The methodological rigour of systematic reviews in environmental health

J. M. L. Menona, F. Strujsa and P. Whaleyb,c

aSystematic Review Center for Laboratory Animal Experimentation, Department for Health Evidence, Radboud University Medical Center, Nijmegen, the Netherlands; bEvidence-Based Toxicology Collaboration, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; cLancaster Environment Centre, Lancaster University, Lancaster, UK

ABSTRACT
While systematic reviews (SRs) are often perceived as a “gold standard” for evidence synthesis in environmental health and toxicology, the methodological rigour with which they are currently being conducted is unclear. The objectives of this study are (1) to provide up-to-date information about the methodological rigour of environmental health SRs and (2) to test hypotheses that reference to a pre-published protocol, use of a reporting checklist, or being published in a journal with a higher impact factor, are associated with increased methodological rigour of a SR. A purposive sample of 75 contemporary SRs were assessed for how many of 11 recommended SR practices they implemented. Information including search strategies, study appraisal tools, and certainty assessment methods was extracted to contextualise the results. The included SRs implemented a median average of 6 out of 11 recommended practices. Use of a framework for assessing certainty in the evidence of a SR, reference to a pre-published protocol, and characterisation of research objectives as a complete Population- Exposure-Comparator-Outcome statement were the least common recommended practices. Reviews that referenced a pre-published protocol scored a mean average of 7.77 out of 10 against 5.39 for those that did not. Neither use of a reporting checklist nor journal impact factor was significantly associated with increased methodological rigour of a SR. Our study shows that environmental health SRs omit a range of methodological components that are important for rigour. Improving this situation will require more complex, comprehensive interventions than simple use of reporting standards.

Table of contents

1. Introduction .................................................................................. 168
2. Materials and methods ................................................................. 169
  2.1. Sampling method .................................................................... 169
    2.1.1. Search strategy .................................................................. 169
    2.1.2. Eligibility criteria and screening methodology .................. 169
  2.2. Development of the SR assessment instrument ...................... 170
    2.2.1. Sources for CREST_Basic ............................................... 170
    2.2.2. Questions and scoring method ......................................... 170
    2.2.3. Interpreting a CREST_Basic score ................................... 170
  2.3. SR assessment and data extraction ......................................... 170
    2.3.1. Blinding the assessment .................................................. 170
    2.3.2. Assessment and data extraction ....................................... 171
  2.4. Analysis .................................................................................. 171
    2.4.1. Summary of SR performance against CREST_Basic ........ 171
    2.4.2. Hypothesis tests ............................................................. 171
    2.4.3. Exploratory analyses ....................................................... 171
  2.5. Deviations from protocol ....................................................... 171
3. Results ....................................................................................... 171
  3.1. General characteristics of the included SRs ......................... 171
  3.2. General performance of included SRs against CREST_Basic ... 173
  3.3. Results for individual questions ............................................. 173
    3.3.1. Objectives ....................................................................... 173
    3.3.2. Protocols ......................................................................... 173
    3.3.3. Databases ........................................................................ 173
    3.3.4. Search strategies .......................................................... 174
    3.3.5. Eligibility criteria ......................................................... 174
    3.3.6. Study appraisal ............................................................ 174
    3.3.7. Certainty assessment .................................................. 175
    3.3.8. Use of reporting checklists .......................................... 175
  3.4. Results of hypothesis tests .................................................... 175
    3.4.1. Pre-published protocol and higher number of recommended SR practices .................................................. 176
    3.4.2. Reporting checklist and higher number of recommended SR practices .................................................. 176
    3.4.3. Higher impact factor of publishing journal and higher number of recommended SR practices .................................................. 176
  3.5. Exploratory analyses ........................................................... 176

CONTACT P. Whaley paul@whaleyresearch.uk

Supplemental data for this article is available online at https://doi.org/10.1080/10408444.2022.2082917

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
4. Discussion 176
4.1. Potential for unclear objectives and ambiguity about eligible evidence 177
4.2. Potential selectivity in retrieval of relevant evidence 178
4.3. Potential for mischaracterisation of validity of studies and overall certainty of evidence 178
4.4. The unclear role of protocol pre-publication in improving the rigour of SRs 179
4.5. The unclear role of reporting checklists in improving the rigour of SRs 180
4.6. Impact factor and publication quality 181
4.7. Strengths and limitations of our study 181
5. Conclusions 182
Acknowledgements 183
Declaration of interest 183
Supplemental material 183
Data availability statement 183
References 183

1. Introduction

Systematic review (SR) is a study design that aims to minimise bias and maximise transparency when answering research questions using existing evidence (Higgins, Thomas, et al. 2019; Whaley, Edwards, et al. 2020). To achieve its aims, SR methodology is rigorous, consisting of a standardised series of steps including: defining specific research objectives; prespecifying a detailed protocol; conducting comprehensive searches for relevant literature; screening and extracting data for analysis; assessing the potential for systematic error in the included studies; and using quantitative, qualitative, and/or narrative methods to synthesise the included evidence and determine the level of certainty with which the research question has been answered (IOM 2011; Hoffmann et al. 2017; Higgins, Lasserson, et al. 2019; Whaley, Aiassa, et al. 2020). When conducted well, SRs provide trustworthy summaries of evidence, helping identify knowledge gaps, providing insight into the validity of study methods in a given area of research, preventing unnecessary primary research being conducted when answers are already known, and cutting through uncertainty when studies are individually inconclusive (Chalmers et al. 2002). SRs also play an important role in improving study design and reporting, the translation of findings from experimental to target contexts, and informing decision-making (Ritskes-Hoitinga et al. 2014; Ritskes-Hoitinga and van Luijk 2019).

The value of systematic review in healthcare, as championed by organisations such as Cochrane, is well-established (Dickersin 2010). The potential value of SR methods for the toxicological and environmental health sciences (henceforth “environmental health”) was arguably first raised in the mid-2000s (Guzelian et al. 2005; Hoffmann and Hartung 2006). While there were precursors to SR in elements of the IARC Monographs (Samet et al. 2020) and disciplines that overlap environmental health, such as public and occupational health, few environmental health publications that explicitly describe themselves as SRs were published before 2005 (Figure 1). In 2010, the first agency-level guidance on using SR in regulatory risk assessment was issued (EFSA 2010) and publication rates of environmental health SRs were beginning to see a significant increase (Figure 1). In 2014, the first formal SR frameworks for chemical risk assessment and environmental health questions were published (Rooney et al. 2014; Woodruff and Sutton 2014). The increase in rate of publication of SRs since 2014 has been rapid, with regional, national, and international agencies and research groups developing guidance and implementing varying interpretations of systematic review methodology for environmental health questions (for example, see Vandenberg et al. 2016; NAS 2017; Schaefer and Myers 2017; Radke et al. 2020; Pega et al. 2021; Whaley, Aiassa, et al.

![Figure 1. Chart showing annual increase in number of publications on topics related to EH research with the term “systematic review” in the title, indexed in Web of Science. Updated from Whaley et al. (2020). Search: TITLE: (“systematic review”). Refined by: [excluding] DOCUMENT TYPES: (MEETING ABSTRACT) AND WEB OF SCIENCE CATEGORIES: (PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH OR TOXICOLOGY) AND [excluding] WEB OF SCIENCE CATEGORIES: (PHARMACOLOGY PHARMACY), Timespan: All years (1995–2020 shown). Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC. Date of search: 21 April 2021.](image)
2020; WHO 2021). In 2020, approximately 1750 environmental health SRs were published (see Figure 1).

