Review Article

Systematic Review in Evidence-Based Risk Assessment

Nawal Farhat[1,2], Katya Tsaioun[3], Patrick Saunders-Hastings[1], Rebecca L. Morgan[4], Siva Ramoju[5], Thomas Hartung[6,7] and Daniel Krewski[1,2,5]

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada; [2]McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, Canada; [3]Evidence-Based Toxicology Collaboration, Johns Hopkins University, Baltimore, MD, USA; [4]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada; [5]Risk Sciences International, Ottawa, Canada; [6]Chair for Evidence-based Toxicology and Center for Alternatives to Animal Testing (CAAT), Johns Hopkins University, Baltimore, MD, USA; [7]CAAT-Europe, University of Konstanz, Konstanz, Germany

Abstract

Systematic reviews provide a structured framework for summarizing the available evidence in a comprehensive, objective, and transparent manner. They inform evidence-based guidelines in medicine, public policy, and more recently, in environmental health and toxicology. Many regulatory agencies have extended and adapted the well-established systematic review methods, initially developed for clinical studies, for their assessment needs. The use of systematic reviews to summarize evidence from existing human, animal, and mechanistic studies can reduce reliance on animal test data in risk assessment and can help avoid unnecessary duplication of animal experiments that have already been conducted. As alternative test methods can be expected to play an increasing role in human health risk assessment in the future, systematic reviews can be particularly helpful in validating these alternatives. The field of evidence-based toxicology has undergone extensive development since its first meeting in 2007 as a result of collaborative efforts among international experts and public health agencies, particularly with respect to the use of mechanistic data and evidence integration. The continued development and wider adoption of systematic review methodology can lead to better 3R implementation. As undertaking a systematic review can be a complex and lengthy process, it is important to understand the main steps involved. Key steps, along with current best practices, are described with references to guidance from organizations with expertise in evidence synthesis. Applications of systematic reviews in clinical, observational, and experimental studies are presented. Finally, software tools available to facilitate and increase the efficiency of completing a systematic review are described.

1 Introduction

Systematic reviews are comprehensive summaries of existing literature where the available evidence on a clearly formulated research question is synthesized using a transparent and objective stepwise process. Currently, systematic reviews serve as essential tools for policymakers making decisions to improve health outcomes and cost-effectiveness of interventions and programs (Fox, 2010). Although early applications of systematic reviews in the field of medicine formed the foundation for evidence-based medicine (EBM), recent applications are found in many fields, including observational and toxicological investigations. Regulatory agencies, such as the US Environmental Protection Agency (EPA), the European Food and Safety Authority (EFSA), and the National Toxicology Program (NTP), rely on findings from systematic reviews as the best available evidence to support public health policies and strategies. Further, the Institute of Medicine/National Academy of Medicine requires the use of systematic reviews to inform recommendations in their mandate for the development of trustworthy clinical practice guidelines (Institute of Medicine, 2011).

Other forms of evidence-based methodologies such as systematic scoping reviews and systematic evidence maps (Wolffe et al., 2020) are used in identifying knowledge gaps to be addressed

This content is part of the *Special Issue* Development of an Evidence-Based Risk Assessment Framework. doi:10.14573/altex.22S2

Received April 11, 2020; Accepted June 11, 2021;
Epub September 27, 2021; © The Authors, 2022.

ALTEX 39(3), 463-479. doi:10.14573/altex.2004111

Correspondence: Nawal Farhat, PhD
McLaughlin Centre for Population Health Risk Assessment, University of Ottawa
600 Peter Morand Crescent, Room 216, Ottawa, ON K1G 5Z3, Canada
(nfarhat@uottawa.ca)

in future research. These types of studies may be performed before a full systematic review is undertaken because they allow for a broader scope and identification of areas with sufficient literature base to justify performing a systematic review. In cases where time or resources do not permit a full systematic review, such a scoping review can provide a useful, although not exhaustive, summary of the available evidence.

1.1 Contributions of systematic reviews to the 3Rs

Systematic reviews of epidemiological and animal experimental studies have the potential to lead to more humane animal research. Ritskes-Hoitinga and van Luijk (2019) note that systematic reviews will help fulfill 3R legislative requirements as formulated in EU Directive 2010/63. Systematic reviews can support the 3Rs – particularly reduction – in several ways (de Vries et al., 2014b). By identifying relevant human evidence through the process of systematic review, the need for reliance on animal evidence can be substantially reduced. Systematic review of findings from animal studies can lead to a reduction in unnecessary animal use and duplication of studies (de Vries et al., 2011; Hooijmans et al., 2010) and even to replacement of animal testing in instances where animal data does not translate well to human health risks (Ritskes-Hoitinga and van Luijk, 2019). Findings from systematic reviews can be used to target only those critical remaining information gaps that require animal testing (McCann et al., 2016). Systematic reviews can also serve as important tools to summarize findings from alternative test methods, including new approach methodologies (NAMs), which are *in vitro* and *in silico* methods. The use of NAMs is promoted by the US EPA's Strategic Plan to specifically reduce animal testing and rely on NAMs for making decisions under the Toxic Substances Control Act (US EPA, 2018a). The expanded use of NAMs should contribute to the 3Rs by reducing the need for animal testing (Andersen et al., 2019). Evidence from multiple streams, including human and animal studies along with findings from NAMs can be integrated within a systematic review (Krewski et al., 2022). Appreciating the importance of systematic reviews in the context of the 3Rs, recent publications have called for making systematic reviews of animal experiments a standard practice (Russel and Burch, 1959; de Vries et al., 2011; Hooijmans et al., 2010; Ritskes-Hoitinga et al., 2014; Ritskes-Hoitinga and Wever, 2018).

Some examples can illustrate where systematic reviews have already contributed to the 3Rs. A systematic review on intestinal anastomosis assessed 350 animal studies and documented poor reporting quality and internal validity of the studies reviewed (Yauw et al., 2015). The findings can lead to improvements in study quality and reporting of future research results on intestinal anastomosis. The authors also encouraged investigators to thoroughly review available animal studies before proceeding with new projects to avoid unnecessary duplication of experiments involving animals. In another example, McCann et al. (2016) reported that an earlier systematic review (Banwell et al., 2009) summarizing findings from animal models of stroke related to the efficacy of interleukin-1 receptor antagonist use in arthritis treatment had guided further animal testing in a more focused manner. The 2016 update of the original systematic review further served to identify gaps remaining on efficacy data of interleukin-1 receptor antagonist (McCann et al., 2016). A systematic review and meta-analysis (Currie et al., 2019) of 337 studies on chemotherapy-induced peripheral neuropathy (CIPN) found that many of the studies assessed in the systematic review had methodological deficiencies, predominantly included only male animals, and often assessed outcomes that were not ideal in terms of clinical relevance to CIPN. This finding can lead to more efficient and effective animal testing in CIPN-related research and reduce animal testing for irrelevant outcomes.

Quantitative synthesis as part of systematic review can also contribute to a reduction in animal experimentation. Sena et al. (2010) performed a meta-analysis to estimate the efficacy of tissue plasminogen activator (tPA) in reducing infarct volume in thrombotic occlusion models of ischemic stroke. Using cumulative meta-analysis, the authors examined the effect of sequentially adding findings from studies in a chronological order (based on date of publication) on the overall estimate of efficacy in reducing infarct volume (202 comparisons were collected from eligible studies and included in the meta-analysis). The overall effect decreased with time but stabilized once the effect estimate was based on approximately 1,500 animals (around year 2001). The systematic review, however, identified studies eligible for the meta-analysis published until 2008. Thus, cumulative meta-analysis can be used to reduce unnecessary animal experiments by providing evidence as to when sufficient data from animal testing has been collected.

1.2 Development of systematic review methods in EBT

Guidelines for performing systematic reviews of randomized controlled trials (RCTs) were initially developed by the Cochrane Collaboration and are detailed in the Cochrane Handbook (Higgins and Green, 2011). Systematic review methodology has been adopted by other fields, particularly evidence-based toxicology (EBT) (Stephens et al., 2016); methods based on EBM have been adapted and extended to better accommodate evidence from different study designs and populations of interest (EFSA, 2010; NTP, 2019).

Systematic review methods in EBT have advanced significantly since the need to adapt evidence-based approaches for use in toxicology was recognized in 2005 (Hoffmann and Hartung, 2005), along with the need for a collective effort within the scientific community and other stakeholders (Stephens et al., 2018). Numerous developments have since contributed to advancing EBT (Fig. 1). The First International Forum Towards EBT, organized by the European Commission in 2007, brought together 170 scientists from all continents and launched an initiative to formally apply evidence-based methods in toxicology (Hoffmann et al., 2007; Griesinger et al., 2009). The benefits of introducing evidence-based approaches to toxicological decision-making were explored, and a list of ten defining characteristics of EBT (Hoffmann et al., 2014) was endorsed by the participants.

In 2009, the development of EBT was further facilitated by creating the first chair for evidence-based toxicology at the Johns Hopkins Bloomberg School of Public Health, Baltimore, USA, endowed by the Doerenkamp-Zbinden Foundation, Kreu-

