130**, 104851.
- Pool, R., and Rusch, E. (2014). *Identifying and Reducing Environmental Health Risks of Chemicals in Our Society: Workshop Summary*. The National Academies Press, Washington, DC.
- Randall, N. P., Donnison, L. M., Lewis, P. J., and James, K. L. (2015). How effective are on-farm mitigation measures for delivering an improved water environment? A systematic map. *Environ. Evid.* **4**, 18.
- Randall, N. P., and James, K. L. (2012). The effectiveness of integrated farm management, organic farming and agri-environment schemes for conserving biodiversity in temperate Europe—A systematic map. *Environ. Evid.* **1**, 4.
- Rhomberg, L. R., Goodman, J. E., Bailey, L. A., Prueitt, R. L., Beck, N. B., Bevan, C., Honeycutt, M., Kaminski, N. E., Paoli, G., Pottenger, L. H., et al. (2013). A survey of frameworks for best practices in weight-of-evidence analyses. *Crit. Rev. Toxicol.* **43**, 753–784.
- Robinson, I., Webber, J., and Eifrem, E. (2015). *Graph Databases*. O'Reilly Media, New York. doi: 10.1016/B978-0-12-407192-6.00003-0.
- Rus, I., and Lindvall, M. (2002). Knowledge management in software engineering. *IEEE Softw.* **19**, 26–38.
- Sadowski, G., and Rathle, P. (2015). Fraud Detection: Discovering Connections with Graph Databases. Neo Technology. Available at: <https://neo4j.com/use-cases/fraud-detection/>. Accessed October 2019.
- Schultz, T. W., Cronin, M. T. D., Walker, J. D., and Aptula, A. O. (2003). Quantitative structure-activity relationships (QSARS) in toxicology: A historical perspective. *J. Mol. Struct.: THEOCHEM* **622**, 1–22.
- Segaran, T., Evans, C., and Taylor, J. (2009). *Programming the Semantic Web: Traditional Data-modeling Methods*. O'Reilly Media, USA.
- Sexton, K., and Hattis, D. (2007). Assessing cumulative health risks from exposure to environmental mixtures—Three fundamental questions. *Environ. Health Perspect.* **115**, 825–832.
- Singhal, A. (2012). *Introducing the Knowledge Graph: Things, Not Strings*. Available at: <https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>. Accessed October 2019.
- Sola, P., Cerutti, P. O., Zhou, W., Gautier, D., Iiyama, M., Shure, J., Chenevoy, A., Yila, J., Dufe, V., Nasi, R., et al. (2017). The environmental, socioeconomic, and health impacts of woodfuel value chains in Sub-Saharan Africa: A systematic map. *Environ. Evid.* **6**, 1–16.
- Stardog. (2019). Stardog. Available at: <https://www.stardog.com/>. Accessed October 2019.
- The Endocrine Disruption Exchange. (2019). TEDX Publications. Available at: <https://endocrinedisruption.org/interactive-tools/publications/>. Accessed October 2019.
- Thorn, J. P. R., Friedman, R., Benz, D., Willis, K. J., and Petrokofsky, G. (2016). What evidence exists for the effectiveness of on-farm conservation land management strategies for preserving ecosystem services in developing countries? A systematic map. *Environ. Evid.* **5**, 1–29.
- United States Environmental Protection Agency. (2016). *The Frank R. Lautenberg Chemical Safety for the 21st Century Act*. Available at: <https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/frank-r-lautenberg-chemical-safety-21st-century-act/>. Accessed October 2019.
- Vandenberg, L. N., Ågerstrand, M., Beronius, A., Beausoleil, C., Bergman, Å., Bero, L. A., Bornehag, C.-G., Boyer, C. S., Cooper, G. S., Cotgreave, I., et al. (2016). A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. *Environ. Health* **15**, 1–19.
- Vinken, M., Whelan, M., and Rogiers, V. (2014). Adverse outcome pathways: Hype or hope? *Arch. Toxicol.* **88**, 1–2.
- Walker, V. R., Boyles, A. L., Pelch, K. E., Holmgren, S. D., Shapiro, A. J., Blystone, C. R., Devito, M. J., Newbold, R. R., Blain, R., Hartman, P., et al. (2018). Human and animal evidence of potential transgenerational inheritance of health effects: An evidence map and state-of-the-science evaluation. *Environ. Int.* **115**, 48–69.


- Webber, J. (2018). *Powering Real-time Recommendations With Graph Database Technology Powering Real-time. Neo4j*.
- Wolffe, T. A. M., Whaley, P., Halsall, C., Rooney, A. A., and Walker, V. R. (2019). Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. *Environ. Int.* **130**, 104871.
- World Health Organization. (2019). *Framework for Use of Systematic Review Methods in Chemical Risk Assessment—Authors Meeting*. Available at: http://who.int/ipcs/events/SRmeeting_US/en/. Accessed October 2019.
- Zynda, M. (2013). The first killer app: A history of spreadsheets. *Interactions* **20**, 68–72.