While there has been a large increase in publications of SRs in environmental health, there is a lack of up-to-date empirical information about their methodological rigour. To our knowledge, three previous studies have directly investigated this. Sheehan and Lam (2015) analysed 48 environmental health SRs and meta-analyses published between 2001 and 2013. Sheehan et al. (2016) analysed 43 SRs and meta-analyses of ambient air pollution epidemiology published between 2009 and 2015. Sutton et al. (2021) analysed 13 SRs published between 2011 and 2017 in the course of a broader assessment of the extent to which reviews of three exposure-outcome pairs (air pollution and autism spectrum disorder, polybrominated diphenyl ethers and attention deficit hyperactivity disorder, and formaldehyde and asthma) can be characterised as systematic. All three studies found a high prevalence of a range of important shortcomings in adherence to accepted systematic review methodology. These include ambiguous formulation of review objectives, lack of pre-published protocols, and a lack of critical appraisal of the included evidence utilising valid instruments. The findings of these studies echo empirical evidence of widespread limitations in the conduct and reporting of SRs in healthcare (Ioannidis 2016; Page et al. 2016; Pussegoda et al. 2017; Gao et al. 2020) and socioeconomic research (Wang et al. 2021).

In this study, we aimed to provide up-to-date information about the methodological rigour of environmental health SRs. To do this, we conducted a survey of recently published, self-described systematic reviews that sought to quantify associations between environmental exposures and health outcomes. We assessed the SRs for methodological rigour, analysing them for frequency of implementation of methodological features that are generally considered to contribute to the transparency, utility, and validity of SRs. We tested three hypotheses: that a pre-published protocol, the use of a reporting checklist, or higher journal impact factor are associated with increased methodological rigour of a SR. We chose these hypotheses because reporting standards and pre-published protocols are intended to improve the rigour of SRs, and higher impact-factor journals may be assumed to be publishing more rigorous research. We also conducted some exploratory analyses of relationships in the data to help us interpret our results and develop hypotheses for potential future testing.

For ease of reading, SR-related tools, instruments, and frameworks are identified throughout the manuscript and supplemental materials as their abbreviation. A guide to abbreviations is provided in Appendix A. Supplemental materials for this manuscript are extensive and have been deposited in a structured Open Science Framework (OSF) archive (https://osf.io/j2d5v/). Supplemental material is hyperlinked throughout the manuscript, and a complete hyperlinked index of supplemental materials is presented in Appendix B.

2. Materials and methods

This study was conducted according to a pre-published protocol registered 19 November 2020 on the Open Science Framework (registration DOI: 10.17605/OSF.IO/MNPVS). The protocol was not externally peer-reviewed; however, registration allowed changes to planned analyses that were made after data had been collected to be transparently identified. The methodology consists of the following components: the selection of a purposive sample of environmental health SRs; the development of a tool, CREST_Basic, for rapid assessment of the methodological rigour of a SR; assessment of the SRs using CREST_Basic and the gathering of additional data to inform the discussion; testing of the primary hypotheses and conduct of the exploratory analyses; and analysis and interpretation of the results. Each is discussed below.

2.1. Sampling method

2.1.1. Search strategy

We designed a search strategy to retrieve environmental health SRs. The strategy had three components: a general environmental health concept; a systematic review concept; and a date filter to capture an 18-month period between 1 January 2019 and 30 June 2020. The search was performed in PubMed on 19 November 2020 and yielded 4882 results. The search strategy was validated by checking recall against a set of 20 SRs eligible for inclusion in our study, that were manually selected from a range of environmental health journals. The search retrieved 100% of the validation set. The search strategy and its results are available in the online supplemental materials SM01 (https://osf.io/85c6d/) and SM03 (https://osf.io/j59ku/) respectively.

2.1.2. Eligibility criteria and screening methodology

We employed a purposive SR sampling strategy aimed at covering a range of environmental health topics in a narrow, recent time window. We defined environmental health as “the investigation of associations between exposures to exogenous environmental challenges and health outcomes,” including toxicology and environmental epidemiology, as per our protocol (Menon et al. 2020). To be eligible for inclusion in the SR sample, documents had to fulfil the following criteria:

1. identify explicitly as a “systematic review” in their title;
2. assess the effect of a non-acute, non-communicable, environmental exposure on a health outcome;
3. include studies in people or mammalian models;
4. be available in HTML format;
5. be published between 1 January 2019 and 30 June 2020.

We excluded umbrella reviews and review protocols, SRs of the wrong population (e.g. plants), and SRs of the following exposures: pathogens; poisons; smoking and tobacco-related products, except for exposure to second-hand smoke (to prevent our sample being dominated by studies of direct tobacco exposure); social, psychological, or behavioural risk factors; housing conditions; SRs of exposure only (e.g. biomonitoring); and protocols for SRs. We did include diet and addiction as exposures, as these are considered by some researchers as being environmental exposures (Bazzano et al. ...
SRs had to be available in HTML format to allow copy-pasting of manuscripts into a uniform format stripped of visual cues, and allow easy removal of identifying information, as per our blinding strategy.

Title and abstract screening was performed independently by two reviewers (PW, JM) in Rayyan QCRI (https://rayyan.qcri.org/). SRs were sorted in reverse chronological order using Rayyan. Eligible studies were screened sequentially for inclusion by JM and PW until a sample of 75 SRs was reached. This was the maximum number of SRs we could evaluate in our budgeted time of 80 investigator hours. Discrepancies in decision were resolved by discussion; 902 SRs were screened to generate the sample. The final set of 75 SRs is available in SM06 (https://osf.io/mgxwj/) and the excluded SRs in SM13 (https://osf.io/y83tu/).

2.2. Development of the SR assessment instrument

In order to achieve our research objectives with the resources we had available, we determined that we required a SR assessment instrument that could be applied by two Masters-level investigators who are not experts in environmental health (JM and FS), would take less than 30 min to apply to a score that would represent the methodological rigour of a SR without falling foul of issues with scores and scales as presented by some appraisal instruments (see e.g. Greenland and O’Rourke 2001; Frampton et al. 2022).

Due to the high number of SR appraisal tools that have been published, we determined it would not be feasible to comprehensively review existing tools for applicability to our task. Instead, we decided to design a new tool, “CREST_Basic,” for our specific objectives and capacity. We summarise key information about CREST_Basic below. A complete description of the development of CREST_Basic, including pilot testing, are in the “Data Collection” section of our study protocol SM00 (https://osf.io/c8khr/).

2.2.1. Sources for CREST_Basic

We based CREST_Basic on three sources familiar to us and judged to be informative for our task: CREST_Triage, a rapid appraisal tool that supports editorial triage decisions for environmental health SRs and has been in use at the journal Environment International since 2018 (Whaley and Isalski 2019); AMSTAR-2, the updated version of an established SR appraisal tool that has been cited over 2000 times (Shea et al. 2017); and COSTER, the first set of good practice recommendations published specifically for environmental health SRs (Whaley, Aiassa, et al. 2020). We felt the sources and consensus processes behind COSTER and AMSTAR-2, combined with our experience of evaluating the rigour of hundreds of SRs from CREST_Triage (PW) and PROSPERO registrations (JM), would provide sufficient topic coverage and experience in structured appraisal of SRs from which to derive an instrument suited to our research objectives.

2.2.2. Questions and scoring method

CREST_Basic consists of 11 questions, each asking whether a methodological component associated with SR rigour is present. The questions either have yes/no or yes/no/unclear answers designed to be quickly and objectively answered by evaluators. Each question can be flagged by an assessor for “unusual practices” to facilitate follow-up analysis and interpretation. Answers of “yes” are awarded one point, and “no” and “unclear” are awarded zero points. A CREST_Basic score thereby represents how many of 11 methodological components are unambiguously present in a single SR, and for a set of SRs allows the prevalence of each of these 11 components to be calculated. Table 1 shows the questions, the reason for each question, criteria for the answers to each question, and data extracted for additional analysis.