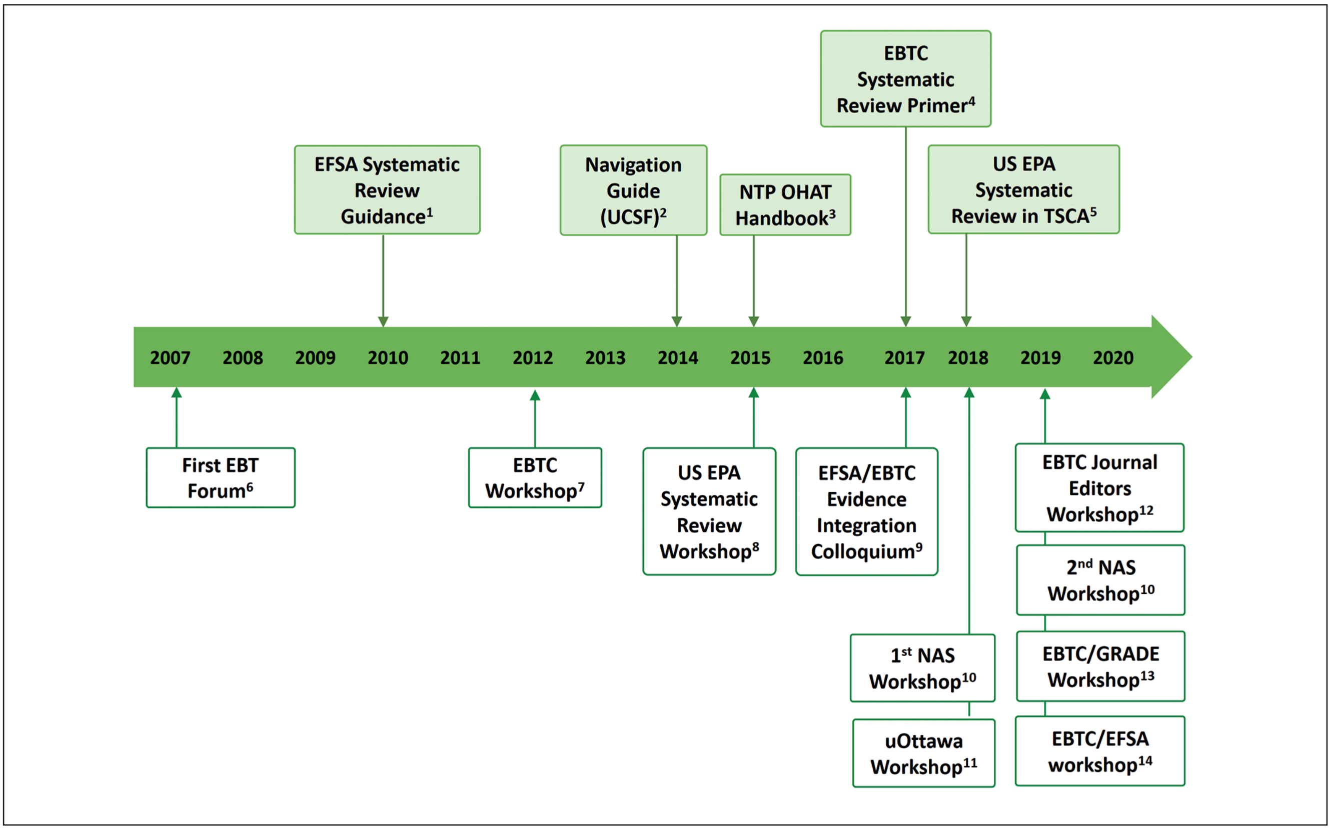
EBTC Systematic Review Primer[4]
EFSA Systematic Review Guidance[1]
Navigation Guide (UCSF)[2]
NTP OHAT Handbook[3]
US EPA Systematic Review in TSCA[5]
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
First EBT Forum[6]
EBTC Workshop[7]
US EPA Systematic Review Workshop[8]
EFSA/EBTC Evidence Integration Colloquium[9]
EBTC Journal Editors Workshop[12]
2nd NAS Workshop[10]
1st NAS Workshop[10]
EBTC/GRADE Workshop[13]
uOttawa Workshop[11]
EBTC/EFSA workshop[14]

Fig. 1: Incorporation of systematic review into evidence-based risk assessment
[1]European Food Safety Authority (EFSA, 2010); [2]University of California San Francisco Navigation Guide (Lam et al., 2014); [3]The National Toxicology Program Office of Health Assessment and Translation Handbook, updated in 2019 (NTP, 2019); [4]Evidence Based Toxicology Collaboration (Hoffman et al., 2017); [5]United States Environmental Protection Agency Toxic Substances Control Act (US EPA, 2018); [6]First International Forum Toward Evidence-Based Toxicology organized by the European Commission; Como, Italy on October 15-18, 2007; [7]Evidence Based Toxicology Collaboration workshop: Evidence-based toxicology for the 21st century – opportunities and challenges; Research Triangle Park, USA on January 24-25, 2012; [8]United States Environmental Protection Agency: Advancing systematic review workshop; Arlington, USA on December 16-17, 2015; [9]Joint European Food Safety Authority and Evidence-Based Toxicology Collaboration Colloquium: Evidence integration in risk assessment: The science of combining apples and oranges; Lisbon, Portugal on October 25-26, 2017; [10]National Academy of Sciences workshops: 1) Strategies and tools for conducting systematic reviews of mechanistic data to support chemical assessments and 2) Evidence integration workshop; Washington, USA on December 10-11, 2018 and June 3-4, 2019, respectively; [11]University of Ottawa workshop: Development of an evidence-based risk assessment framework; Ottawa, Canada on December 17-18, 2018; [12]Evidence Based Toxicology Collaboration workshop for editors of toxicology journals: Assuring the quality of systematic reviews published in toxicology an environmental health journals; Research Triangle Park on May 29-31, 2019; [13]Evidence Based Toxicology Collaboration and GRADE Working Group workshop: The application of evidence-based methods to construct mechanistic frameworks for the development and use of non-animal toxicity tests; Hamilton, Canada on June 13-14, 2019; [14]Evidence Based Toxicology Collaboration and European Food Safety Authority workshop: Advancing the application of evidence-based methods to construct mechanistic frameworks for the development and use of non-animal toxicity tests; Parma, Italy on October 2-3, 2019.

zlingen, Switzerland. A 2012 workshop organized by the Evidence-Based Toxicology Collaboration (EBTC), founded in 2011, was instrumental in clarifying evidence-based approaches and identifying the next steps required to facilitate their application in toxicology (Judson et al., 2013; Silbergeld and Scherer, 2013; Stephens et al., 2013, 2018).

A workshop convened by the US EPA in 2015[1] discussed developments in systematic review methods, particularly in identifying and evaluating evidence from multiple streams – epidemiology, animal toxicology, and mechanistic studies (Stephens et al., 2013). In 2017, a Joint Colloquium between the EFSA and the EBTC brought together experts to address challenges in evi-

[1] Advancing Systematic Review Workshop (2015). https://www.epa.gov/iris/advancing-systematic-review-workshop-december-2015

dence integration in chemical risk assessments and explored future directions (EFSA, 2018). Qualitative and quantitative methodological approaches were explored, along with the need for testing and validating these approaches in toxicology.

More recently, a series of workshops that focused on data integration and quality assessment of data from multiple evidence streams were held. At the request of the US EPA, the US National Academy of Sciences Engineering, and Medicine (NAS) hosted two workshops in 2018[2] and 2019[3] to explore approaches to overcome challenges in evidence integration in chemical risk assessments, particularly when dealing with mechanistic data. State-of-the-art strategies and tools for systematic review of mechanistic data were explored, including screening the literature, study validity assessment, assessing the certainty in individual studies and bodies of evidence, and methods for integration of mechanistic data.

In 2018, the University of Ottawa hosted a workshop that brought together experts to develop a framework for evidence-based risk assessment that ensures that data from multiple evidence streams for health risk assessment is evaluated using an objective, transparent, and comprehensive approach (Krewski et al., 2022). In 2019, a workshop by EBTC and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group, explored how mechanistic data can be better integrated into systematic reviews of chemical risk assessments[4]. This was followed by a collaborative workshop, organized by EBTC and EFSA, to further explore the integration of mechanistic data and examine evidence-based methods to construct mechanistic frameworks (such as adverse outcome pathways (AOPs)) that can be used in the development of non-animal toxicity tests (report in preparation). Additionally, journal editors met in 2019 to discuss strategies that assure the quality of systematic reviews in toxicology and environmental health[5]. Several publications providing helpful guidance on EBT have emerged from this series of meetings and workshops (EFSA, 2010; Woodruff and Sutton, 2014; Hoffmann et al., 2017; US EPA, 2018b; NTP, 2015, 2019).

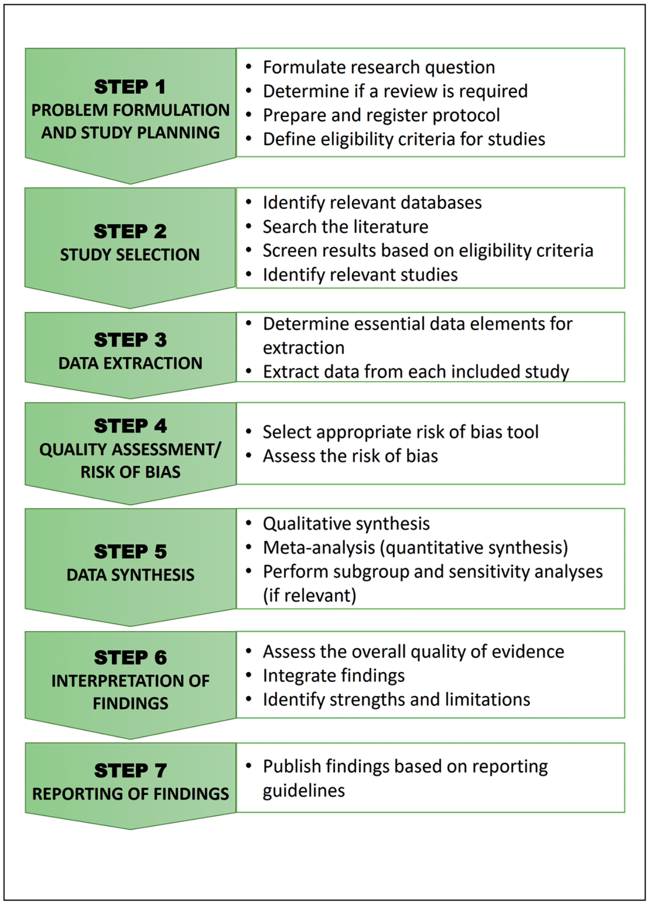
STEP 1
PROBLEM FORMULATION AND STUDY PLANNING
• Formulate research question
• Determine if a review is required
• Prepare and register protocol
• Define eligibility criteria for studies
STEP 2
STUDY SELECTION
• Identify relevant databases
• Search the literature
• Screen results based on eligibility criteria
• Identify relevant studies
STEP 3
DATA EXTRACTION
• Determine essential data elements for extraction
• Extract data from each included study
STEP 4
QUALITY ASSESSMENT/ RISK OF BIAS
• Select appropriate risk of bias tool
• Assess the risk of bias
STEP 5
DATA SYNTHESIS
• Qualitative synthesis
• Meta-analysis (quantitative synthesis)
• Perform subgroup and sensitivity analyses (if relevant)
STEP 6
INTERPRETATION OF FINDINGS
• Assess the overall quality of evidence
• Integrate findings
• Identify strengths and limitations
STEP 7
REPORTING OF FINDINGS
• Publish findings based on reporting guidelines

Fig. 2: Flow diagram of key steps in systematic review

1.3 Approach to performing systematic reviews

In contrast to traditional literature reviews, the approach followed in a systematic review is based on rigorous, objective, and reproducible methods for identifying relevant studies, assessing their quality, and summarizing their findings. Systematic reviews, which may include a meta-analysis (quantitative synthesis), require substantial preparation and planning. Therefore, authors planning to conduct systematic reviews need a clear understanding of the methodology.

This document outlines the key steps for performing a systematic review based on current best practices recommended by experts in the field of evidence synthesis, incorporating recent developments in this area. Systematic review software that has the potential to improve the efficiency at every stage of systematic reviews is also described, and key applications are listed.

2 Key steps in systematic review

Systematic review methods are designed to minimize bias in the selection of studies from the literature to address a clearly defined research question. Methodological guidance for undertaking a systematic review is available from several groups specializing in evidence synthesis, such as the Cochrane Collaboration, the Center for Reviews and Dissemination (CRD), the GRADE Working Group, and the Joanna Briggs Institute (JBI). In addition, many organizations and regulatory agencies have refined the methods to fit their assessment needs. Table 1 lists some of these groups alongside their published guidance or handbooks.