Appendix E: NASEM Presentation

09/09/2020


Practical challenges in assessing indirectness and the implications for integrating multiple streams of evidence in systematic reviews

Paul Whaley
Evidence Based Toxicology Collaboration Research Fellow
Lancaster Environment Centre, UK
National Academy of Sciences, Engineering and Medicine, Washington DC
Evidence Integration Workshop, 3 June 2019



1

About me



- Researcher at Lancaster University and the Evidence-Based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health
- Editor for Systematic Reviews at *Environment International* (IF 7.297)
- Focus on systematic review methods for environmental health research: frameworks for systematic evidence surveillance and synthesis; critical appraisal tools; research standards; quality assurance and control in SR publishing

2

2

Today's themes

- Systematic review as a grounded approach to evidence review
- A PECO-based framework for assessing external validity of studies
- Evidence that successfully grounding SRs is extremely challenging
- How our PECO framework anticipates a computational approach to SR
- Research needs for delivering grounded, computational SRs

3

Recap of systematic review and evidence integration

A PECO-based framework for evidence integration
Practical challenges in achieving grounded analysis
Solution: a computational approach
Conclusions and credits

4

What is a systematic review?

- A systematic review is a research project which tests a hypothesis using pre-existing evidence instead of conducting a novel experiment
- The test should minimise bias introduced by (a) the evidence included in the review, and (b) by the performance of the review
 - Include all the evidence relevant to testing the hypothesis (search and screening)
 - Appraise the quality of the evidence (at level of individual study and body of evidence)
 - Synthesise the evidence into a summary result (qualitative & quantitative methods)

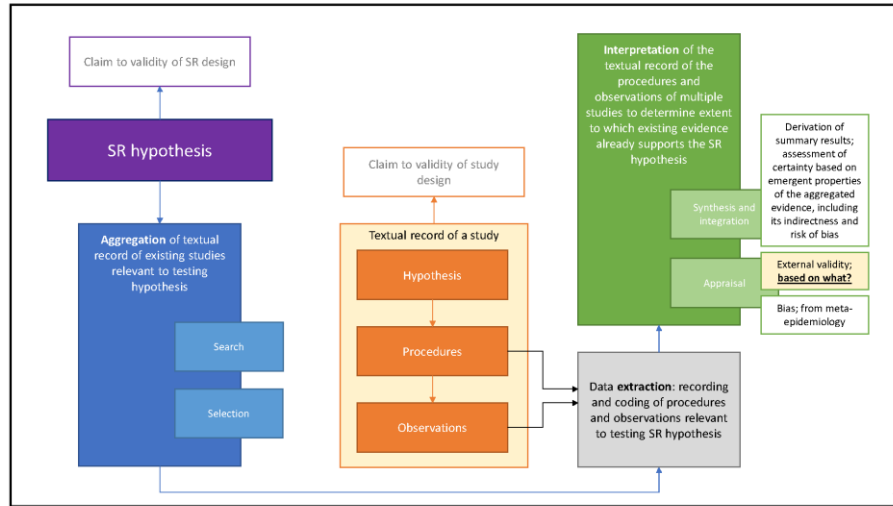


5

Systematic review = grounded interpretation

- SR is an advance on traditional narrative review because it uses explicit, discussable methods to **ground** the test of the hypothesis
- SRs are grounded when they connect interpretation of the validity of study procedures and observations with:
 - a. the textual record in the study documents of those procedures & observations
 - b. empirical evidence of the validity of the procedures described in that record
- Can't take grounding for granted, but because SR methods are explicit, they can be repeated, evaluated and deliberately changed

6



7

What is “evidence integration”?

- Evidence integration is based on a concept of dividing evidence into streams (or lines) of readily-comparable populations – usually animal vs. human, though could be a species, genus, or family
- Evidence is synthesised to produce summary results of effect of exposure in each stream
- Certainty of the evidence for the effect is assessed for each stream
- Integration is a function of combined certainty across each stream, generating a judgement of the overall level of evidence
- In the OHAT framework, mechanistic data can inform changes to the level of evidence; in the 2019 update to the IARC preamble, mechanistic evidence is a distinct stream in its own right



8

Integrating mechanistic information in SRs

- Current approaches were designed to support qualitative hazard classification, not obviously applicable to complex analysis objectives (e.g. quantifying health effects of exposures)
- We already exclude or combine multiple study designs according to principles of relevance or similarity which are informed by mechanistic data
- Mechanistic studies are conducted because they describe and/or predict health outcomes in a target population – why separate them from the whole-organism models of which they are intended to be informative?
- Can we do more to systematically incorporate mechanistic evidence into systematic reviews of exposures?

9

Recap of systematic review and evidence integration
A PECO-based framework for evidence integration
Practical challenges in achieving grounded analysis
Solution: a computational approach
Conclusions and credits

10

The role of PECO statements in SRs

- SR = test of a hypothesis using existing evidence
- Hypothesis interpreted as a research question, formulated as a Population-Exposure-Comparator-Outcome statement
- Common research scenario in environmental health: there is a suspected relationship between an exposure and an outcome, but the nature of the relationship is unknown (scenario 1, right)

P: Among adult females, what is the effect of
E: 1 µg/kg bw childhood organochlorine levels in blood, versus
C: 1 µg/kg bw incremental increase on
O: endometriosis?