The questions were presented to assessors as a Google Form, of which a PDF copy is available in SM02 (https://osf.io/y7w29/). The reason some questions have “unclear” as an answer is to allow evaluators to indicate uncertainty when an answer to a question depends on the testimony of the authors rather than being directly observable as a feature of a SR manuscript. For example, a missing element of a Population-Exposure-Comparator-Outcome (PECO) statement or presence of a complete search string, can be directly observed. In contrast, whether 100% of search results were screened in duplicate may be stated by SR authors but is not easily observed and may be reported ambiguously.

2.2.3. Interpreting a CREST_Basic score

CREST_Basic provides a measure of the methodological rigour of a SR, defined as the number of a given set of recommended methodological components that are present in a SR. Methodological rigour is a different concept to validity: the presence of a component associated with rigour does not guarantee the validity of a SR, as the component can be implemented to varying degrees of validity; nor does the absence of a component invalidate a SR. However, the absence of components associated with methodological rigour does indicate a potential threat to the validity or utility of a SR. For example, not searching in at least two databases increases risk of selection bias in a SR, and not providing a complete electronic search string means a database search strategy cannot be validated.

2.3. SR assessment and data extraction

2.3.1. Blinding the assessment

A blinding strategy was applied to minimise the potential for knowledge of e.g. publishing journal or author team to influence the way in which CREST_Basic was applied by the investigators. SRs were retrieved by PW and anonymised by copy-pasting article text from the HTML web version of a manuscript into a Microsoft Word document, clearing all formatting to a default of 11 pt Arial text, and removing identifiers for authors (names, affiliations, and CReDiT statement) and publication (e.g. journal title, article title, document formatting, links to article online). Any supplemental material needed for answering CREST_Basic questions was
screen-grabbed and pasted at the end of the word document. Each Word document was allocated a random number generated by Random.org (https://www.random.org/) as a unique identifier. Investigators were instructed not to conduct internet searches on text strings in the Word documents but to request additional anonymised materials from PW. An example of a blinded document is shown in SM08 (https://osf.io/62tpq/), and the text selection from HTML in SM09 (https://osf.io/e8s7c/). Due to probable prior knowledge of included manuscripts and potential conflicts of interest arising from their position as a specialist SR editor, PW did not conduct data extraction, and did not view extracted data until JM and FS had completed disagreement resolution. None of the investigators had published a SR eligible for inclusion in this study, so measures to prevent the assessment of their own work were unnecessary.

2.3.2. Assessment and data extraction
The anonymised manuscripts were assessed independently by two investigators (FS, JM) using the Google Forms version of CREST_Basic. The investigators were trained on a set of three SRs that met the eligibility criteria for this study, except for being outside the target date range. To reduce the risk of learning effects influencing manuscript assessments, one investigator assessed SRs in ascending order of numerical identifier while the other worked in descending order. Discrepancies in assessments were resolved by discussion. No third-party arbitration was required. After the assessment stage was completed, one investigator (JM) extracted additional information including journal title, impact factor (taken from the journal website), topics of the SRs, included populations of the SRs, and the presence of a CRediT statement or equivalent. Raw data from the data extraction process is available in SM05 (https://osf.io/uv5dg/).

2.4. Analysis

2.4.1. Summary of SR performance against CREST_Basic
Included SRs were allocated 1 point per affirmative answer for each CREST_Basic question. To summarise the performance of each individual SR against CREST_Basic, the points were summed to give a score out of 11. To summarise the performance of all SRs against each CREST_Basic question, we counted the frequency of affirmative answers across the set of included SRs. Because these are counts of document features, the issues with using scores and scales that would apply if e.g. validity of the findings of individual SRs were represented by a score are avoided (Greenland and O’Rourke 2001).

2.4.2. Hypothesis tests
We tested for inequality in sample size and variance using the Shapiro–Wilk test (Shapiro and Wilk 1965; Field 2009), Anderson-Darling test (Anderson and Darling 1952), and probability plots. The results of these tests allowed us to assume normal distribution; we therefore used the Mann–Whitney U test (Mann and Whitney 1947; Field 2009), computed in XLSTAT for Microsoft Excel (Addinsoft 2021), to test for an association between CREST_Basic score and the presence of a pre-existing protocol and the use of a reporting checklist. The Student T-test was used as a secondary test (Field 2009). To test for an association between score and journal impact factor, we used Pearson’s correlation coefficient (Field 2009).

2.4.3. Exploratory analyses
To explore our data, we created 2 × 2 contingency tables to identify correlations between characteristics of the included studies (e.g. the presence of a protocol and the use of a reporting checklist). Correlations were evaluated using the chi-square test in IBM SPSS version 27 (IBM 2020). When the expected numbers for a group was less than 5, we used Fisher’s exact test.

2.5. Deviations from protocol
The protocol for this study is available from the original OSF registration (https://osf.io/mkg9f/) and duplicated as SM00 on the OSF archive for this project (https://osf.io/c8khr/). We did not conduct a search using the Web of Science platform (see SM00 lines 95–96) due to it returning an excessively high number of irrelevant articles that we did not have capacity to screen. In our analysis we pooled rather than kept separate the “testimony” and “direct” question types that we defined in the protocol (see SM00 lines 151–156 and lines 268–274). We decided this was an artificial and potentially confusing distinction that would be better handled through subgroup analysis; however, in the supplemental materials the distinction between the two types is preserved, via colour coding, to facilitate separate analysis if needed, and testimony questions are marked with an asterisk in Table 1.

3. Results

3.1. General characteristics of the included SRs
We included 75 SRs in our study (see SM06 for the full list, https://osf.io/mgxwj/). The 75 SRs were published in 50 different journals, with impact factors ranging from 0.66–12.38 (median 3.32). The most prevalent journals were Environmental Research (n = 10), the International Journal of Environmental Research and Public Health (n = 6), and Science of the Total Environment (n = 4). Sixty-seven SRs included exclusively human evidence, 3 included exclusively animal evidence, and 5 were of both human and animal evidence. Four SRs included in vitro studies. Sixty-four SRs were published within the target date range, 10 in 2019, and 1 in 2018 (see section 4.7 for a discussion of this discrepancy). One hundred percent of the SRs we screened were available as HTML. Thirty-five environmental exposures and 25 health outcomes were covered. The three most commonly studied
Table 1. Questions and explanatory notes for CREST_Basic.