[2] NAS (2018). Strategies and Tools for Conducting Systematic Reviews of Mechanistic Data to Support Chemical Assessments. http://dels.nas.edu/Upcoming-Workshop/Strategies-Tools-Conducting-Systematic/AUTO-5-32-82-N?bname=best

[3] NAS (2019). Evidence Integration Workshop. http://dels.nas.edu/Upcoming-Event/Evidence-Integration-Workshop/AUTO-0-96-15-Q

[4] EBTC (2019). Integrating mechanistic evidence into toxicology systematic reviews. https://youtu.be/NNa0r2qL4pI

[5] EBTC (2019). News. http://www.ebtox.org/news/

Tab. 1: List of select organizations with the corresponding guidance for systematic reviews

Organization	Focus	Published guidance for systematic reviews
Agency for Healthcare Research and Quality	Health care	Methods Guide for Effectiveness and Comparative Effectiveness Reviews (AHRQ, 2014) Methods Guide for Comparative Effectiveness Reviews – Quantitative synthesis (Morton et al., 2018)
Center for Reviews and Dissemination	Health care	Systematic Reviews: CRD's guidance for undertaking reviews in health care (Centre for Reviews and Dissemination, 2008) Guidance specific for reviews of clinical tests, public health interventions, adverse effects, economic evaluations, and qualitative evidence is available
The Cochrane Collaboration	Health care	Cochrane Handbook for Systematic Reviews of Interventions Version (Higgins and Green, 2011) Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (The Cochrane Collaboration, n.d.)
European Food Safety Authority	Food and feed safety assessments	Application of systematic review methodology to food and feed safety assessments to support decision making (EFSA, 2010)
International Agency for Research on Cancer	Cancer hazards	IARC Monographs on the Identification of Carcinogenic Hazards to Humans – Preamble (IARC, 2019)
Joanna Briggs Institute	Health care	Joanna Briggs Institute Reviewer's Manual (Aromataris and Munn, 2017) Guidance specific for reviews of qualitative evidence, quantitative evidence/effectiveness, economic evidence, and text and opinions is available
National Toxicology Program (Office of Health Assessment and Translation)	Chemical risk evaluations	Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration (NTP, 2019)
Navigation Guide[a]	Environmental health	Overview of the Navigation Guide Methodology in environmental health available in Woodruff and Sutton (2014)
U.S. Environmental Protection Agency	Chemical risk evaluations	Application of Systematic Review in TSCA Risk Evaluations (US EPA, 2018b); (*Intended for risk evaluations under the Toxic Substances Control Act*)
SYstematic Review Center for Laboratory animal Experimentation (SYRCLE)	Animal studies	Meta-analyses of animal studies: An introduction of a valuable instrument to further improve healthcare (Hooijmans et al., 2014a).

[a] The Navigation Guide represents a methodology and not a specific organization.

Systematic review methods can be applied to synthesize findings from different types of primary research, including RCTs, observational (non-randomized) human studies, diagnostic test accuracy, animal studies, and mechanistic studies. As the methods employed in these studies can differ extensively, techniques have been adapted to accommodate these variations, such as considering study limitations (i.e., risk of bias, RoB) specific to the type of primary research being reviewed.

Bringing together the best available research on a research question can be challenging and requires substantial preparation and planning. Authors planning a review should ensure they have a qualified and multidisciplinary team. It is recommended that a systematic review team include members with experience in conducting literature searches, including medical librarians, statisticians, methodologists, and content experts, to ensure a high-quality review. In interpreting the results of a systematic review, it is important to consider the methodological quality of the review. The AMSTAR2 (A MeaSurement Tool to Assess systematic Reviews 2) tool (Shea et al., 2017) is useful in helping readers identify high-quality reviews; the ROBIS tool (Whiting et al., 2016) is useful in assessing particularly the overall RoB of a systematic review. An example of the use of AMSTAR2 is presented in (Hersi et al., 2017).

The key steps based on current guidelines involved in a systematic review are summarized in Figure 2.

2.1 Problem formulation and study planning

2.1.1 Research question

The first step in performing a systematic review is to clearly define the research question that will be addressed. It is essential that the question be specific (Khan et al., 2011). A clear, specific, and answerable research question can be guided by the PICO or PECO (Population, Intervention/Exposure, Comparison, Outcome) framework (Moher et al., 2009; Wright et al., 2007; Morgan et al., 2018b) to define the following elements of the question: 1) the population; 2) the intervention (or exposure); 3) the comparator (or control); and 4) the outcome(s) of interest. In a research question for a systematic review in toxicology, the intervention component of PICO is substituted with exposure, which is more appropriate in this context. Formulating a clear question that includes these elements is essential for systematic reviews of effects of an intervention or exposure and is helpful in subsequent steps when determining the study eligibility criteria and developing the literature search strategy (EFSA, 2010). When formulating a PECO for a review of exposure, it may also be helpful to define how much is known about the exposure of interest to clarify the scope of the review (Morgan et al., 2018b).

A well-formulated research question also sets the scope of the review; thus, it should be determined at this stage whether the review is intended to have a narrow or broad scope. A broad-scoped review can serve as a comprehensive summary of a large amount of data; however, it would have a higher likelihood of heterogeneity among studies and would require more resources to perform. A review with a narrower scope might consider a more manageable amount of data, but its findings may not be generalizable beyond a specific population or context (Wright et al., 2007; Counsell, 1997; Higgins and Green, 2011).

The research question facilitates the development of study eligibility criteria for inclusion of studies in the systematic review. Inclusion and exclusion criteria, formulated prior to initiating the review, are features that distinguish systematic reviews from narrative reviews (Hoffmann et al., 2017). Using the PICO components, explicit criteria defining the population (having specific characteristics or conditions, or setting from which they are selected), intervention (mode of delivery, frequency, duration), comparison groups (placebo, standard of care, active control), and outcomes of interest (outcome measure, primary and secondary outcomes) are developed. The types of study designs of interest (RCTs, observational, animal) can also be prespecified (Higgins and Green, 2011). Clearly defined criteria are key to the subsequent screening stage and serve as a guide for reviewers when deciding whether to include (or exclude) a study.

The number of systematic reviews published annually is high. Fontelo and Liu (2018) reported that, in 2015, over 10,000 systematic reviews were identified in PubMed published from the US, UK, China, Australia and Canada. Authors planning a review should therefore confirm whether their research question has already been answered in an existing review, or whether a similar review is underway. A search in PubMed and other bibliographic databases can identify whether recently published systematic reviews have addressed the research question of interest. This important initial check can prevent inefficient use of resources and duplication of effort. Registries of systematic reviews and protocols, such as the International Prospective Register of Systematic Reviews (PROSPERO), the Cochrane Database of Systematic Reviews (CDSR), and the Database of Abstracts of Reviews of Effects (DARE), are all useful sources for identifying ongoing and published systematic reviews. Updating an existing review is a much easier task and may be performed in the interest of efficiency if a high-quality systematic review is available.

2.1.2 Protocol

A protocol for a planned systematic review lays out the research question, review objectives, inclusion and exclusion criteria, and describes all methods to be followed. Methods for the identification of studies to be included in the review based on detailed inclusion and exclusion criteria, the data extraction process, the quality assessment of individual studies, and the qualitative and quantitative analysis should all be described in the protocol.

The protocol should be published in a protocol registry and/or in a peer-reviewed journal. This *a priori* outline of the review methodology supports transparency and reduces potential bias. In addition, the protocol – when published – can be tracked by those planning to conduct a systematic review on the same topic, allowing them to avoid the potential duplication of efforts or support already ongoing activities (Higgins and Green, 2011). Although review authors should adhere to the methods outlined in the protocol while conducting the review, changes and adjustments may be required. In this case, any deviations from the initial protocol should be reported to further promote transparency in the review process.

Protocols for Cochrane reviews must be peer-reviewed and published in the Cochrane Database of Systematic Reviews. For non-Cochrane systematic reviews, registration and publication of the protocols is encouraged before starting the review. Having registered a protocol before starting a review is one of the key items in the critical appraisal tool AMSTAR2, commonly used to assess the quality of systematic reviews (Shea et al., 2017).

Protocols can be registered in public registries available specifically for review protocols such as PROSPERO for reviews of human health research and preclinical animal intervention studies[6], CAMARADES[7] (Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies) for reviews of animal studies, SYRCLE[8] for reviews of non-intervention animal studies and of non-animal/animal-free studies, or in digital archives such as Zenodo[9].

Although there is no standard format for a systematic review protocol, authors can refer to the PRISMA-P (Preferred reporting items for systematic review and meta-analysis protocols) statement (Moher et al., 2015). This framework contains a checklist of 17 items related to administrative information, the introduction, and the methods recommended to be included in a systematic review protocol. While these instruments capture whether specific items in the review were addressed or certain steps were taken, improving the transparency of the document, they do not ensure the methodological quality of those items. Guidance pre-

[6] https://www.crd.york.ac.uk/prospero/

[7] http://www.dcn.ed.ac.uk/camarades/

[8] www.syrcle.nl

[9] https://zenodo.org

sented by de Vries et al. (2015) provides a standard format that includes all the important elements that should be included in a systematic review protocol for animal studies.

2.2 Study selection

The selection of studies to include in a systematic review should involve a structured process that entails a comprehensive search for potentially relevant studies and screening of the search results based on predefined eligibility criteria. This screening stage can be a lengthy process depending on the number of articles retrieved.

2.2.1 Selecting bibliographic database(s)

Performing a thorough search supports efforts to identify studies relevant to the review research question (Khan et al., 2011). Bibliographic databases that are useful for health-related systematic reviews include Embase, MEDLINE, PubMed, Scopus, Web of Science, PsycINFO, CENTRAL, TOXLINE, and CINAHL. Many other subject-specific databases exist. For example PubMed, Embase, and Web of Science are ideal databases when searching for animal studies (Hooijmans et al., 2010). The selection of databases should be made based on relevance to the review topic and is best informed by information specialists or healthcare librarians.

2.2.2 Designing the search strategy

Designing the search strategy requires consideration of the main concepts of the review (the *P*, *I*, *C*, and *O*) and the eligibility criteria defined in earlier stages. The goal of the search is to identify all potentially relevant literature through a well-designed and comprehensive search strategy. The search strategy can be developed by a librarian as it requires a good understanding of search terms and the indexing used by the various databases to be searched. When designing a search strategy, the topic of interest can be divided into major concepts, where relevant search terms are then listed to describe each concept. In addition to the eligibility criteria, the search criteria can include the year of publication, design of studies, and language of publication. The process may require several iterations before a final strategy is adopted. Chapter 6 of the Cochrane Handbook (Higgins and Green, 2011) provides detailed guidance on the selection of search terms and on the combination of the various concepts of the search strategy, all of which need to be carefully considered in order to optimize the search quality. The use of search filters is helpful in narrowing the number of search results to include only potentially relevant references. Many filters are available that can identify studies with certain characteristics such as study design or target population. These filters are usually unique to specific databases, such as the search filters for animal studies available for searches in PubMed (Hooijmans et al., 2010) and Embase (de Vries et al., 2014a). As indexing processes and MESH headings are updated regularly, only up-to-date search filters that reflect these changes should be used in a systematic review to ensure optimal search results.