11

Including indirect evidence

- Necessary in a SR of an exposure-outcome relationship when we do not have certain evidence within the strict confines of the PECO
- Look at intermediate outcomes, disease markers, animal models, similar chemicals (read-across), etc. etc.
- All indirect evidence but still relevant to the question, and therefore could increase certainty in test of hypothesis
- There are lots of ways in which this evidence can be organised

12

Example: Matta et al. (2019)

K. Matta et al.

Environment International 124 (2019) 400–407

Table 2

Body of evidence structure based on major experimental outcomes of endometriosis to guide grouping endpoints and experiments.

| Level | Endometriosis-related outcomes | Endpoint/assay examples | Body of evidence grouping examples |
|-------------------------|---|--|---|
| Primary/apical outcomes | Spontaneous endometriosis | <i>In vivo</i> : onset after chronic/transgenerational exposure in non-human primates | 1- Spontaneous endometriosis in animals |
| | Migration/attachment | <i>In vivo</i> : experiments evaluating the invasiveness of implants in rodents or primates <i>In vitro</i> : migration assays in cell models | 2- Invasiveness of endometriotic tissue in animals 3 - Invasiveness of endometriotic tissue in cell cultures |
| | Survival/proliferation/apoptosis | <i>In vivo</i> : experiments on proliferation/expansion of endometriotic lesions in rodents and/or primates <i>In vitro</i> : proliferation/viability/apoptosis cell assays | 4 - Survival/proliferation of lesions in animals 5- Proliferation in cell culture |
| Intermediary /secondary | Progesterone resistance | <i>In vivo</i> : PR-B/A expression | 6- Progesterone resistance in animals |
| | Aromatase/steroidogenic pathway | <i>In vitro</i> : CYP19A1 expression | 7. Disruption of aromatase pathway in cell culture |
| | Inflammatory cytokines | <i>In vivo</i> : IL6 levels | 8 - Inflammation in animals |
| | Other outcomes: immunosuppression, oxidative stress | | |

13

13

Interpreting Matta et al. into PECO's

| Level | Population | Exposure | Comparator | Outcome |
|-------|---|----------|------------|---|
| 1* | Non-human primate | Chronic | ? | Spontaneous endometriosis |
| | Non-human primate with implanted tissue | Transgen | ? | Invasiveness of implanted tissue |
| | Rodent | Chronic | ? | Proliferation of endometriotic tissue |
| 2* | In vivo | ? | ? | PR-B/A expression (progesterone resistance) |
| | In vitro | ? | ? | CYP19A1 expression (aromatase pathway) |
| | In vivo | ? | ? | Inflammation |

- As described, relationship between included studies, hypotheses under test and the relevant PECO's are ambiguous – characteristics need to be more tightly defined
- In actuality, we probably don't need to define in advance all the potentially relevant sub-PECO's (cumbersome, p-hacking) – can't we just observe how direct the evidence is?

14

14

Proposal: PECO as a directness framework

Relative to the PECO which is the target of a SR, all evidence is to some extent indirect, and may therefore be evaluated as follows:

- Define the target PECO (tPECO) for the SR, as we do already
- Extract the experimental PECO (ePECO) from each included study
- Evaluate the similarity of each ePECO to the SR tPECO (ePECO→tPECO)
- Describe directness of the evidence overall as a function of how the ePECOs map in aggregate onto the SR tPECO

15

15

What this might look like...

| Study | P features | | | | E features | | | C features | O features |
|--------|------------|----------------|---------------|--------|--------------|---------------------|--------------------|-----------------------|-------------------|
| | Specie | L. Org. | Age | Sex | Chem | Dose | Timing | Dose | Outcome |
| Target | Human | Whole organism | Pre-menopause | Female | OC | 1 µg/kg bw | Pre-puberty | 1 µg/kg bw increments | Endometriosis |
| Ref013 | Human | Whole organism | Adult | Female | Furan mix | High exposure group | Up to 16 years age | Low exposure group | Endometriosis |
| Ref852 | Human | HESC cells | - | Female | TCDD | 10uM solution | - | 10 uM increments | Migration |
| Ref134 | Wistar Rat | Whole organism | 24 months | Male | Chlorpyrifos | 1000 µg/kg bw/d | Until weaning | Vehicle | PR-B/A expression |

- Allows us to describe all types of study design using the same set of categories
- We can make comparisons between experimental PECO and our target question, without having to divide evidence up into streams beforehand
- Makes explicit the information being interpreted (if not yet the rules for interpretation)

16

16

| P features | | | | | E features | | | C features | O features |
|------------|------------|----------------|---------------|--------|--------------|---------------------|--------------------|-----------------------|-------------------|
| Study | Specie | L. Org. | Age | Sex | Chem | Dose | Timing | Dose | Outcome |
| Target | Human | Whole organism | Pre-menopause | Female | OC | 1 µg/kg bw | Pre-puberty | 1 µg/kg bw increments | Endometriosis |
| Ref013 | Human | Whole organism | Adult | Female | Furan mix | High exposure group | Up to 16 years age | Low exposure group | Endometriosis |
| Ref852 | Human | HESC cells | - | - | TCDD | 100µM solution | - | 10 µM increments | Migration |
| Ref134 | Wistar Rat | Whole organism | 24 months | Male | Chlorpyrifos | 1000 µg/kg bw/d | Until weaning | Vehicle | PR-B/A expression |

Judgement of similarity at level of

- A. whole study
- B. broad PECO element
- C. individual PECO sub-element

How do we ensure these judgements are valid?