| Question | Reason for question | Criteria for answers | Additional extracted data |
|----------|---------------------|----------------------|----------------------------|
| 1. PECO statement: Is there a complete PECO statement? | The Population-Exposure-Comparator-Outcome ("PECO") framework is fundamental to structuring the objectives of a SR of exposure-outcome relationships. Absence of any element of the PECO framework has a number of consequences, including unambiguous assessment of comprehensives of a SR and the applicability of its findings. The PECO statement should be explicitly stated and cannot be inferred, for example, from the eligibility criteria of the SR. | To answer "yes," all four PECO elements must be explicit and stated as objectives independently of the eligibility criteria. A complete, stand-alone PECO statement is not necessary for a "yes," so long as all PECO elements are explicit in the text. | The investigators extracted the full PECO statement as presented in the manuscript and noted any missing elements. |
| 2. Protocol: Is there reference to a pre-published protocol? | A pre-published SR specifies the methods according to which a SR will be conducted. Pre-publication should involve posting a version-controlled, read-only document to a third-party preprint repository. Pre-publication of protocol allows potential for bias via changes in methods and/or selective reporting based on knowledge of results to be audited. Self-publication of a protocol without third-party version control does not allow for the pre-publication status of a protocol to be fully audited (NASEM 2022). | To answer "yes," the SR must provide a unique identifier for a protocol on a third-party website that the authors cannot directly edit (e.g. PROSPERO or preprint repository). | The investigators extracted the location, registration number and/or DOI of the protocol. The investigators also noted if a protocol was referenced but no location was provided. |
| 3. Databases: Were at least two databases or platforms searched? | Two databases or research platforms is a minimum for a comprehensive search. This helps ensure that no evidence relevant to the SR is overlooked due to differences in literature indexed by the databases. | To answer "yes," at least two research databases or platforms must have been searched. | The investigators extracted the names of the databases or platforms. |
| 4. Search strings: Have the authors presented a complete search string for at least one database? | Providing an exact search string is a necessary component of a reproducible search strategy that can be validated for comprehensiveness. | To answer "yes," the exact search string used in one of the included databases must be presented. | The investigators extracted the full search string, either from the main text or supplementary material, and noted whether there were any readily observable errors in syntax or issues with conceptual coverage. |
| 5. Eligibility criteria: Are the eligibility criteria described in terms of Population, Exposure and Outcome? | Interpretation of the objectives into eligibility criteria, explicitly based on the PECO statement, allows the reader to judge the validity of the SR authors' decisions on including studies, given the review objectives. | To answer "yes," eligibility criteria for the P, E and O elements of the PECO statement must be explicitly described. | The investigators extracted the eligibility criteria as a paragraph or list, noting any missing elements. |
| 6. Duplicate screening*: Were all screening decisions made in duplicate? | Duplicate screening reduces the risk of systematic and random error in screening decisions. | To answer "yes," the SR must state that 100% of screening was conducted in duplicate. | The investigators extracted relevant text describing whether duplicate screening was conducted. |
| 7. Duplicate data extraction*: Was data extraction conducted in duplicate? | Duplicate data extraction reduces the risk of systematic and random error in the data extraction process. | To answer "yes," the SR must state that 100% of data extraction was conducted in duplicate. | The investigators extracted relevant text describing whether duplicate data extraction was conducted. |
| 8. Use of critical appraisal tool: Was study quality assessed using a critical appraisal tool? | Critical appraisal helps contextualise the findings of a SR with the limitations of the included studies. Critical appraisal should be conducted using an appraisal instrument or tool, so the same criteria are consistently applied to each category of study included in a SR. | To answer "yes," the SR must appraise 100% of included studies using a critical appraisal tool. | The investigators extracted the names of any tools used for critical appraisal. If appropriate, the investigators classified tools as "home-made" (if created from scratch) or "modified" (if changes were made to a tool beyond those specified in the instructions for using the tool). |

(continued)
exposures were air pollution, pesticides, and occupational risk factors. The three most commonly studied health outcomes were female reproductive outcomes (primarily infant birthweight), neurodevelopment, and cancer. The data tables and analyses for this study are in SM10, SM11, and SM12 of the project OSF archive (https://osf.io/9krhz/).

3.2. General performance of included SRs against CREST_Basic

The included SRs scored a mean of 5.97 points out of 11 in the CREST_Basic questionnaire. Out of the 75 included SRs, only one scored the maximum 11 points. Three SRs scored 10 points, and six SRs scored 9 points. The median and mode scores were 6. Eighty-seven percent (65 of 75) of the included SRs omitted three or more components associated with methodological rigour. Figure 2 shows the overall scores and their distribution across the sample.

The most commonly omitted SR components were a complete PECO statement (57 of 75 SRs), a link to a pre-published protocol (62 of 75 SRs), and the lack of use of a certainty assessment framework (65 of 75 SRs). The three questions against which the included SRs performed the best were the use of at least two databases for search (67 of 75 SRs), provision of a complete search string (57 of 75 SRs) and the use of a critical appraisal tool to assess individual included studies (57 of 75 SRs). Figure 3 shows the performance of the included SRs against each individual CREST_Basic question.

3.3. Results for individual questions

3.3.1. Objectives

Seventy-six percent (57 of 75) of SRs had incomplete PECO statements. Of those that were incomplete, 100% (57 of 57) omitted a comparator, and 65% (37 of 57) omitted the population of interest. The exposure of interest was omitted once, and the outcome of interest was omitted once. This is shown in Figure 4.

3.3.2. Protocols

Seventeen percent (13 of 75) of SRs provided the location of a pre-published protocol (62 of 75 SRs), and the lack of use of a certainty assessment framework (65 of 75 SRs). The three questions against which the included SRs performed the best were the use of at least two databases for search (67 of 75 SRs), provision of a complete search string (57 of 75 SRs) and the use of a critical appraisal tool to assess individual included studies (57 of 75 SRs). Figure 3 shows the performance of the included SRs against each individual CREST_Basic question.

3.3.3. Databases

Eighty-nine percent (67 of 75) of SRs searched at least two literature databases or platforms. A total of 40 different databases or platforms were used. The most commonly used was PubMed/Medline (n = 68), followed by Web of Science (n = 44) and Embase (n = 33). Chinese databases were used 13 times. Registries and specialist grey literature databases
were used 15 times. A full list of the databases and platforms and the number of times they were used by the included SRs is in the “Data Set and Charts” sheet of SM10 in the project OSF archive (https://osf.io/6emp8/).

3.3.4. Search strategies
Seventy-six percent (57 of 75) of SRs provided at least one complete electronic search string. However, 56% of the reported search strings (32 of 57) were flagged by the assessors as potentially problematic, with concerns including lack of use of MeSH terms for PubMed, incomplete coverage of topic concepts, and inappropriate database syntax. Thus, while search strings were generally provided, there are concerns about the comprehensiveness of the searches in at least 67% (50 of 75) of the included SRs (18 without a search string plus 32 with a questionable search strategy).

3.3.5. Eligibility criteria
Seventy-one percent (53 of 75) of SRs reported their eligibility criteria in terms of population, exposure and outcome. Fifteen SRs omitted population, 8 omitted exposure, and 8 omitted outcomes. PECO elements were more widely implemented for describing the eligibility criteria for a SR than for framing its objectives.

3.3.6. Study appraisal
Seventy-six percent (57 of 75) of SRs used an instrument to facilitate critical appraisal of the included studies. The most
prevalent instrument was the Newcastle-Ottawa Scale (n = 24) (Wells et al. 2009) followed by home-made or modified instruments (n = 8), and either ROBINS-I or ROBINS-E (n = 5) (Sterne et al. 2016). Otherwise, choice of instrument was heterogeneous, with 31 instruments used across the 75 included SRs (see Figure 5). There were occasional examples of clearly inappropriate study appraisal methods. These included use of the reporting checklist STROBE, even though it instructs authors not to use STROBE as an appraisal tool (Vandenbroucke et al. 2007), and use of the GRADE certainty assessment framework, which applies to bodies of evidence not individual studies (Guyatt et al. 2008).

3.3.7. Certainty assessment
Thirteen percent (10 of 75) of SRs applied a formal framework for characterising certainty in their results. The GRADE Framework was used in 7 SRs, the NTP OHAT Framework (based on GRADE) in 2 SRs (NTP-OHAT 2019), and a home-made approach based on the Bradford Hill considerations in 1 SR. No SRs directly applied the Bradford Hill criteria (Bradford Hill 1965).

3.3.8. Use of reporting checklists
Sixty-nine percent (52 of 75) of SRs claimed to use a reporting checklist. Of these, only 60% (31 of 52) actually provided a completed checklist. Five checklists were referenced, the most common being PRISMA (n = 45) (Moher et al. 2009). The NTP OHAT Framework was referenced once, even though it is not a reporting checklist (NTP-OHAT 2019).