The search process and strategies should be well-documented with sufficient detail to allow the search to be reproduced by other researchers (Higgins and Green, 2011). In addition to the detailed search strategy, data on the number of records identified from each source database and the date(s) on which each search was performed should be documented.

Review authors may decide to search other sources for potentially relevant studies, such as conference proceedings, reference lists of included articles, reference lists of relevant reviews, grey literature sources (technical reports from government and public health organizations), and registries of ongoing studies such as clinicaltrials.gov[10] (Higgins and Green, 2011; Counsell, 1997). Searching these sources can complement the bibliographic database search to ensure all potentially relevant studies are identified and screened for relevance.

2.2.3 Reference management

With the large volume of literature accessible through the databases queried, it is possible that the search returns hundreds or thousands of references, each of which will need to be assessed for eligibility. For example, in a recent systematic review by the NTP on the effects of fluoride on learning and memory, over 4,500 animal studies were screened for eligibility (NTP, 2016). Articles retrieved from all databases should be imported into a reference management software. Such software offers many features, including automated organization, management, and citation functions, which are essential when managing large bibliographic databases. Duplicate references will likely be present in the search results if multiple sources or databases are queried, and detection and removal of duplicates can be performed within most reference management tools.

2.2.4 Screening results

Potentially relevant studies found during searches are screened to determine whether each meets the eligibility criteria. Typically, this is performed in two stages. The first stage involves a review of the titles and abstracts of all studies retrieved, with those clearly not meeting eligibility criteria being excluded. Studies that likely meet the inclusion criteria or studies whose eligibility cannot be assessed based on the title and abstract are moved to full text review.

Identified references are typically independently screened by two reviewers. This stage can be the most time-intensive stage that can take hundreds of hours, depending on the number of references identified (Carver et al., 2013). Disagreements between reviewers are resolved by discussion or by a third review author. The assessment by two reviewers is intended to prevent the introduction of bias in the selection of studies, where any study found relevant by either reviewer is included for further assessment (Wright et al., 2007). The screening process can be structured by

[10] https://clinicaltrials.gov/

the creation of screening forms that serve to guide reviewers' decisions when assessing a study. The number of studies at each screening stage and reasons for exclusions should be recorded and reported in subsequent reports and publications.

Both stages of the screening process can be facilitated using software designed for study screening or conducting systematic reviews. The use of such software greatly increases the efficiency of the screening stage, particularly for systematic reviews with many potentially relevant studies. Systematic review software typically records reviewers' evaluations and automatically moves articles through the various stages of the review based on user-defined settings and criteria. The software also records reasons for study exclusion that can be summarized. Software tools that can be used in the systematic review process, including screening results, are described in (Tab. S1, S2[11]).

2.3 Data extraction

All studies retained following full-text review are eligible for data extraction. Relevant data are extracted from each study, ideally independently by two reviewers where disagreements are resolved by discussion or by a third person. Alternatively, at a minimum, the second extractor checks data extracted by the first reviewer for accuracy (Centre for Reviews and Dissemination, 2008), or two reviewers independently extract only those data (for example, outcome data) that are critical for interpreting the results (Higgins and Green, 2011). Although duplicate data extraction may be time-intensive, it has been shown to result in fewer data errors than the latter approaches. In a study comparing the frequency of errors and time requirement for single and double data extraction, Buscemi et al. (2006) report that single data extraction required 36.1% (relative difference) less time but resulted in 21.7% (relative difference) more errors compared to duplicate extraction.

The extracted data will vary based on the scope of the review. In general, data should be collected on the methods, participants/population, intervention or exposure, outcomes, and results. Many reviews also extract data related to the study authors' conclusions and funding details. Other items for which reviewers may collect data depend on the review research question, objectives, and search strategy. The Cochrane Handbook and the OHAT Handbook list key data extraction elements that are typically extracted from human studies (NTP, 2019) as well as from animal and *in vitro* studies (Higgins and Green, 2011; NTP, 2019; Hooijmans et al., 2014a). In some instances, there may be missing information in the full-text publication. It may therefore also be useful to extract study authors' contact information for requesting missing data. Rules for the number of times a study author will be contacted and the eligibility of the study for the review if the author does not provide the information needed should be decided and recorded in the protocol.

Studies included in a toxicological systematic review are likely to be of complex study design, rendering extraction of relevant data a complex task. For example, animal and *in vitro* studies require extraction of data on the animal model or cell/tissue type, source of animals or cells, treatment characteristics (purity, dose, vehicle, route of administration), guideline compliance, results (e.g., no observed effect level (NOEL), lowest observed effect level, (LOEL), benchmark dose (BMD)). Although many of the software applications available support systematic reviews in any field, software tools such as HAWC (Health Assessment Workspace Collaborative) are available for risk assessments and human health assessments, allow the use of flexible data entry forms, and can handle complex data extraction.

Structured electronic forms (spreadsheets, word processing documents) for collecting data can be designed for all data items determined essential for the review. At this stage, information required for quality assessment or RoB can also be extracted from included studies. It is recommended that reviewers pilot the data extraction forms on a small sample of studies before adopting the forms for the full review (Wright et al., 2007). This will allow revisions to be made before the final version is used for all included studies.

The data extraction process can also be performed by using systematic review software. Customized extraction forms can be designed to facilitate this stage by storing and validating extracted information. Recent software applications include features that allow automated extraction of pre-specified data fields.

2.4 Quality assessment

Assessment of the limitations of each included study is a fundamental step in performing a systematic review, as including studies with higher RoB can bias the findings of a review (Egger and Smith 2001). Many instruments are available for evaluating the RoB, some of which are specific to a particular study design while others are intended for a range of designs (Viswanathan et al., 2017).

Earlier quality assessment tools represented scales or checklists and included items that do not assess the internal validity of studies (such as power calculations and inclusion/exclusion criteria) (Higgins et al., 2011). Newer quality assessment instruments have focused on the internal validity by assessing RoB of the reported findings. The Cochrane Collaboration developed the RoB tool, intended for randomized trials, which has become popular in Cochrane and non-Cochrane systematic reviews (Jorgensen et al., 2016). This tool is neither a scale nor a checklist but rather a domain-based instrument that assesses each of five potential sources of bias within a study: selection bias, performance bias, detection bias, attrition bias, and reporting bias. The RoB in each domain is rated as either *high risk*, *low risk* or *unclear risk* by answering specific questions for each domain. For example, if a study did not blind participants, personnel, and outcome assessors, it would be judged to have *high risk* of performance bias. Clear guidance and examples for each level of bias for each domain are presented in the Cochrane Handbook (Higgins et al., 2011).

[11] doi:10.14573/altex.2004111s

Two scoring tools that are extensively used to assess the methodological quality of non-randomized studies (NRS) are the Newcastle-Ottawa Scale (NOS)[12] and the Downs and Black (Downs and Black, 1998). Although the use of scales that yield a summary score is not encouraged in the current Cochrane handbook, the NOS is preferred over the Downs and Black due to its simpler application and time efficiency (Higgins and Green, 2011). Both instruments include criteria that correspond to aspects of the internal validity (findings are free from bias) as well as the external validity (generalizability of the findings) of a study (Sterne et al., 2016a).

More recently, however, the ROBINS-I (Risk Of Bias In Non-randomized Studies of Interventions) tool (Sterne et al., 2016a) was developed to provide a RoB tool that is suited for NRS. ROBINS-I assesses the potential for bias within NRS by evaluating them against the ideal target trial. ROBINS-I assesses seven domains through which bias may be introduced to findings of NRS: confounding, selection of participants, classification of interventions, deviation from intended interventions, missing data, outcome assessment, and selection of reported results. For each domain, the tool includes a series of signaling questions that are intended to guide the selection of the level of potential risk. Explicit judgements for the RoB in each domain allow review authors to evaluate the overall RoB for findings in a study. Additional information is available in a detailed ROBINS-I guidance document (Sterne et al., 2016b).

Assessing the RoB in toxicological studies is also a crucial step in systematic reviews, which has been adapted for use with toxicological evidence largely based on the RoB tool for clinical studies (Woodruff and Sutton, 2014; Hooijmans et al., 2014b; Krauth et al., 2014). For example, the NTP developed the OHAT RoB tool to facilitate the assessment of internal validity using common methods and categories across the various evidence streams (NTP, 2015). The OHAT tool consists of 11 questions that assess the potential bias from selection, confounding, performance, attrition/exclusion, detection, and selective reporting. Questions are available to guide the assessment of the level of bias for each bias domain. The RoB is assessed for each outcome of interest in a study. Other relevant approaches include the RoB assessment within the Navigation Guide methodology for human and animal studies (Woodruff and Sutton, 2014) and SYRCLE's RoB tool, a version of the Cochrane RoB tool, adapted for experimental animal studies (Hooijmans et al., 2014b).

Science in Risk Assessment and Policy (SciRAP[13]) is a useful, publicly available web-based tool for evaluation of the reliability and relevance of toxicity data. The tool provides clear criteria for evaluating the reporting quality and the methodological quality of both *in vitro* and *in vivo* toxicity studies. It also provides checklists of items that should be reported in *in vitro* and *in vivo* studies such that researchers report their findings in a structured and transparent manner that meets regulatory requirements (Molander et al., 2015; Beronius et al., 2018).

Review authors interested in quality assessment of toxicological studies should be aware of the diverse methods available for quality assessment. Samuel et al. (2016) summarize and provide guidance on the extensive literature related to assessing the methodological and reporting quality of relevant studies.

RoB assessment for each study included in the review should be completed independently by two reviewers where disagreements are resolved by discussion or by a third person (Higgins and Green, 2011; NTP, 2019). It is also important to note that the RoB should be assessed for each outcome reported in the study, since each outcome may have different sources of bias (Viswanathan et al., 2017; Higgins and Green, 2011).

2.5 Data synthesis

The characteristics and findings of studies included in a systematic review are typically summarized qualitatively in tabular form. The format and level of detail of the descriptive tables vary considerably between reviews. Decisions on which items to include in the tables will depend on the study characteristics deemed important based on the research question (Khan et al., 2011).

Quantitative analysis may also be performed in the form of meta-analysis, where effect estimates from two or more studies are pooled using statistical methods. Results of meta-analysis can be graphically displayed using forest plots to allow examination of the differences in effects across studies as well as the overall combined effect estimate (Higgins and Green, 2011).

The decision on whether to synthesize findings from all included studies can be subjective and often requires careful consideration and input from topic-specific experts (AHRQ, 2014; Higgins and Green, 2011). The AHRQ (2014) suggests that findings from comparative effectiveness research should be combined only when they are similar clinically (e.g., in terms of characteristics of participants, study setting, types of outcomes, interventions) and methodologically (e.g., in terms of study design and RoB). Guidance from the NTP (2019) states that combining studies may not be appropriate in cases where exposure or outcome data vary considerably, the RoB is high, or in cases where combining findings does not lead to meaningful results. In instances where significant heterogeneity among included studies is present, it may be appropriate to present the evidence collected in the form of a narrative synthesis (NTP, 2019).