17

Rules for interpretation? Maybe in AOPs

- Is the observed intermediate event strongly predictive of the target outcome?
- Do the mechanisms in the observed population also happen in the target population?
- Intuition: the more certain the answer, the lower the sense that the evidence is indirect
- If true, maybe judgement of similarity can be derived from a function of certainty in the AOP network
- Potential for grounding judgements of directness in biological knowledge (so long as that knowledge is gathered systematically)

The diagram illustrates an Adverse Outcome Pathway (AOP) network. It starts with 'Exposure' (green box) leading to 'Initiating Event' (blue box). From 'Initiating Event', the pathway proceeds to 'Event 2' (blue box), which is influenced by two yellow circles representing uncertainty levels. 'Event 2' leads to 'Event 3' (blue box), which is also influenced by a yellow circle. 'Event 3' leads to 'Event 4' (blue box), which is influenced by a grey circle. Finally, 'Event 4' leads to 'Outcome' (orange box). A red triangle above the pathway indicates the 'Level of cellular organisation' increasing from left to right. A dashed arrow labeled 'Uncertainty' points from the grey circle to the pathway.

18

Recap of systematic review and evidence integration
A PECO-based framework for evidence integration
Practical challenges in achieving grounded analysis
Solution: a computational approach
Conclusions and credits

19

Two major, practical threats to grounded SR

- Implementing valid processes
- Overwhelming data volume

20

Prepublication data on EH systematic reviews

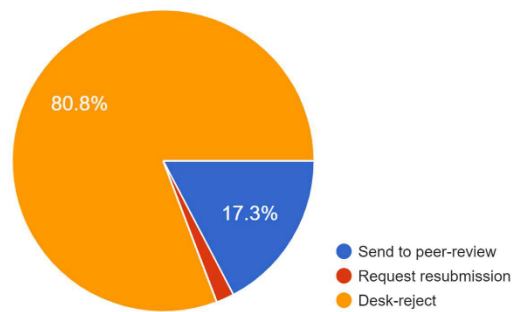
- At *Environment International*, we triage submissions on six key features of a SR:
 1. Are objectives appropriate to investigating research question?
 2. Does the search methodology miss relevant evidence?
 3. Do the exclusion criteria and screening process exclude relevant evidence?
 4. Have included studies been appraised using a valid risk of bias instrument?
 5. Have appropriate quantitative and qualitative been used to synthesise the evidence?
 6. Has certainty in the evidence been assessed using appropriate, defined criteria?
- We score the methods on a Likert scale of 1-5 (1 = serious concerns)
- A score of 1 or 2 in any domain is a critical shortcoming and results in desk-rejection*

*Authors receive a detailed triage report and editor feedback on identified issues; as often as possible issues are discussed with authors with a view to enabling resubmission