3.4. Results of hypothesis tests
The tests for distribution of our data gave mixed results regarding normality (Shapiro–Wilk test W:0.970, p = .05; Anderson–Darling test A²:0.834, p = .030). Based on histogram, P–P and Q–Q plots (see SM11), we assumed that our overall scores were normally distributed.
3.4.1. Pre-published protocol and higher number of recommended SR practices

Thirteen SRs provided the location of a pre-published protocol. The mean CREST_Basic score for SRs with a pre-published protocol was 7.769 (SD 1.928) out of 10 against 5.387 (SD 1.166) without (shown in Figure 6). The median score was 8 for SRs with protocol against 6 without. Significance was confirmed with a two-tailed t-test ($p < 0.0001$) and Mann-Whitney test ($p < 0.0001$), although the standard deviations overlap. Therefore, we could reject the null hypothesis that providing a pre-published protocol is not associated with increased methodological rigour of a SR, as represented by a CREST_Basic score.

3.4.2. Reporting checklist and higher number of recommended SR practices

Fifty-two SRs referred to a reporting checklist. Reviews were set in two groups: studies which referred to a reporting checklist ($n = 52$) and ones that did not ($n = 23$). The mean CREST_Basic score for SRs that referred to a reporting checklist was 5.5 (SD 1.93) out of 10 against 4.78 (SD 2.43) ($p = 0.323$) for those that did not (shown in Figure 7). The median was 5 against 5. Overlapping standard deviations and a two-tailed t-test ($p = 0.18$) indicate the association is not statistically significant. Therefore, the null hypothesis that reference to a reporting checklist is not associated with increased methodological rigour of a SR, as represented by a CREST_Basic score, could not be rejected.

3.4.3. Higher impact factor of publishing journal and higher number of recommended SR practices

Reviews were ranked by impact factor and divided into three tertiles: low (IF $\leq 2.849$, $n = 27$); intermediate ($2.849 < $IF $\leq 5.248$, $n = 21$); and high ($5.248 < $IF, $n = 26$). One review (ID443725738) was removed from analysis as the publishing journal had no impact factor (Current Developments in Nutrition). The mean CREST_Basic score for low tertile against high tertile was 5.44 (SD 2.7) against 6.58 (SD 2.37) ($p = 0.085$), shown in Figure 8. The median score was 6 against 6.5. Overlapping standard deviations and a two-tailed t-test ($p = 0.065$) indicate the association is not statistically significant. Pearson’s correlation coefficient was also non-significant ($p = 0.185$). Therefore, the null hypothesis that higher journal impact factor is not associated with increased methodological rigour of a SR, as represented by a CREST_Basic score, could not be rejected.

3.5. Exploratory analyses

Exploratory analyses of answers to CREST_Basic questions, conducted using $2 \times 2$ contingency tables, found a number of recommended practices to be significantly associated. The three practices carrying the most associations were reference to a pre-published protocol (four other practices were associated), use of a critical appraisal instrument (four other practices were associated), and duplicate conduct of critical appraisal (three other practices were associated). All statistically significant associations are shown in Table 2. The contingency tables are in Supplementary Materials 12.

4. Discussion

Our work raises a number of thematic issues relating to the rigour of systematic reviews in environmental health. We discuss six of these, speculate as to their implications for
eventually improving the rigour of published SRs, and describe some of the strengths and limitations of our study.

4.1. Potential for unclear objectives and ambiguity about eligible evidence

The value of PECO statements in characterising objectives and supporting structured analysis of evidence in SRs was recently reinforced by the US National Academies of Sciences, Engineering, and Medicine (NASEM) reviews of the US EPA TSCA SR protocols (NASEM 2021) and US EPA IRIS Handbook (NASEM 2022), and the World Health Organisation (WHO) framework for use of SR in chemical risk assessment (WHO 2021). Of the three studies directly comparable to ours, Sheehan and Lam (2015) found that 54% of environmental health SRs included all four elements of a PECO

---

**Figure 7.** Comparison of mean CREST_Basic score for SRs that reference a reporting checklist (n = 52) against those that do not (n = 23). Error bars show 1 standard deviation.

**Figure 8.** Comparison of mean CREST_Basic scores for low (n = 27) vs. high (n = 26) journal impact-factor tertiles. Error bars show 1 standard deviation.
Table 2. Significant associations between recommended practices implemented by the included SRs.

| CREST_Basic question                                      | Number of significant associations | Associated recommended practices                                                                 |
|-----------------------------------------------------------|------------------------------------|---------------------------------------------------------------------------------------------------|
| (1) Complete PECO statement                                | 1                                  | • Duplicate data extraction                                                                     |
| (2) Presence of a pre-published protocol                   | 4                                  | • Use of a critical appraisal instrument                                                          |
|                                                           |                                    | • Duplicate screening                                                                             |
|                                                           |                                    | • Duplicate critical appraisal                                                                  |
|                                                           |                                    | • Reporting checklist                                                                            |
| (3) Use of at least two databases                          | 2                                  | • Use of a critical appraisal instrument                                                          |
|                                                           |                                    | • Duplicate screening                                                                            |
| (4) At least one complete search string                    | 0                                  | • None                                                                                           |
| (5) Eligibility criteria cover P, E, O                     | 1                                  | • Certainty assessment                                                                           |
| (6) Use of a critical appraisal instrument                 | 4                                  | • Presence of a pre-published protocol                                                            |
|                                                           |                                    | • Use of at least two databases                                                                  |
|                                                           |                                    | • Use of certainty assessment framework                                                           |
|                                                           |                                    | • Duplicate critical appraisal                                                                  |
|                                                           |                                    | • Eligibility criteria cover P, E, O                                                             |
| (7) Use of certainty assessment framework                  | 1                                  |                                                                                                  |
| (8) Screening in duplicate                                 | 1                                  | • Presence of a pre-published protocol                                                            |
| (9) Data extraction in duplicate                           | 1                                  | • Presence of a pre-published protocol                                                            |
| (10) Duplicate critical appraisal                          | 3                                  | • Presence of a pre-published protocol                                                            |
|                                                           |                                    | • Use of a critical appraisal instrument                                                          |
|                                                           |                                    | • Duplicate data extraction                                                                     |
| (11) Reference to a reporting checklist                    | 1                                  | • Duplicate data extraction                                                                     |

statement, while Sutton et al. (2021) found 22% of SRs had a complete PECO statement or equivalent. Sheehan et al. (2016) did not evaluate PECO statements. In our sample, the populations, exposures, comparators, and outcomes of interest that form the objectives of a SR were only reported in 24% of the included SRs. In the incomplete PECO statements, comparators were always overlooked, and populations were overlooked in 65% of the time. Population, exposure, and outcome characteristics were better reported in eligibility criteria, where populations were reported in 75% of the included SRs and exposures and outcomes were each reported 90% of the time. Given the relative frequency of providing PECO information, we believe that SR authors may often be conflating eligibility criteria with the objectives of a review.

PECO statements are a formalism that facilitates unambiguous characterisation of the knowledge goal of a SR, interpreting a research question into an exact statement of the variables that are to be investigated (Morgan et al. 2018). Eligibility criteria, while also characterised in terms of PECO elements, are importantly different: rather than characterising the question being asked by a SR, they characterise the evidence that is considered informative for answering it. As such, although eligibility criteria need to be consistent with the PECO, they should not necessarily match it. This is a common necessity in environmental health research. Evidence from human studies is often limited, and eligibility criteria in a SR therefore may often need to be extended to include mammalian and other study models as informative surrogates for the specific PECO elements that characterise the SR’s objectives (Whaley et al. 2022). Conflating eligibility criteria and research objective reduces the clarity with which SR objectives are communicated to the reader, and neutralise a potentially valuable interpretative framework that can facilitate the conduct of the SR as a whole.

4.2. Potential selectivity in retrieval of relevant evidence

That SR methods may reduce selectivity in use of evidence has been an important driver in their uptake in environmental health research and risk assessment (e.g. Radke et al. 2020; Pega et al. 2021). Agency-led assessments of health risks posed by environmental exposures have often been criticised for being insufficiently comprehensive in their coverage of relevant evidence (e.g. Buonsante et al. 2014; Robinson et al. 2020; NASEM 2021). Sheehan and Lam (2015) found that 48% of SRs “fully described the screening, text review, and selection processes”; Sheehan et al. (2016) found 51% of SRs provided transparent search methods and 63% clear study selection criteria and procedures; Sutton et al. (2021) found 85% of their sample provided a “satisfactory” search strategy and 61% a “satisfactory” selection process.