When similar studies are combined by meta-analysis, the statistical heterogeneity, which represents the variability of the effect estimates across combined studies (Higgins and Green, 2011), should be examined. In forest plots, heterogeneity can be visually inspected by examining the overlap of confidence intervals of individual effect estimates across studies. Quantitatively, heterogeneity can be assessed by the I^2 statistic, which is a "*measure of proportion of total variance in the pooled effects that is due to variance, as opposed to random variation quantitatively*" (Higgins and Green, 2011). I^2 values are usually automatically calculated in meta-analysis software such as Review Manager.

[12] The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

[13] https://www.scirap.org

Cochrane proposes several options for review authors when there is considerable statistical heterogeneity among a group of studies. These include:

1. Do not perform a meta-analysis, as the summary effect may be misleading;
2. Ignore heterogeneity and perform a fixed-effect meta-analysis;
3. Perform a random effects meta-analysis in cases where heterogeneity cannot be explained;
4. Exclude outliers if there are apparent reasons for the corresponding extreme results; or
5. Explore heterogeneity by subgroup analysis or meta-regression (Higgins and Green, 2011).

Meta-regression is not recommended in meta-analyses that consist of less than ten studies (Higgins and Green, 2011). However, when feasible, this method can describe between-study variation (AHRQ, 2014) by examining the effect of study characteristics (explanatory variables) on the summary effect estimate of a meta-analysis.

Examining experiences with previous meta-analyses can also help guide the practical application of this technique. For example, the case studies presented by Goodman et al. (2015) are instructive in appreciating how meta-analyses can be used in different research contexts.

2.6 Interpretation of findings

Findings of a systematic review should be presented and discussed considering all the factors that may have led to the observed findings, particularly the overall quality and strength of the evidence. This step is important as the discussion can serve as a guide for policymakers to make regulatory decisions based on the best available scientific evidence (Wright et al., 2007; EFSA, 2010). Biologic variations that may influence the effect of the intervention, variations in the context and culture, patient adherence to the intervention, and the values and preferences of groups of patients should also be considered when discussing review findings (Higgins and Green, 2011). Other factors to consider when interpreting the findings are the potential limitations of the review process and agreements or disagreements of the findings with the other studies or reviews (EFSA, 2010).

Assessing the certainty of the evidence (e.g., quality or confidence in the estimated effects) for each outcome in a systematic review can be presented in a summary of findings (SoF) table using the GRADE approach (Guyatt et al., 2011). A SoF table succinctly presents the body of evidence from each outcome within a systematic review, specifically including study design, evidence assessment ratings, relative and absolute effects, and the certainty of the evidence for an outcome (Guyatt et al., 2013). Within the GRADE approach, a body of evidence corresponding to randomized studies is assigned an initial certainty of *high*. On the other hand, a body of evidence corresponding to NRS is assigned an initial certainty of *low* to account for the lack of a prognostic balance between the comparison groups.

The SoF table presents the results of the evidence assessment, which is based on five factors that reduce certainty in the estimate of effect: RoB, inconsistency (i.e., heterogeneity), indirectness, imprecision, and publication bias. There are three factors that may increase certainty in the estimate of effect: large or very large magnitude of effect, dose-response gradient, and opposing residual confounding. The factors that may increase certainty apply to those outcomes informed by NRS. The overall GRADE certainty ratings are (Schünemann et al., 2013a):

1. *high* suggesting that we are very confident that the true effect lies close to that of the estimate of the effect;
2. *moderate* suggesting that the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different;
3. *low* suggesting that the true effect may be substantially different from the estimate of the effect; and
4. *very low* suggesting that the true effect is likely to be substantially different from the estimate of effect.

The GradePro software[14] is publicly available for creating a *synthesis of evidence* table summarizing the GRADE process. Detailed guidance on GRADE is available in the GRADE Handbook (Schünemann et al., 2013a).

In recent years, the Environmental Health Project Group within the GRADE Working Group has developed additional approaches and examples for the assessment of evidence from environmental and occupational health studies. In particular, Hooijmans et al. (2018) explain how the GRADE approach can be used to evaluate the certainty of evidence of preclinical animal studies that assess the efficacy and safety of therapeutic interventions. In their first version showing how the main principles of GRADE can be used to assess evidence from preclinical animal studies, the authors explain that further research is required to better operationalize the specific domains used in GRADE.

Other organizations have adapted GRADE for use in environmental health and animal studies, such as NTP OHAT and the Navigation Guide (NTP, 2019; Morgan et al., 2016; Woodruff and Sutton, 2014; Morgan et al., 2019). The NTP OHAT has adapted the GRADE method to use with observational epidemiological studies and to integrate evidence from multiple streams (human, animal, *in vitro*) (NTP, 2019).

2.7 Reporting of findings

It is recommended that review authors refer to the PRISMA guidelines to ensure high-quality reporting of a systematic review (Moher et al. 2009). In addition to the PRISMA flowchart (Fig. 3), the guidelines include a list of 27 items that should be reported in a published systematic review. Items are categorized by each section that should be included in the systematic review (abstract, introduction, methods, results, discussion, and funding). The PRISMA statement has been endorsed[15] by numerous journals and editorial organizations including the CRD and Cochrane. COSTER (The conduct of systematic reviews in toxicology and environmental health research), which offers equiva-

[14] https://gradepro.org/

[15] http://prisma-statement.org/Endorsement/PRISMAEndorsers

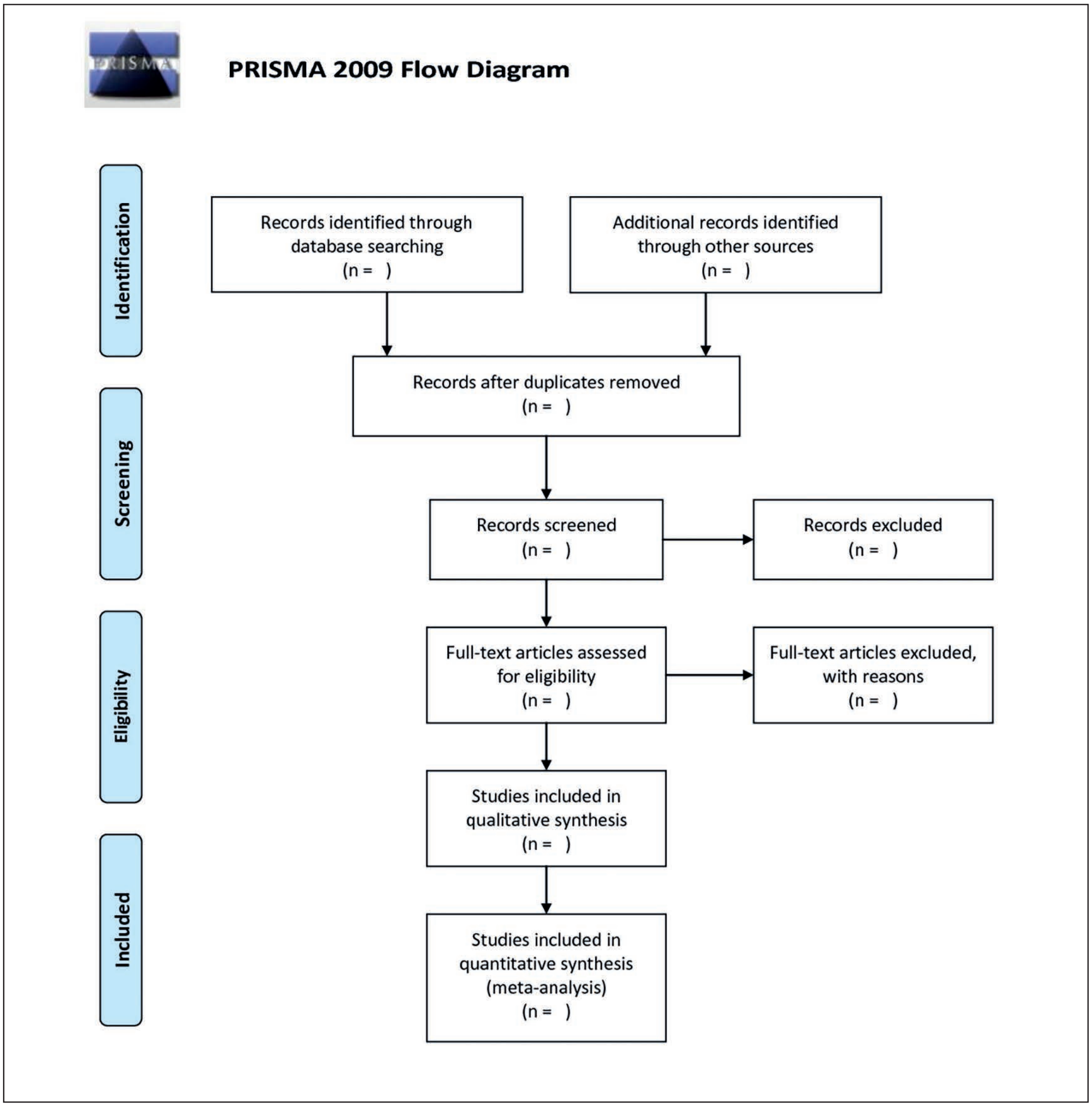
PRISMA
PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through database searching (n =)
Additional records identified through other sources (n =)
Records after duplicates removed (n =)
Records screened (n =)
Records excluded (n =)
Full-text articles assessed for eligibility (n =)
Full-text articles excluded, with reasons (n =)
Studies included in qualitative synthesis (n =)
Studies included in quantitative synthesis (meta-analysis) (n =)

Fig. 3: PRISMA flow chart
Source: Moher et al. (2009)

lent guidance on designing and conducting systematic reviews in toxicology and environmental health research, will be available soon. The guidelines consist of 70 requirements for sound systematic reviews agreed on by a group of experts (Whaley et al., in preparation).

3 Applications in evidence-based risk assessment

3.1 Clinical studies

Systematic review methods were first developed to support evidence-based decisions in healthcare. Thus, methods are well established and have been widely applied to assess evidence from clinical studies. The steps described in the previous section, including assessment of RoB and evidence assessment using GRADE methods, are largely applicable to clinical studies.

A recent Cochrane review that evaluated the effectiveness of prophylactic vaccination against the most carcinogenic human papillomavirus (HPV) types examined evidence from 26 RCTs (Arbyn et al., 2018). Using the Cochrane RoB tool, a rating of *low risk* was assigned to the majority of included studies. Approximately one third of the studies were rated as having an *unclear risk* for the bias domains of allocation concealment, blinding of participants and personnel, and blinding of outcome assessment. Using the GRADE approach, the certainty of the evidence was initially rated as *high* (all evidence was from RCTs) and downgraded in cases where there were concerns with RoB, inconsistency, imprecision, indirectness, or publication bias, as appropriate. For the main comparison, the effect of HPV vaccine (compared to placebo) on cervical lesions in adolescent girls and women who were negative for hrHPV DNA at baseline, the certainty (quality) of evidence was graded to be *moderate* or *high* for all of the outcomes assessed and was downgraded only for serious imprecision in some of the studies. Multiple analyses for various HPV types, vaccine doses, baseline HPV DNA status, and age groups were presented. Among other findings, the review authors conclude that HPV vaccines protect against cervical precancer (RR 0.05; 95% confidence interval (CI): 0.03 to 0.10) in women between 15-26 years of age who are negative HPV types 16 and 18 (HPV16/18) at baseline, with the evidence being of *high* certainty. Among women 24-45 years of age who are HPV16/18 negative, the protective effect was lower (RR 0.30; 95% CI: 0.11 to 0.81), with the evidence being of *moderate* certainty.