21

21

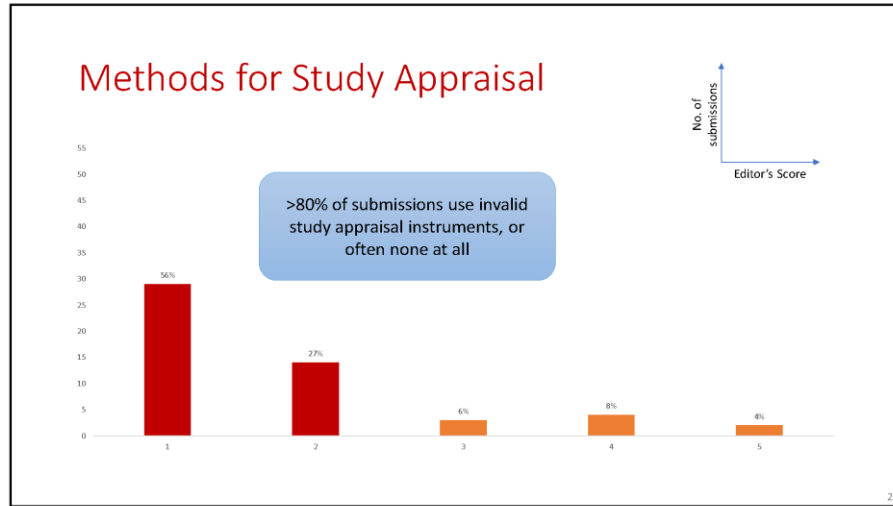
Summary of Triage Decisions



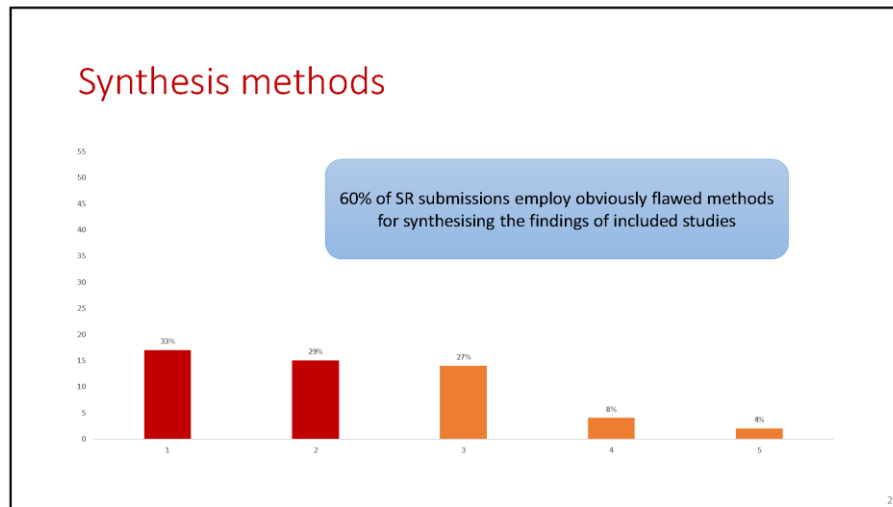
Period April 2018 - May 2019, since introduction of triage tool. n=52

22

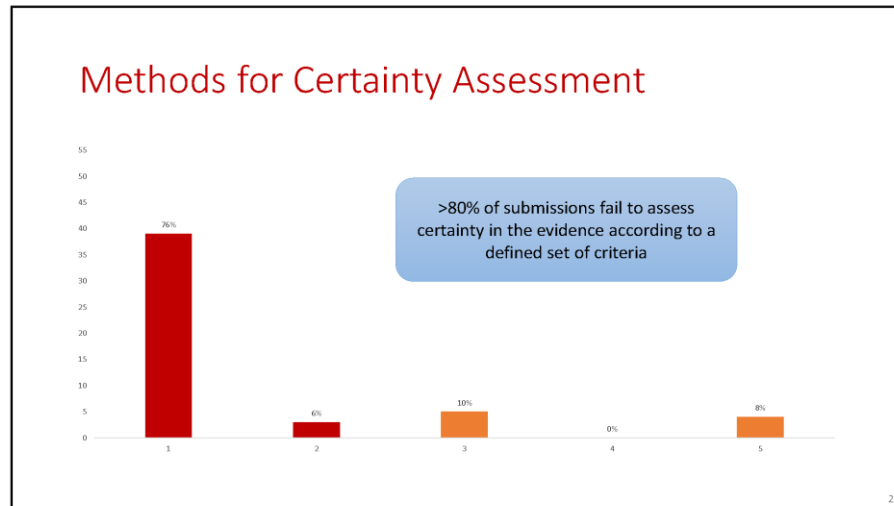
22



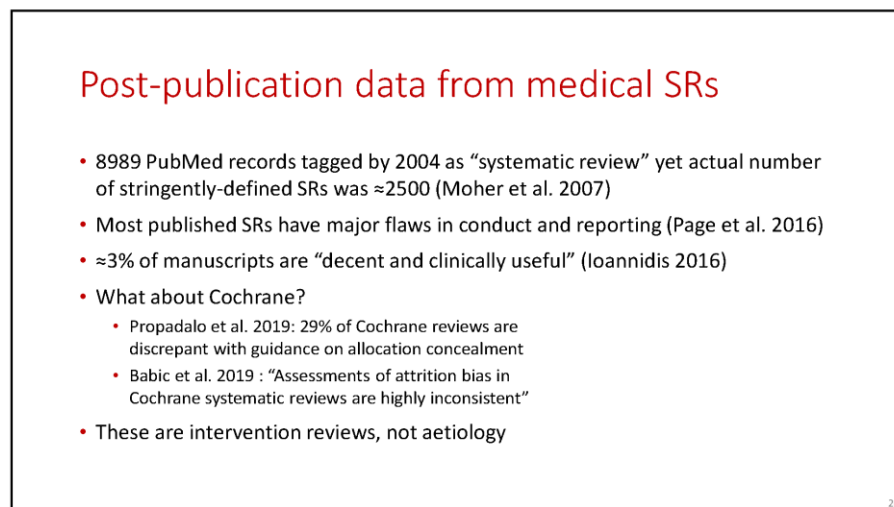
23



24



25



26

Educating our way out of this challenge?