In our sample, most SRs searched at least two databases (only 8 of 75 did not), and only three SRs explicitly stated that they did not conduct screening in duplicate (although 28 of 75 did not report this clearly, and we did not validate claims of duplicate screening). Fifty-seven of 75 SRs provided their search strategy, enabling validation of their approach.

Superficially, this seems encouraging. However, if we count SRs that either searched only one database or platform, or did not present a full search electronic string, or were unclear about screening evidence in duplicate, or were flagged for potential issues in search strategy, then 76% (57 of 75) of the included SRs may be vulnerable to partial inclusion of evidence relevant to their objectives. This is without attempting to validate in detail the search strategies and screening approaches of the remaining 24%. We believe this raises serious questions about whether the majority of published environmental SRs have sufficiently comprehensive coverage of the evidence they should be including.

4.3. Potential for mischaracterisation of validity of studies and overall certainty of evidence

Critical appraisal of the included evidence is considered to be an essential component of SR, because it provides the information that allows the credibility of the findings of a SR to be characterised. It should occur at two levels: the individual...
study, and the body of evidence on which the findings of the SR are based (Higgins, Thomas, et al. 2019). The SRs in our sample were seeking to quantify associations between exposures and outcomes. Critical appraisal at the level of individual study would therefore be expected to at least address internal validity, i.e. the potential for the way in which a study is conducted and reported to introduce systematic error into its results (Frampton et al. 2022). Many SR frameworks recommend that appraisal of internal validity be conducted using an instrument that assesses risk of bias (Hoffmann et al. 2017). Appraisal at the level of body of evidence concerns features of the evidence base as a whole that may affect certainty in the findings of a SR. This would include the results of the internal validity assessment and other features of the evidence base such as unexplained heterogeneity, the precision of the pooled results, and the generalisability of the evidence to the target question, among others (Morgan et al. 2016; Whaley, Aliassa, et al. 2020).

Contemporary approaches to SR in agency-led assessments increasingly include some kind of study appraisal and certainty assessment method. The specific methods vary, with subtle but important differences for both processes between, for example, the US EPA TSCA SR protocol (NASEM 2021), US EPA IRIS Handbook (NASEM 2022), US NTP OHAT Handbook (Rooney et al. 2014), and the Navigation Guide (Woodruff and Sutton 2014). For individual study appraisal there is a much greater diversity of instruments, with 31 used in our sample alone. Agency approaches tend to converge on risk of bias assessment, though implemented in different ways (NASEM 2021, 2022; WHO 2021), and there is some principled opposition to assessment of risk of bias at all (e.g. Steenland et al. 2020; Sheehan et al. 2016, and Sutton et al. 2021)., found that 38%, 28%, and 38% of their samples of SRs conducted some form of individual study appraisal respectively. None of these studies assessed the use of a certainty appraisal framework in a way that is comparable to our approach.

In our sample, there was high prevalence (76%) of the use of a critical appraisal instrument to assess individual included studies. However, we noted that many of the instruments were of unclear validity for use in quantitative SRs of associations. For example, the Newcastle-Ottawa Scale was the most prevalent instrument (used by 24 SRs) but as an approach to assessing internal validity it presents several issues. These include: incomplete coverage of important factors affecting the internal validity of a study (e.g. selective reporting is not assessed); the treatment of each shortcoming as being of equal weight regardless of true impact on internal validity (all shortcomings are scored equally); and a lack of transparency (users do not report reasons for their judgements) (Frampton et al. 2022). Of the 23 tools cited (8 were homemade), the six that in our opinion show the most validity (the Cochrane Risk of Bias instrument; ROBINS; SYRCLE; OHAT; and the Navigation Guide) were used in only 11 of the included SRs.

Our survey also shows that, while using instruments for assessing individual studies is at least relatively common, the use of a framework for appraisal of the overall body of evidence is currently only employed in a small minority of environmental health SRs (13%, 10 of 75). It was surprising that the Bradford Hill criteria were only used once, and then in modified form, although it should be noted that the GRADE Framework, and the Navigation Guide and NTP OHAT Handbook that are adaptations thereof, are directly derived from the Bradford Hill criteria (Schunemann et al. 2011). The reasons for applying a framework to assessing the body of evidence are the same as those for using an instrument for appraising individual studies: it promotes transparency and consistency of judgements, and discourages ad-hoc interpretation of the evidence that may be vulnerable to inconsistency of judgement and confirmation bias (NHMRC 2019). The lack of prevalence of formal approaches to assessing certainty is therefore an important limitation in current practices in conducting environmental health SRs.

Overall, we believe our survey shows that systematic review teams are challenged in selecting appropriate study appraisal instruments and certainty assessment frameworks. It is not clear to us if this is a cause or an effect of a lack of consensus on best practices. Either way, guidance on how to choose and/or modify study appraisal instruments that are valid and useful for a given SR context would be valuable. Frampton et al. (2022) recently proposed the FEAT mnemonic for this purpose, breaking down the evaluation of study appraisal instruments into four criteria of Focus (target of appraisal), Extent (comprehensiveness of appraisal), Application (generation of valid appraisal results that can be integrated into review findings), and Transparency (clear documentation of judgements made during the appraisal process). Incorporation of this kind of overarching guidance into academic training and the SR planning process may support SR authors in selecting appraisal tools, assist peer-reviewers and editors involved in checking the validity of methods used by SR authors, and also help move towards an improved consensus view of which tools are best used for a given research context. Similar work could be done for certainty assessment. A recent protocol for a WHO SR of the effect of electromagnetic frequency radiation on cancer in animal studies was developed by a mix of authors favouring and sceptical of risk of bias assessment and GRADE-style approaches to certainty assessment (Mevissen et al. 2022). This may provide some indication of where future consensus may reside.

4.4. The unclear role of protocol pre-publication in improving the rigour of SRs

Pre-publication of a protocol for a SR has been a common recommendation in healthcare and social science for a long time, with protocol repositories such as PROSPERO established to provide a means for this to happen. Environmental health journals first began to accept SR protocols from around 2016 (Whaley and Halsall 2016). Agency-level assessments such as those by the European Food Safety Authority, US NTP OHAT, US EPA IRIS Program, US EPA TSCA, and WHO now pre-publish protocols for SRs. To facilitate this overall trend, PROSPERO has been extended to environmental health topics and SRs of animal studies, so long as they have a
direct health outcome (https://www.crd.york.ac.uk/prospero/#aboutpage). In absolute terms, however, protocol publication is still rare: in our sample, 17% of SRs linked to a pre-published protocol; Sheehan and Lam (2015) found 8% referred to any kind of protocol; Sheehan et al. (2016) did not look at this; and Sutton et al. (2021) rated 22% of their sample as “satisfactory” for use of protocol.

According to our analysis of $2 \times 2$ contingency tables, SRs that provided a link to a pre-published protocol performed statistically significantly better than those that did not in four areas: duplicate screening; duplicate appraisal; the use of a critical appraisal tool; and use of a reporting checklist. Our results also showed that only one SR that linked to a protocol failed to provide a complete search string, and only one SR that linked to a protocol failed to provide complete eligibility criteria (see Figure 9); however, this potential association between reference to protocol and complete search strategy or eligibility criteria did not reach statistical significance, potentially due to sample size.