Another Cochrane review assessed the efficacy (and other outcomes) of cannabis-based medications for relief of neuropathic pain in adults by examining findings from 16 RCTs (Mücke et al., 2018). Using the RoB tool, the rating of *unclear risk* was assigned to many of the studies for the following bias domains: blinding of participants and personnel, blinding of outcome assessment, and incomplete outcome data domains. Twelve of the studies were judged to be of *moderate* certainty based on the RoB tool. The review concludes, among other findings, that cannabis-based medicines are effective (risk difference 0.05; 95% CI: 0.00-0.09) in achieving 50% or more pain relief in adults (compared with placebo), with the certainty of evidence being *low*. The certainty of evidence for this outcome was assigned a *low* rating as it was downgraded due to concerns with indirectness and imprecision. This certainty rating suggests that future research is very likely to have an impact on the confidence in the estimate of effect calculated.

3.2 Human observational studies

NRS contribute to systematic reviews in three distinct ways, by providing complementary, sequential, or replacement evidence to RCTs (Schünemann et al., 2013b). Many reviews start by trying to identify a body of evidence to answer their PICO/PECO from randomized studies. The rationale for this is that RCT have the potential to provide higher certainty evidence of the effect of an intervention on an outcome of interest; however, in situations when the evidence from RCTs is of lower certainty or non-existent (due to monetary or ethical issues), NRS may be more informative for the review. For example, NRS may complement RCT evidence by providing additional information on the generalizability of an intervention in different populations, possible interaction effects, or different baseline risk estimates (Schünemann et al., 2013b). NRS studies may provide information that has yet to be obtained by RCTs, including long-term outcomes or correlations between surrogate outcomes and patient-important outcomes. Lastly, in situations when RCTs are lacking, NRS can provide higher quality direct evidence to answer the research question.

The grading of a body of evidence from NRS is similar to that from RCTs. The certainty across the body of evidence for an outcome informed by NRS is assessed for concerns with RoB, inconsistency, indirectness, imprecision, and publication bias. If there are no reasons to rate down, then the evidence may be considered for rating up. For example, a systematic review of 18 cohort studies demonstrated that among persons with hepatitis C virus infection, a sustained response to treatment (i.e., sustained virologic response) reduces the risk of development of hepatocellular carcinoma by 76% (relative risk: 0.24; 95% CI: 0.18-0.31) (Morgan et al., 2013). For the outcome of development of hepatocellular carcinoma, the body of evidence started at the initial certainty of *low* because they were NRS. There were no serious concerns about the body of evidence, therefore the evidence was not rated down; however, the magnitude of the reduction of the development of liver cancer was so large that the body of evidence was rated up to *moderate*.

In another example of a recent systematic review of 35 RCTs and 9 NRS, the risk of adverse cardiovascular events in patients treated with non-ergot dopamine agonists was assessed (Crispo et al., submitted). The study examined multiple comparisons. For the main comparison, pramipexole compared to no treatment, two NRS (case-control) studies using electronic health records reported on the risk of heart failure (the main outcome). The pooled effect estimate of two of these studies suggested a moderate increase (adjusted odds ratio 1.46; 95% CI: 1.03-2.08) in the risk of heart failure among pramipexole-treated patients. As this study used the Newcastle-Ottawa Scale for quality assessment of NRS, the evidence for this comparison was assigned an initial *low* rating, based on GRADE guidelines. This rating was neither upgraded nor downgraded during the GRADE, and the final certainty in the evidence was rated as *low*.

With the introduction of the recent ROBINS-I tool for assessing the RoB in observational studies, the evidence from NRS is assigned an initial high certainty rating within the GRADE procedure (Morgan et al., 2018a; Sterne et al., 2016a; Morgan et al., 2019). The reason for this rating is that ROBINS-I assesses NRS against the ideal randomized clinical trial and examines the potential for bias introduced by the lack of a prognostic balance between the intervention and comparison groups, essentially accounting for the potential for residual and unknown confounding. The application of these instruments within a systematic review may be desirable among groups wanting to start at an initial certainty rating of *high* within GRADE; however, it has the potential to lead to the same final certainty in the evidence ratings (Schünemann et al., 2019).

3.3 Experimental animal studies

The application of systematic reviews in the toxicological field has received more attention in recent years as it offers transparent, objective, and reproducible means to inform risk decisions (Hoffmann et al., 2017). Making greater use of findings from published animal studies can avoid unnecessary duplication. The inclusion of human and mechanistic data in risk assessment can also reduce the need for animal test data. Optimizing the use of these data sources in the context of systematic reviews can contribute to the 3Rs. Many methodological tools (such as RoB tools) are currently being adapted from clinical research to better meet the specific characteristics of toxicological research.

Although systematic reviews in toxicology follow the main steps discussed earlier in this article, the specific characteristics of toxicological data require careful consideration, such as extraction of complex study design elements. There are guidelines, specifically for systematic review of animal studies, such as a publication by SYRCLE (SYstematic Review Centre for Laboratory animal Experimentation) on how to identify all potentially relevant evidence (Leenaars et al., 2012), guidance for conducting meta-analysis on findings from animal studies, as well as exploring heterogeneity between studies (Hooijmans et al., 2014a; Vesterinen et al., 2014), and the SYRCLE RoB tool (Hooijmans et al., 2014b).

One example of a systematic review that included toxicological studies assessed evidence from 21 animal studies and 18 observational studies using methods of the Navigation Guide (Woodruff and Sutton, 2014) to investigate the association between developmental exposure to perfluorooctanoic acid (PFOA) and fetal growth outcomes (Lam et al., 2014). To accommodate evidence from human and animal studies, two PECO questions were specified at the onset, and data from each stream was kept separate in the initial stages of the review. The RoB for each study was rated by assessing the risk (*low*, *probably low*, *probably high*, or *high* risk) for each domain specified in the Navigation Guide methodology (similar domains to the Cochrane RoB tool). The quality of the evidence was then rated across all studies within each evidence stream. The nonhuman evidence, assigned a prespecified initial rating of *high* quality, was further evaluated separately for mammalian data and non-mammalian data. The quality rating across all the nonhuman mammalian body of evidence was downgraded to *moderate* due to RoB concerns across the individual studies, while the rating across the non-mammalian evidence was downgraded to *low* due to concerns with RoB across individual studies in addition to indirectness. The strength of human and nonhuman evidence streams was rated separately, considering the quality of the body of evidence, the direction of the effect, the confidence in the effect estimate, and other factors that may impact certainty. The quality of both evidence streams was rated as *moderate*, along with a high level of confidence in the association of increased exposure to PFOA and decreased birth weight. Finally, human and nonhuman evidence streams were integrated, concluding that PFOA be classified as *known to be toxic* based on the availability of *sufficient evidence of toxicity*.

3.4 New approach methodologies

With the increasing use of NAMs in evidence-based risk assessment, systematic review methods can be applied to summarize findings from these studies as well as to validate and support the use of these alternative methods with greater confidence.

Kemppinen et al. (2011) conducted a systematic review of genome-wide expression studies to better understand the molecular basis of multiple sclerosis (MS) and to identify genes that show differential expression. After screening identified records from the database searches, the review identified eight (microarray) studies that reported on 2,017 unique genes with increased expression and 1,860 genes with decreased expression in MS. A total of 229 genes with differential expression to the same direction were found to have been reported in at least two studies, 12 of which were reported in at least three studies. The review further explored the relationship between the 229 genes identified with known immunological pathways. Twenty pathways – the glucocorticoid receptor signalling pathway being the most common – were found to be significantly associated with the differentially expressed genes. The authors noted that the findings extracted were not directly comparable due to heterogeneity among studies, including differences in samples, data quality control, and the definition of differential expression. This review was the first systematic review conducted on microarray studies in MS: its findings can be used to direct future *in silico* studies to focus on specific genes and pathways when studying MS.

In another example, Bronsveld et al. (2015) conducted a systematic review to examine the available evidence on the association between insulin analogues and breast cancer in diabetics after several epidemiological studies questioned the link between insulin and cancer risk. The review included *in vitro*, animal, and human studies. Sixteen *in vitro* studies in which protein and gene expression for breast cell lines was assessed to determine the mitogenic properties of insulin analogues were identified. Among the marketed insulin analogues, an increased proliferative potential was reported for glargine in seven of the *in vitro* studies. However, findings from the epidemiological studies reviewed did not support an association between glargine use in diabetic patients and increased breast cancer risk. The authors note that these studies had significant differences in study design and had relatively short follow-up times. Insulin-induced breast cancer mitogenesis is poorly understood, and the clinical relevance of increased mitogenicity for glargine remains unclear. Based on the findings of this review, the possibility for glargine to induce breast tumor progression by upregulating mitogenic signaling pathways cannot be excluded.

4 Systematic review software

The use of software when preparing systematic reviews has become essential, particularly for the management of the large numbers of references identified by literature searches. Software for systematic reviews has features that can support multiple stages of the review and increase overall efficiency such as literature search, importing references, detection of duplicate references, screening, managing included/excluded references, reviewer conflict resolution, data extraction, RoB assessment, meta-analysis, diagram preparation (funnel plots, Prisma diagrams), in addition to having multiple users working on the same project.

New features are constantly being integrated into the software to facilitate the review process. For example, screening based on natural language processing technology has recently emerged and is being incorporated into systematic review software. In this case, the automated reference screening can function as a second reviewer for improved time efficiency. This feature is relatively recent and current assessments on automated screening indicate suboptimal performance. For example, two recent studies tested the automated screening process in three systematic reviews and found that the software did not correctly screen all relevant articles (Gartlehner et al., 2019; Gates et al., 2019). Until the performance of this feature improves, such tools may be beneficial in supporting rapid reviews, which do not attempt to identify the total body of evidence on a specific topic (Gates et al., 2019). Tools that currently support automation/machine learning include Distiller, EPPI-Reviewer, Rayyan, and SWIFT Active Screener (Van der Mierden et al., 2019).

A recent paper assessed 16 tools based on the number of mandatory, desirable, and optional features each application has (Van der Mierden et al., 2019). Mandatory features included multiple users, importing/exporting references, distinct screening stages (title/abstract and full text) and reference allocation, while desirable and optional features included keyword highlight, free to use, attaching pdfs, flow diagram creation, and machine learning/automation. Of the software that requires a paid license, Distiller was found to support the most features, followed by EPPI-Reviewer, SWIFT Active Screener, and Covidence. Of the free tools, Rayyan supported the highest number of mandatory features, followed by SysRev and CADIMA. Distiller and Rayyan are briefly described below.