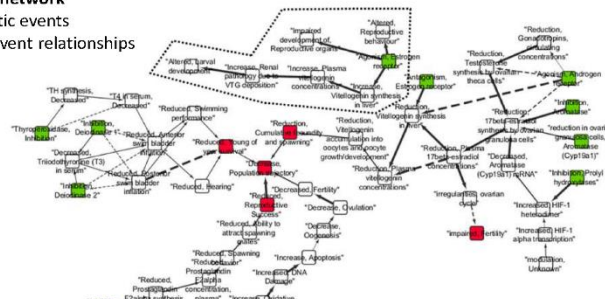
- Most EH research teams do not successfully apply even the simpler, well-documented instruments (e.g. OHAT, Navigation Guide, GRADE) which would better ground their SR methods
- Even if we ended up doing as well on average as the medics, we wouldn't be doing well enough
- Doing as well as the outlier (setting up a Cochrane for EH research) is not a near-future event
- Complex tools like ROBINS-E: what prospects for successful use given the above?



27

The data volume problem

CYP19 AOP network
 >50 biokinetic events
 >65 event/event relationships



From: Villeneuve et al. (2018) Adverse Outcome Pathway Network Analytics

28

The integration challenge, in a nutshell

- If the stars align, simple SRs can successfully be conducted
- But in most normal scenarios, SR methods are out of reach of most researchers' capacity to apply them successfully
- Methods for integrating mechanistic data into SRs are unlikely to be any easier to apply successfully – plus, they overwhelm us with data
- We can't escape this challenge: the methods need to be applied in order for SRs to be grounded
- So we need a scalable approach to grounded integration methods

29

Recap of systematic review and evidence integration
A PECO-based framework for evidence integration
Practical challenges in achieving grounded analysis
Solution: a computational approach
Conclusions and credits

30

In favour of algorithms

- By turning features into numbers, we can make processes repeatable and scalable (i.e. computers can do the work for us)
- Discussable inputs which can be changed deliberately
- The challenge is preserving the links in the chain of evidence that keeps the process grounded (score-text-design-validity)
- How do we do that for complex SR questions, e.g. predicting dose-response relationships in human populations using indirect evidence?



31

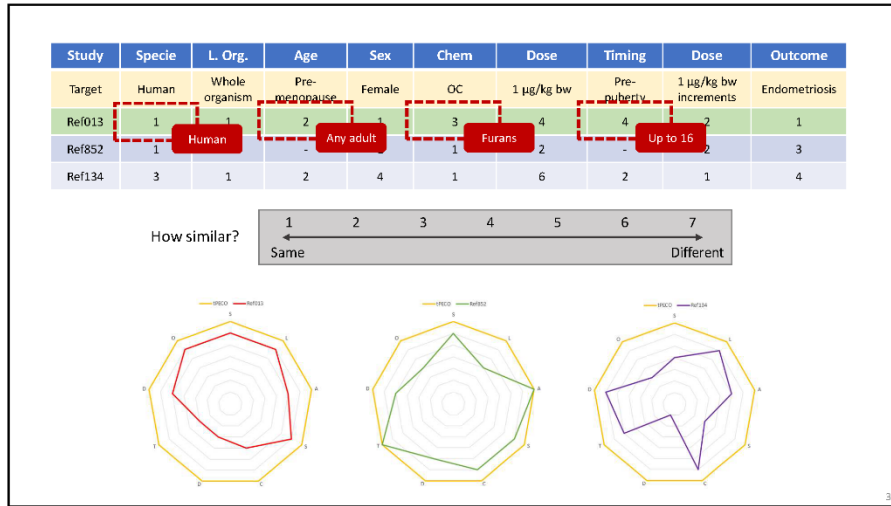
31

| Study | Specie | P features | | | E features | | | C features | O features |
|--------|------------|----------------|---------------|--------|--------------|---------------------|--------------------|-----------------------|-------------------|
| | | L. Org. | Age | Sex | Chem | Dose | Timing | Dose | Outcome |
| Target | Human | Whole organism | Pre-menopause | Female | OC | 1 µg/kg bw | Pre-puberty | 1 µg/kg bw increments | Endometriosis |
| Ref013 | Human | Whole organism | Adult | Female | Furan mix | High exposure group | Up to 16 years age | Low exposure group | Endometriosis |
| Ref852 | Human | HESC cells | - | Female | TCDD | 10µM solution | - | 10 µM increments | Migration |
| Ref134 | Wistar Rat | Whole organism | 24 months | Male | Chlorpyrifos | 1000 µg/kg bw/d | Until weaning | Vehicle | PR-B/A expression |

We can readily turn judgements of similarity into numbers within our tPECO framework

32

32



33

How do we ground similarity scores?