Issues with incomplete PECO statements and a lack of certainty appraisal framework were still prevalent among the protocol-linked SRs. Of the 13 SRs that referenced a protocol and provided a complete search strategy, 6 were flagged for potential issues with the validity of their search strategy. While all the SRs that referenced a protocol used a critical appraisal tool to assess individual studies, 7 of the 13 used a tool of unclear validity for assessing risk of bias in a SR, such as the Newcastle-Ottawa Scale. Therefore, while a link to a protocol is associated with improved methodological rigour, there are still important shortcomings in the absolute rigour of protocol-based SRs, and there are questions about the validity of the methods that are being used.

Because our study is cross-sectional in design, we cannot determine the direction of cause between link to a protocol and CREST_Basic score, i.e. whether registration of protocol causes authors to conduct more components of a SR, or whether a more comprehensive awareness of the multiple components of a SR leads to recognition of the importance of registering review protocols. While we would speculate it is the latter, we also acknowledge the possibility that the act of deliberate planning implied by protocol registration may increase the number of CREST_Basic domains that are covered by a SR team. This would be especially likely if the protocol registration system prompts for text for each major step of the SR (as PROSPERO does, for example). We note, however, that registration alone does not offer an obvious, active mechanism for improvement, as improvement in planned methods would be due to internal actions on the part of the project team. This is in contrast to journal peer-review, where there is an external process for identifying issues in a manuscript and incentive for addressing them. We therefore speculate that external peer-review may provide a mechanism by which protocol development could cause improvement in SR rigour, but our sample did not include enough SRs that used this approach to test for an association.

Given the potential for improvement in the methods of planned SRs, and the potential costs (to both researchers and funders) of implementing a general policy of peer-reviewing SR protocols, it may be justified to conduct a randomised controlled trial of the efficacy of peer-review of protocols in improving SR methods. Alternative observational study designs, that may nonetheless provide evidence of sufficient certainty to support a policy of peer-reviewing protocols, could be before-after comparison of the quality of submitted protocols with accepted protocols, or a cross-sectional comparison of submitted or accepted protocols with protocols that have only been registered.

4.5. The unclear role of reporting checklists in improving the rigour of SRs

SRs which claimed to have followed reporting checklists were slightly, but not significantly, associated with better
CREST_Basic scores. Among the SRs that did refer to a reporting checklist, a majority did not provide a complete PECO statement, a link to a protocol, or use a certainty framework. A large minority did not report duplicate screening, duplicate data extraction, or duplicate critical appraisal (see Figure 10), even though these are key elements of reporting checklists such as PRISMA. In general, while authors were claiming to have used reporting checklists, they were selective in their compliance with them. This is consistent with the findings of Page et al. (2016) in healthcare research. Sheehan and Lam (2015), Sheehan et al. (2016), and Sutton et al. (2021) also indicate that SRs often cite reporting standards while exhibiting poor compliance with them, though none of these studies assessed SRs citing reporting standards as a distinct subgroup.

There is evidence from some reviews that SRs published in journals that recommend or require adherence to a reporting checklist are better reported, though the effect size may be small (Agha et al. 2016, Cullis et al. 2017, Nawi et al. 2019). In other areas, the effectiveness of reporting checklists for improving research has been shown to be limited. For example, in pre-clinical research, Hair et al. showed that asking for a completed checklist of the ARRIVE guidelines did not alter the reporting quality of studies (Hair et al. 2019). Researchers have highlighted a number of challenges in adhering to reporting checklists, finding that lack of awareness of checklists, checklist complexity, and time or word limitations, can reduce compliance (Burford et al. 2013). Overall, it remains unclear how effective reporting checklists are for improving the quality of conduct or reporting of SRs in environmental health. If they are helpful, we speculate that it would be when they are included in a broader strategy for improving publishing standards for SRs rather than being applied in isolation, through e.g. editorial enforcement of standards and rigorous gatekeeping. This is discussed in more detail in Whaley et al. (2021), in the context of interventions that journal editors might make to improve SR publishing standards.

4.6. Impact factor and publication quality

Higher impact factor of a journal is in some contexts taken as a measure of the higher quality of the research it publishes. While the link between the two is tenuous (for discussion, see e.g. Haustein and Larivière 2015; McKiernan et al. 2019), it could be speculated that higher impact journals may be more likely to impose quality control processes that increase the rigour of the SRs they are publishing. We followed other studies of comparable type to ours in testing for an association between impact factor and some measure of research quality (see e.g. Fleming et al. 2014); in our case, the number of recommended methodological practices being implemented in published SRs. In general, environmental health journals are publishing SRs that omit a number of important methodological components, and higher impact journals seem to perform about as well as lower impact journals in this regard. The use of journal impact factor as an indicator of the rigour of a systematic review would at this time seem unsupported.

4.7. Strengths and limitations of our study

A limitation of our study lies in its sample of SRs. We aimed for a comprehensive sample in a time window of several months, counting back from June 2020. However, because of apparent inconsistencies in recording of publication dates in PubMed and RIS exports, and subsequent handling in RAYYAN, we have a non-random sample of SRs spread across a wider time window. The reason for this happening may be due to ambiguity around online vs. print publication, which
is why we did not exclude one SR (ID 551204262) that appears to fall outside our date range (this SR was e-pub-
lished and indexed in PubMed in 2018 but only listed as published in the journal in 2020). Due to our blinding stra-
tegy, the issue only became apparent after we had completed the additional data extraction step of the study, and we did not have resources to resample and reconduct data extrac-
tion. If time of publication is associated with methodological rigour or affects prevalence of topics covered within a given time period, then random sampling from a specified time period would improve the generalisability of our results, though would not avoid the issue of some SRs being outside the target time window. Searching PubMed only may also present a generalisability issue, if there are significant differ-
ces between SRs that are indexed in PubMed and SRs that are indexed in other databases or literature platforms. Our sample did consist mainly of SRs of human epidemiology studies; this may reflect their predominance in the literature, but we do not know the ratio of epidemiology to non-epi-
demiology SRs in environmental health. We could arguably have excluded SRs of drug exposure, but only two papers on this topic were included. We did not provide data on SRs that did not declare themselves as such. Since we excluded SRs of direct tobacco exposure, our findings do not general-
ise to that topic.

Nonetheless, we believe we have presented the largest evaluation of environmental health SRs to date, with more generalisability than the topic-focused studies of Sutton et al. (2021) and Sheehan et al. (2016), and presenting a more con-
temporary picture of SR practices than either of these two studies or Sheehan and Lam (2015). Our assessment directly covers a set of well-established characteristics that a SR ought to have, updating the Literature Review Appraisal Tool (LRAT) applied by Sutton et al. (2021) based on contemporary guidance for conduct of environmental health SRs that was unavailable to Sheehan and Lam (2015) and the original LRAT authors (Whaley and Halsall 2014).

Our methodology has prioritised replicability over poten-
tially subjective or controversial judgements about the valid-
ity of the methods of SRs. While this puts limits on detailed insight into the impact of methodological rigour on the validity of the findings of included SRs (e.g. we can only comment on choice of study appraisal instruments, not the validity of their application), CREST_Basic is quick to apply (10 min per evaluation) and requires only Masters-level expertise and approximately 3 h of training. Because the questions of CREST_Basic have been designed to be unambiguous and have objective answers, they should be applicable by other researchers to generate results directly comparable with our study here (we observed 107 discrepancies or 13% raw dis-
agreement between reviewers in our data set, that primarily related to issues such as confusing GRADE with risk of bias assessment, that are easily resolved via consensus between evaluators—see SM07 https://osf.io/dz4v8/). This may help with reproducing our findings in a more representative sample of SRs, a particular subset under-represented in our study (e.g. SRs of animal or in vitro studies), assessing changes in SR quality over time, and extending in other ways our initial dataset of 75 SRs. To facilitate this process, we have made our entire project available as a template on the Open Science Framework (https://osf.io/j2d5v/).