Distiller[16] is a web-based tool that supports many stages of a systematic review and is not specific to a particular research discipline. Settings for the screening stages (title/abstract and full text) can be controlled by the user to assign references to reviewers. Screening forms can be created and rules for allocating references as included, excluded, or conflicted can be customized. The program provides the users flexibility in creating data extraction forms and quality assessment forms. A wide range of customizable reports can be generated from the screening process to display extracted data. A recently added module (DistillerAI) integrates artificial intelligence to increase the efficiency of screening by having the program function as a second reviewer. The software requires a paid license but currently offers a limited free subscription to students.

Rayyan[17] is a free web-based tool that allows multiple users to collaborate on the same project. One downside of this software is the lack of distinct screening stages (title/abstract and full text). Instead, screening of studies is performed in one stage where users classify references as included or excluded. A mobile application is also available that allows screening of references with the option of working offline. Rayyan also can predict which studies should be included based on decisions on studies that have already been screened. Other features including better detection of duplicates, RoB assessment, and automatic extraction of data may be added in the future (Ouzzani et al., 2016) in addition to having distinct title and abstract and full text screening stages (Van der Mierden et al., 2019).

When selecting a tool, review authors are encouraged to consider the complexity and size of the review, as well as the available funding. Software tools for systematic reviews are described in Table S1[11], and the review stages supported by selected software are summarized in Table S2[11].

5 Conclusions

Systematic reviews are becoming increasingly common in healthcare, public policy, and evidence-based toxicology. Methods for systematic reviews in toxicology and environmental health have evolved over the last decade, and numerous collaborative efforts continue to adapt approaches to better accommodate evidence from multiple evidence streams. The use of systematic reviews to identify all relevant published findings supports the optimal use of evidence from human, animal, and mechanistic studies, thereby contributing to the 3Rs. By relying more on human and mechanistic evidence and including data derived from the increasing number of validated new approach methodologies, reliance on animal test data can be reduced. High-quality reviews of animal studies can also prevent unnecessary duplication in animal experiments conducted in support of human health risk assessment. Replacement of animal studies may also be achieved by using systematic reviews to summarize findings from the application of NAMs and provide validation for these new approaches.

When conducted following a structured, transparent, and systematic process, as outlined in this paper, findings from systematic reviews can represent the best available evidence to answer a clearly defined question. Methods for RoB assessment, certainty assessment, and interpretation of findings are evolving as international regulatory agencies are in the process of incorporating the human observational as well as *in vivo* and *in vitro* toxicological evidence in regulatory decision frameworks. Available software tools, continually being developed to include more features, provide support to many steps of a systematic review, allowing faster completion of the review and management of large numbers of references. In addition, software programs can assure transparency and independence of reviewers in the process. Several published guidelines and handbooks referenced in this document describe the methods and techniques for conducting a systematic review in detail and represent a valuable resource for authors planning a systematic review.

References

AHRQ (2014). *Methods Guide for Effectiveness and Comparative Effectiveness Reviews.* Publication No. 10(14)-EHC063-EF. Rockville, MD, USA: Agency for Healthcare Research and Quality.

Andersen, M. E., McMullen, P. D., Phillips, M. B. et al. (2019). Developing context appropriate toxicity testing approaches using new alternative methods (NAMs). *ALTEX 36*, 523-534. doi:10.14573/altex.1906261

Arbyn, M., Xu, L., Simoens, C. et al. (2018). Prophylactic vaccination against human papillomaviruses to prevent cervical cancer and its precursors. *Cochrane Database Syst Rev 5*, CD009069. doi:10.1002/14651858.CD009069.pub3

Aromataris, E. and Munn, Z. (2017). *Joanna Briggs Institute Reviewer's Manual*. The Joanna Briggs Institute. https://reviewersmanual.joannabriggs.org/

Banwell, V., Sena, E. S. and Macleod, M. R. (2009). Systematic review and stratified meta-analysis of the efficacy of interleukin-1 receptor antagonist in animal models of stroke. *J Stroke Cerebrovasc Dis 18*, 269-276. doi:10.1016/j.jstrokecerebrovasdis.2008.11.009

Beronius, A., Molander, L., Zilliacus, J. et al. (2018). Testing and refining the science in risk assessment and policy (SciRAP) web-

[16] www.evidencepartners.com

[17] https://rayyan.qcri.org

based platform for evaluating the reliability and relevance of in vivo toxicity studies. *J Appl Toxicol 38*, 1460-1470. doi:10.1002/jat.3648

Bronsveld, H. K., ter Braak, B., Karlstad, Ø. et al. (2015). Treatment with insulin (analogues) and breast cancer risk in diabetics; a systematic review and meta-analysis of in vitro, animal and human evidence. *Breast Cancer Res 17*, 100. doi:10.1186/s13058-015-0611-2

Buscemi, N., Hartling, L., Vandermeer, B. et al. (2006). Single data extraction generated more errors than double data extraction in systematic reviews. *J Clin Epidemiol 59*, 697-703 doi:10.1016/j.jclinepi.2005.11.010

Carver, J. C., Hassler, E., Hernandes, E. et al. (2013). Identifying barriers to the systematic literature review process. Paper presented at the 2013 ACM / IEEE international symposium on empirical software engineering and measurement. https://doi.org/10.1109/esem.2013.28

Centre for Reviews and Dissemination (2008). *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care*. Centre for Reviews and Dissemination, University of York. https://www.york.ac.uk/crd/guidance/

Counsell, C. (1997). Formulating questions and locating primary studies for inclusion in systematic reviews. *Ann Intern Med 127*, 380-387. doi:10.7326/0003-4819-127-5-199709010-00008

Crispo, J., Farhat, N., Fortin, Y. et al. (submitted). Non-ergot dopamine agonists and the risk of heart failure and other adverse cardiovascular reactions in Parkinson's disease.

Currie, G. L., Angel-Scott, H. N., Colvin, L. et al. (2019). Animal models of chemotherapy-induced peripheral neuropathy: A machine-assisted systematic review and meta-analysis. *PLoS Biol 17*, e3000243. doi:10.1371/journal.pbio.3000243

de Vries, R. B., Hooijmans, C. R., Tillema, A. et al. (2011). A search filter for increasing the retrieval of animal studies in Embase. *Lab Anim 45*, 268-270. doi:10.1258/la.2011.011056

de Vries, R. B., Hooijmans, C. R., Tillema, A. et al. (2014a). Updated version of the Embase search filter for animal studies. *Lab Anim 48*, 88. doi:10.1177/0023677213494374

de Vries, R. B. M., Wever, K. E., Avey, M. T. et al. (2014b). The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *ILAR J 55*, 427-437. doi:10.1093/ilar/ilu043

de Vries, R. B. M., Hooijmans, C. R., Langendam, M. W. et al. (2015). A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies. *Evid Based Preclin Med 2*, e00007. doi:10.1002/ebm2.7

Downs, S. H. and Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health 52*, 377-384. doi:10.1136/jech.52.6.377

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

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*, 1396E. doi:10.2903/sp.efsa.2018.EN-1396

Egger, M. and Smith, G. D. (2001). *Systematic Reviews in Health Care: Meta-Analysis in Context*. 2nd edition. London, UK: BMJ Books.

Fontelo, P. and Liu, F. (2018). A review of recent publication trends from top publishing countries. *Syst Rev 7*, 147. doi:10.1186/s13643-018-0819-1

Fox, D. (2010). *The Convergence of Science and Governance: Research, Health Policy, and American States*. Berkeley, CA, USA: University of California Press.

Gartlehner, G., Wagner, G., Lux, L. et al. (2019). Assessing the accuracy of machine-assisted abstract screening with distillerai: A user study. *Syst Rev 8*, 277. doi:10.1186/s13643-019-1221-3

Gates, A., Guitard, S., Pillay, J. et al. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Syst Rev 8*, 278-278. doi:10.1186/s13643-019-1222-2

Goodman, J. E., Petito Boyce, C., Sax, S. N. et al. (2015). Rethinking meta-analysis: Applications for air pollution data and beyond. *Risk Anal 35*, 1017-1039. doi:10.1111/risa.12405

Griesinger, C., Hoffmann, S., Kinsner, A. et al. (2009). Preface – Proceedings of the 1st international forum towards evidence-based toxicology. *Hum Exp Toxicol 28*, 83-86. doi:10.1177/0960327109105753

Guyatt, G., Oxman, A. D., Akl, E. A. et al. (2011). Grade guidelines: 1. Introduction-grade evidence profiles and summary of findings tables. *J Clin Epidemiol 64*, 383-394. doi:10.1016/j.jclinepi.2010.04.026

Guyatt, G. H., Oxman, A. D., Santesso, N. et al. (2013). Grade guidelines: 12. Preparing summary of findings tables – Binary outcomes. *J Clin Epidemiol 66*, 158-172. doi:10.1016/j.jclinepi.2012.01.012

Hersi, M., Quach, P., Wang, M. D. et al. (2017). Systematic reviews of factors associated with the onset and progression of neurological conditions in humans: A methodological overview. *Neurotoxicology 61*, 12-18. doi:10.1016/j.neuro.2016.06.017

Higgins, J. and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions.Version 5.1.0*. The Cochrane Collaboration and John Wiley & Sons Ltd.

Higgins, J. P., Altman, D. G., Gotzsche, P. C. et al. (2011). The Cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ 343*, d5928. doi:10.1136/bmj.d5928

Hoffmann, S. and Hartung, T. (2005). Diagnosis: Toxic! – Try-ing to apply approaches of clinical diagnostics and prevalence in toxicology considerations. *Toxicol Sci 85*, 422-428. doi:10.1093/toxsci/kfi099

Hoffmann, S., Griesinger, C., Coecke, S. et al. (2007). First international forum towards evidence based toxicology. *ALTEX 24*, 354-355. https://www.altex.org/index.php/altex/article/view/740/756

Hoffmann, S., Stephens, M. and Hartung, T. (2014). Evidence-based toxicology. In P. Wexler (ed.), *Encyclopedia of Toxicology* (565-567). Elsevier Inc. doi:10.1016/b978-0-12-386454-3.01060-5

Hoffmann, S., de Vries, R. B. M., Stephens, M. L. et al. (2017). A primer on systematic reviews in toxicology. *Arch Toxicol 91*, 2551-2575. doi:10.1007/s00204-017-1980-3

Hooijmans, C. R., Tillema, A., Leenaars, M. et al. (2010). Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. *Lab Anim 44*, 170-175. doi:10.1258/la.2010.009117

Hooijmans, C. R., IntHout, J., Ritskes-Hoitinga, M. et al. (2014a). Meta-analyses of animal studies: An introduction of a valuable instrument to further improve healthcare. *ILAR J 55*, 418-426. doi:10.1093/ilar/ilu042

Hooijmans, C. R., Rovers, M. M., de Vries, R. B. et al. (2014b). SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol 14*, 43. doi:10.1186/1471-2288-14-43

Hooijmans, C. R., de Vries, R. B. M., Ritskes-Hoitinga, M. et al. (2018). Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLoS One 13*, e0187271. doi:10.1371/journal.pone.0187271

IARC (2019). IARC Monographs on the Identification of Carcinogenic Hazards to Humans – Preamble.