- In our mechanistic study, what makes a rat score a 3? Or PR-B/A a 4?
- The million (multi-trillion?) dollar question

| Study | P features | | | | E features | | | C features | O features |
|--------|------------|----------------|---------------|--------|------------|------------|-------------|-----------------------|---------------|
| | Specie | L. Org. | Age | Sex | Chem | Dose | Timing | Dose | Outcome |
| Target | Human | Whole organism | Pre-menopause | Female | OC | 1 µg/kg bw | Pre-puberty | 1 µg/kg bw increments | Endometriosis |
| Ref013 | 1 | 1 | 2 | 1 | 3 | 4 | 4 | 2 | 1 |
| Ref852 | 1 | 3 | - | 1 | 1 | 2 | - | 2 | 3 |
| Ref134 | 3 | Rat | 2 | 4 | 1 | 6 | 2 | 1 | 4 |

PR-B/A expression

34

Research for grounding similarity scores

- Grounding requires us to connect the numbers to the textual record, and to the empirical evidence for their interpretation (their value)
- There are at least three big jobs that need to be done
 1. Systematic methods for AOP development
 2. Automated data extraction
 3. Machine-learning models for weighting evidence
- Probably all three need doing, because it looks like a big-data challenge

35

35

1. Systematic approach to AOP development

- Data model for external validity is underpinned by AOPs
- But we haven't formalised the key features from which AOPs are built
 - What information in the textual record should we use when developing an AOP?
 - What rules should we follow in developing valid AOPs / determining their plausibility?
- This will need to be grounded, and therefore systematic*
- If we figure this out, we will know what rules the machines should be following when identifying and evaluating putative AOPs for us

*SR approach to AOPs is subject of EBTC GRADE pre-meeting in Hamilton next week

36

36

2. Automated data extraction

- PECO features and AOP information need extracting from narrative text in full study reports
- This will be a very large extraction job: high level of granularity across thousands of documents
- Would require automation to be practically doable, therefore natural language processing (NLP) approach
- NLP methods can't yet differentiate the features we are interested in, at level of full text, with enough reliability to do data extraction for us
- The step-change which is required implies need for a full-text toxicology corpus training set

Solutions

Chlorpyrifos solutions were prepared by dilution of a commercial formulation (CPF, LeSolan 400 BRG, 40% w/v, Dow Agrosciences Industrial Ltda) in saline (NaCl, 0.9%). In order to achieve the specified doses applied for each group (see below), dilution was adjusted based on the content of the active ingredient specified in the formulation. Solutions were always freshly prepared and used on the same day. Control animals were treated with saline.

Experimental Protocol

Considering that exposure to pesticides in farmers may have different cycle lengths depending on the season, number of crops per season, and type of crop [11, 34], we proposed a design of intermittent exposure at two time intervals, but administered with the same number of total doses per group. One group of animals was treated weekly with CPF or saline, for 12 weeks and another group of animals was treated three times a week, on alternating days, for 4 weeks. By adopting the same number of injections (total 12 administrations), within two time intervals, we could test whether longer or shorter intervals between exposures differentially impacts the cardioprotection function. The CPF doses chosen for treatment were 7 mg/kg and 10 mg/kg. The dose of 10 mg/kg corresponds to 1/3 of the dose that impaired cardiovascular function in a model of acute intoxication with CPF previously described by our group [26]. The 7 mg/kg corresponds to 2/3 of the 10 mg/kg dose. Either intraperitoneal or CPF administration was performed through intraperitoneal injection to assure accurate and efficient delivery of doses. The reasons of selection of CPF compounds for being used

37

Teaching computers to read

Rats have four legs, big ears and a tail.

We breed Han-Wistars.

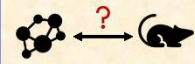
Paul is being ratly because he is tired and hungry.

Paul isn't ratly enough for Warfarin to poison him.

Artists like a golden ratio.

No rats were harmed during filming.

Tree-rats keep stealing food from our bird-feeders.



- Computers "read" by building statistical models to attempt to discern the same regularities in a written document that people respond to when discerning the meaning therein (the written concept "rat" will have a certain statistical shape in a document)
- The problem is there are lots of things which will, to the statistical models, look like regularities which are not meaningful (i.e. look like rats but are not rats), while many meaningful regularities will be invisible to them (are rats, but do not look like them)
- To help, we can manually annotate a large, representative set of documents (a corpus) to show the machines the parts which are meaningful to us (where the rats actually are). The machine can heavily weight this information in its statistical model, massively improving its performance for a data extraction task

38

Machine-learning models for weighting evidence

- Starts off with responding to the features we know are important (blinding, species, vehicle, event, dose regimen, formulation etc.)
- Uses statistical models of those features to repeat human processes at high volume (e.g. judges risk of bias, indirectness, etc.)
- Large datasets yielded by success with NLP implies quantitative models for interpreting meaning of dataset features
- Over time, the machine identifies predictive features we are not aware of, and improves its performance beyond human capability

39

Recap of systematic review and evidence integration
A PECO-based framework for evidence integration
Practical challenges in achieving grounded analysis
Solution: a computational approach
Conclusions and credits

40

Summary

- Successful evidence integration requires us to ground complex judgements of the directness of evidence in (a) the textual record of research and (b) in biological knowledge
- We have proposed a framework for using PECO statements to structure judgments about external validity, which seems to necessitate a computational implementation
- We have outlined a research roadmap toward how such an implementation can be realised and grounded

41

41

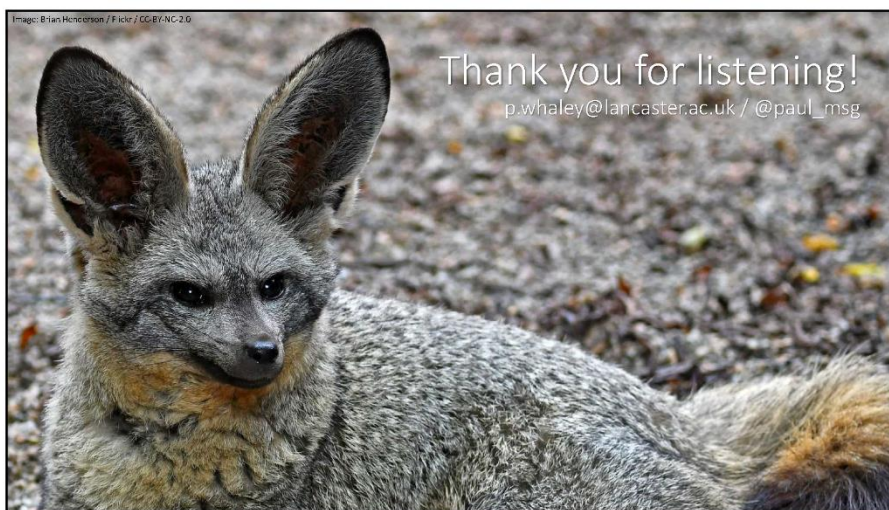
Thanks to...*

- **Stephen Wattam**, WAP Academic Consultancy Ltd
- **Daniele Wikoff**, Toxstrategies LLC
- **Oliver Wild**, Lancaster Environment Centre, UK
- **Taylor Wolffe**, Lancaster Environment Centre, UK
- **Paul Rayson**, Lancaster University School of Computing and Communications
- **John Vidler**, Lancaster University School of Computing and Communications
- **EBTC staff**: Katya Tsaoun, Sebastian Hoffmann, Rob de Vries
- **Patient listeners**: Rebecca Morgan, Michelle Angrish

*Credit is theirs, mistakes are mine

42

42



43

Wikoff Model* for quantitative integration

- Model measures the extent to which a body of evidence relevant to the potential carcinogenicity of a chemical fulfils the KCCs
- Uses three inputs (1-3) and an algorithm (4) to provide a numeric description (5) of how well the evidence "matches" the KCCs
- It works a bit like calculating Flesch-Kincaid readability scores in word processors: overall target characteristic described as a function of some measurable properties, normalised onto a scale

Step 1
Individual Study Assessment

1

Component 1: Reliability
(Internal validity)
How well was the study designed/reported to evaluate the endpoint?
(1/2/3)

2

Component 2: Strength
(External validity)
How good is the model of characterization outcome relative to cancer/KCC?
(1/2/4/8)

3

Component 3: Activity
Result of study by model (Active/Inactive)
(1/0)

4

$$\sum_{i, \text{all}} \left[\left(\frac{w_R(-R_i + 4) + w_M M_i}{w_R R_{\max} + w_M M_{\max}} \right) * \frac{E_{\text{Act},i}}{E_{\text{Total},i}} - \left(\frac{w_R(-R_i + 4) + w_M M_i}{w_R R_{\max} + w_M M_{\max}} \right) * \frac{E_{\text{InAct},i}}{E_{\text{Total},i}} \right]$$

$n_{i, \text{all}}$

5

-1

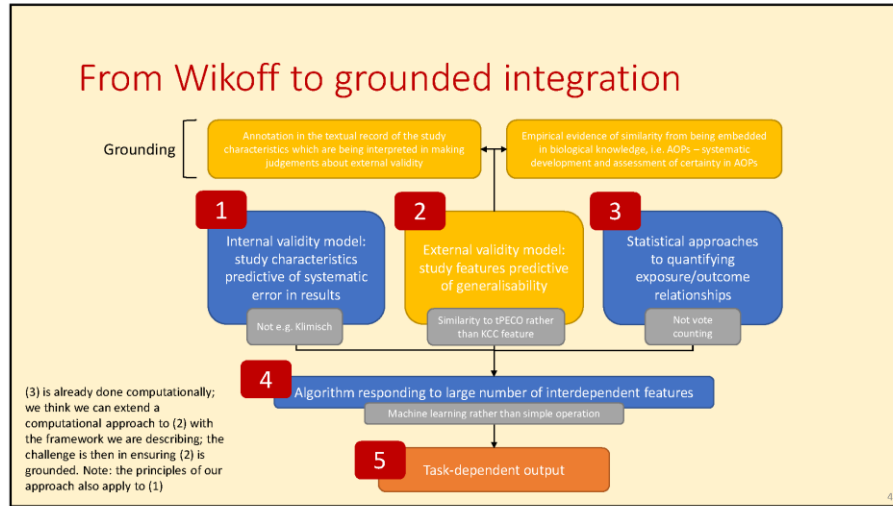
0

1

← KCCs not fulfilled KCCs fulfilled →

*Oversimplified version presented here, see Wikoff et al. (2019) for detail

44



45