Institute of Medicine (2011). *Clinical Practice Guidelines We Can Trust*. Washington, DC, USA: National Academies Press. doi:10.17226/13058

Jorgensen, L., Paludan-Muller, A. S., Laursen, D. R. et al. (2016). Evaluation of the Cochrane tool for assessing risk of bias in randomized clinical trials: Overview of published comments and analysis of user practice in Cochrane and non-Cochrane reviews. *Syst Rev 5*, 80. doi:10.1186/s13643-016-0259-8

Judson, R., Kavlock, R., Martin, M. et al. (2013). Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. *ALTEX 30*, 51-56. doi:10.14573/altex.2013.1.051

Kemppinen, A. K., Kaprio, J., Palotie, A. et al. (2011). Systematic review of genome-wide expression studies in multiple sclerosis. *BMJ Open 1*, e000053. doi:10.1136/bmjopen-2011-000053

Khan, K., Kunz, R., Kelijnen, J. et al. (2011). *Systematic Reviews to Support Evidence Based Medicine: How to Review and Apply Findings of Healthcare Research*. 2nd edition. London, UK: Hodder & Stoughton Ltd.

Krauth, D., Anglemyer, A., Philipps, R. et al. (2014). Nonindustry-sponsored preclinical studies on statins yield greater efficacy estimates than industry-sponsored studies: A meta-analysis. *PLoS Biol 12*, e1001770. doi:10.1371/journal.pbio.1001770

Krewski, D., Saunders-Hastings, P., Baan, R. et al. (2022). Workshop report: Development of an evidence-based risk assessment framework. *ALTEX*, in press.

Lam, J., Koustas, E., Sutton, P. et al. (2014). The navigation guide – Evidence-based medicine meets environmental health: Integration of animal and human evidence for PFOA effects on fetal growth. *Environ Health Perspect 122*, 1040-1051. doi:10.1289/ehp.1307923

Leenaars, M., Hooijmans, C. R., van Veggel, N. et al. (2012). A step-by-step guide to systematically identify all relevant animal studies. *Lab Anim 46*, 24-31. doi:10.1258/la.2011.011087

McCann, S. K., Cramond, F., Macleod, M. R. et al. (2016). Systematic review and meta-analysis of the efficacy of interleukin-1 receptor antagonist in animal models of stroke: An update. *Transl Stroke Res 7*, 395-406. doi:10.1007/s12975-016-0489-z

Moher, D., Liberati, A., Tetzlaff, J. et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med 6*, e1000097. doi:10.1371/journal.pmed.1000097

Moher, D., Shamseer, L., Clarke, M. et al. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev 4*, 1. doi:10.1186/2046-4053-4-1

Molander, L., Ågerstrand, M., Beronius, A. et al. (2015). Science in risk assessment and policy (SciRAP): An online resource for evaluating and reporting in vivo (eco)toxicity studies. *Hum Ecol Risk Assess 21*, 753-762. doi:10.1080/10807039.2014.928104

Morgan, R. L., Baack, B., Smith, B. D. et al. (2013). Eradication of hepatitis C virus infection and the development of hepatocellular carcinoma: A meta-analysis of observational studies. *Ann Intern Med 158*, 329-337. doi:10.7326/0003-4819-158-5-201303050-00005

Morgan, R. L., Thayer, K. A., Bero, L. et al. (2016). Grade: Assessing the quality of evidence in environmental and occupational health. *Environ Int 92-93*, 611-616. doi:10.1016/j.envint.2016.01.004

Morgan, R. L., Thayer, K. A., Santesso, N. et al. (2018a). Evaluation of the risk of bias in non-randomized studies of interventions (ROBINS-I) and the 'target experiment' concept in studies of exposures: Rationale and preliminary instrument development. *Environ Int 120*, 382-387. doi:10.1016/j.envint.2018.08.018

Morgan, R. L., Whaley, P., Thayer, K. A. et al. (2018b). Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environ Int 121*, 1027-1031. doi:10.1016/j.envint.2018.07.015

Morgan, R. L., Thayer, K. A., Santesso, N. et al. (2019). A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE. *Environ Int 122*, 168-184. doi:10.1016/j.envint.2018.11.004

Morton, S., Murad, M., O'Connor, E. et al. (2018). *Quantitative Synthesis – An Update. Methods Guide for Comparative Effectiveness Reviews.* Rockville, MD, USA: Agency for Healthcare Research and Quality. doi:10.23970/ahrqepcmethguide3

Mücke, M., Phillips, T., Radbruch, L. et al. (2018). Cannabis-based medicines for chronic neuropathic pain in adults. *Cochrane Database Syst Rev 3*, CD012182. doi:10.1002/14651858.CD012182.pub2

NTP – National Toxicology Program (2015). OHAT Risk of Bias Rating Tool for Human and Animal Studies. Office of Health Assessment and Translation. https://ntp.niehs.nih.gov/ntp/ohat/pubs/riskofbiastool_508.pdf

NTP (2016). Systematic Literature Review on the Effects of Fluoride on Learning and Memory in Animal Studies. NTP Research Report 1. https://ntp.niehs.nih.gov/ntp/results/pubs/rr/reports/rr01_508.pdf

NTP (2019). Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. Office of Health Assessment and Translation. https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbook march2019_508.pdf

Ouzzani, M., Hammady, H., Fedorowicz, Z. et al. (2016). Rayyan – A web and mobile app for systematic reviews. *Syst Rev 5*, 210. doi:10.1186/s13643-016-0384-4

Ritskes-Hoitinga, M., Leenaars, M., Avey, M. et al. (2014). Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. *Cochrane Database Syst Rev 3*, ED000078. doi:10.1002/14651858.ed000078

Ritskes-Hoitinga, M. and Wever, K. (2018). Improving the conduct, reporting, and appraisal of animal research. *BMJ 360*, j4935. doi:10.1136/bmj.j4935

Ritskes-Hoitinga, M. and van Luijk, J. (2019). How can systematic reviews teach us more about the implementation of the 3Rs and animal welfare? *Animals (Basel) 9*, doi:10.3390/ani9121163

Samuel, G. O., Hoffmann, S., Wright, R. A. et al. (2016). Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: A scoping review. *Environ Int 92-93*, 630-646. doi:10.1016/j.envint.2016.03.010

Schünemann, H., Brożek, J., Guyatt, G. et al. (2013a). *The Grade Handbook*. https://gdt.gradepro.org/app/handbook/handbook.html

Schünemann, H. J., Tugwell, P., Reeves, B. C. et al. (2013b). Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Res Synth Methods 4*, 49-62. doi:10.1002/jrsm.1078

Schünemann, H. J., Cuello, C., Akl, E. A. et al. (2019). Grade guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. *J Cin Epidemiol 111*, 105-114. doi:10.1016/j.jclinepi.2018.01.012

Sena, E. S., Briscoe, C. L., Howells, D. W. et al. (2010). Factors affecting the apparent efficacy and safety of tissue plasminogen activator in thrombotic occlusion models of stroke: Systematic review and meta-analysis. *J Cereb Blood Flow Metab 30*, 1905-1913. doi:10.1038/jcbfm.2010.116

Shea, B. J., Reeves, B. C., Wells, G. et al. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ 358*, j4008. doi:10.1136/bmj.j4008

Silbergeld, E. and Scherer, R. W. (2013). Evidence-based toxicology: Strait is the gate, but the road is worth taking. *ALTEX 30*, 67-73. doi:10.14573/altex.2013.1.067

Stephens, M. L., Andersen, M., Becker, R. A. et al. (2013). Evidence-based toxicology for the 21st century: Opportunities and challenges. *ALTEX 30*, 74-103. doi:10.14573/altex.2013.1.074

Stephens, M. L., Betts, K., Beck, N. B. et al. (2016). The emergence of systematic review in toxicology. *Toxicol Sci 152*, 10-16. doi:10.1093/toxsci/kfw059

Stephens, M. L., Akgun-Olmez, S. G., Hoffmann, S. et al. (2018). Adaptation of the systematic review framework to the assessment of toxicological test methods: Challenges and lessons learned with the zebrafish embryotoxicity test. *Toxicol Sci 171*, 56-68. doi:10.1093/toxsci/kfz128

Sterne, J., Hernán, M., Reeves, B. et al. (2016a). ROBINS-I: A tool for assessing risk of bias in non-randomized studies of interventions. *BMJ 355*, i4919. doi:10.1136/bmj.i4919

Sterne, J., Higgins, J., Elbers, R. et al. (2016b). Risk of Bias in Non-Randomized Studies of Interventions (ROBINS-I): Detailed Guidance, Updated 12 October, 2016. http://www.riskofbias.info

US EPA (2018a). Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the TSCA Program. US EPA. EPA-740-R1-8004. https://www.epa.gov/sites/production/files/2018-06/documents/epa_alt_strat_plan_6-20-18_clean_final.pdf

US EPA (2018b). Application of Systematic Review in TSCA Risk Evaluations. Document# 740-P1-8001. https://www.epa.gov/sites/production/files/2018-06/documents/final_application_of_sr_in_tsca_05-31-18.pdf

Van der Mierden, S., Tsaioun, K., Bleich, A. et al. (2019). Software tools for literature screening in systematic reviews in biomedical research. *ALTEX 36*, 508-517. doi:10.14573/altex.1902131

Vesterinen, H. M., Sena, E. S., Egan, K. J. et al. (2014). Meta-analysis of data from animal studies: A practical guide. *J Neurosci Methods 221*, 92-102. doi:10.1016/j.jneumeth.2013.09.010

Viswanathan, M., Patnode, C., Berkman, N. et al. (2017). Assessing the risk of bias in systematic reviews of health care interventions. Methods guide for comparative effectiveness reviews. (prepared by the scientific resource center under contract no. 290-2012-0004-c). AHRQ Publication No. 17(18)-ehc036-ef. doi:10.23970/ahrqepcmethguide2

Whaley, P., Aiassa, E., Beausoleil, C. et al. (in preparation). A code of practice for the conduct of systematic reviews in toxicology and environmental health research (COSTER).

Whiting, P., Savović, J., Higgins, J. P. et al. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol 69*, 225-234. doi:10.1016/j.jclinepi.2015.06.005

Wolffe, T., Vidler, J., Halsall, C. et al. (2020). A survey of systematic evidence mapping practice and the case for knowledge graphs in environmental health and toxicology. *Toxicol Sci 175*, 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. *Environ Health Perspect 122*, 1007-1014. doi:10.1289/ehp.1307175

Wright, R. W., Brand, R. A., Dunn, W. et al. (2007). How to write a systematic review. *Clin Orthop Relat Res 455*, 23-29. doi:10.1097/BLO.0b013e31802c9098

Yauw, S. T., Wever, K. E., Hoesseini, A. et al. (2015). Systematic review of experimental studies on intestinal anastomosis. *Br J Surg 102*, 726-734. doi:10.1002/bjs.9776

Conflict of interest

The authors declare that they have no conflict of interest.