5. Conclusions

Our study shows that a large number of SRs on environmen-
tal health topics are being published in spite of important shortcomings in methodological rigour. Reporting checklists as currently applied seem ineffective for addressing this, and while pre-publication of protocol has some effect on rigour, it is arguably not effective enough. Higher journal impact fac-
tor does not guarantee higher methodological rigour of SRs. Our analysis also indicates the following issues are prevalent in environmental health SRs:

- PECO statements are under-utilised, with eligibility crite-
ria often being conflated with SR objectives.
- There is significant potential for selective inclusion of evi-
dence in SRs via limitations in search strategy and litera-
ture screening methods.
- There is frequent use of instruments that are inappropri-
ate for critical appraisal of individual studies included in a SR, although there is an encouraging general recogni-
tion of the importance of critical appraisal as a step of the SR process.
- There is a lack of recognition of the importance of sys-
tematic assessment of the certainty of the evidence as part of the findings of a SR.
- Pre-published protocols are under-utilised, but to be fully effective for improving the rigour of SRs they likely need to be accompanied by additional interventions, such as peer-review of protocols and editorial enforcement of SR standards. The same appears to be true for reporting checklists.

Improving the rigour of SRs in environmental health is likely to be a serious challenge, requiring a coordinated, multi-intervention response. We would be cautious about stepping beyond what our immediate study has shown in recommending a response. However, general education of authors, editors, and peer-reviewers would seem sensible, as manuscripts with important limitations in rigour are routinely making it through journal review processes. Instruments that can support peer-review (such as AMSTAR-2) and editorial gatekeeping (such as CREST_Triage) are available but not widely used in SR manuscript handling workflows. Enforcement seems key: PRISMA reports are regularly used, but it seems they are rarely checked. Peer-review of protocols may be especially effective for increasing methodological rig-
our of SRs and could be tested. Overall, the range of poten-
tial interventions is wide; we have considered these in previous work (Whaley et al. 2021). If it could be established, a Cochrane-like organisation for environmental health that
can research, develop, and set SR standards, while building consensus and educating people, might be able to make an important difference. These are exciting times for SR, but steps need to be taken to ensure their rigour if they are to serve as the gold standard in evidence synthesis.

Acknowledgements

We would like to thank Dr. Sebastian Hoffman of the Evidence-based Toxicology Collaboration for reviewing our study protocol. JM participated in writing the protocol, piloted CREST_Basic, designed the search strategy, performed search and screening, data extraction and post-hoc data extraction, data analysis, and wrote the first draft of the manuscript. FS participated in writing the protocol, piloted CREST_Basic, performed data extraction, and reviewed versions of the manuscript. PW supervised the project, wrote the protocol, designed the search strategy, performed the screening, anonymised the reviews, handled the uploading of documents to the Open Science Framework, and reviewed and revised the manuscript. Dr. Katya Tsaioi (KT) and SH were involved in planning the study. SH reviewed the study protocol and KT reviewed the final manuscript and approved it for submission. Lancaster University provided library access and software for PW. The authors are extremely grateful to five anonymous peer-reviewers for extensive, thoughtful comments that resulted in significant improvements to the framing and contextualisation of the manuscript. The authors would also thank Dr. Roger O. McClellan and Dr. Jesse Denson Hesch for their editing of the manuscript.

Declaration of interest

This study is an undertaking of the Evidence-based Toxicology Collaboration (EBTC) at Johns Hopkins Bloomberg School of Public Health, https://www.ebtox.org/. EBTC is a research organisation that works in the toxicological and environmental health sciences on raising research standards, promoting the use of systematic methods in synthesising evidence, improving access to research, and advocating evidence-based decision-making. EBTC receives the majority of its income from private donations to Johns Hopkins University.

This study was self-funded from EBTC’s discretionary funds. The study was initiated by staff in strategic support of the objectives of EBTC. It was conducted to gain insight into current methods being used for conducting systematic reviews (SRs) in environmental health, and will inform EBTC’s education and advocacy work around raising SR standards. PW (Research Fellow at EBTC) led the study. Radboud University Medical Centre (JM and FS) was engaged by EBTC on a consultancy contract to conduct the research. Dr Katya Tsaioi (Executive Director of EBTC) and Dr Sebastian Hoffmann (Research Fellow at EBTC) were involved in planning the study. SH reviewed the study protocol and KT reviewed and approved the submission of the final manuscript.

JM declares they are an editor of submissions to the animal studies journal: a before and after study. Int J Surg. 30:169–212.

Bazzano LA, Pogribna U, Whelton PK. 2008. Hypertension. In: Heggenhougen HK (kris), editor. International Encyclopedia of public health. Oxford (UK): Academic Press; p. 501–513.

Bradford Hill A. 1965. The environment and disease: association or causation? Proc R Soc Med. 58:295–300.

Buonsante VA, Muilerman H, Santos T, Robinson C, Tweedale AC. 2014. Risk assessment’s insensitive toxicity testing may cause it to fail. Environ Res. 135:139–147.

Burford BJ, Welch V, Waters E, Tugwell P, Moher D, O’Neill J, Koehlmoos T, Petticrew M. 2013. Testing the PRISMA-equity 2012 reporting guideline: the perspectives of systematic review authors. PLoS One. 8(10):e75122–e75122.

Chalmers I, Hedges LV, Cooper H. 2002. A brief history of research synthesis. Eval Health Prof. 25(1):12–37.

Cullis PS, Gudlaugsdottir K, Andrews J. 2017. A systematic review of the quality of conduct and reporting of systematic reviews and meta-analyses in paediatric surgery. PLoS One. 12(4):e0175213–e0175213.

Dickersin K. 2010. Health-care policy. To reform U.S. health care, start with systematic reviews. Science. 329(5991):516–517.

Downes MJ, Brennan ML, Williams HC, Dean RS. 2016. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open. 6(12):e011458.

[EFSAs] European Food Safety Authority. 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSAs J. 8(6):1637.

Field A. 2009. Discovering statistics using SPSS. 3rd ed. London (UK): Sage Publications Ltd.

Fleming PS, Koletsi D, Seehra J, Pandis N. 2014. Systematic reviews published in higher impact clinical journals were of higher quality. J Clin Epidemiol. 67(7):754–759.

Frampton G, Whaley P, Bennett M, Bilotta G, Dorne JCM, Eales J, James K, Kohl C, Lind M, Livoreil B, et al. 2022. Principles and framework for assessing the risk of bias for studies included in comparative quantitative environmental systematic reviews. Environ Evid. 11(1).

Gao Y, Cai Y, Yang K, Liu M, Shi S, Chen J, Sun Y, Song F, Zhang J, Tian J. 2020. Methodological and reporting quality in non-Cochrane systematic review updates could be improved: a comparative study. J Clin Epidemiol. 119:36–46.

Greenland S, O’Rourke K. 2001. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics. 2(4):463–471.

Guyatt GH, Oxman GE, Vist R, Kunz Y, Falck-Ytter P, Alonso-Coello HJ, Schünemann AD. 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 336(7650):924–926.

Guzelian PS, Victoroff MS, Halmes NC, James RC, Guzelian CP. 2005. Evidence-based toxicology: a comprehensive framework for causation. Hum Exp Toxicol. 24(4):161–201.

Hair K, Maceood MR, Sena ES. 2019. A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (ICARUs). Res Integr Peer Rev. 4:12.

Hausteins S, Lavriére V. 2015. The use of bibliometrics for assessing research: possibilities, limitations and adverse effects. In: Welpe IM,
Wollersheim J, Ringelhan S, Osterloh M, editors. Incentives and performance: governance of research organizations. Cham (Switzerland): Springer International Publishing; p. 121–139.

Higgins JPT, Lasserson T, Chandler J, Tovey D, Thomas J, Flemyng E, Churchill R. 2019. Methodological expectations of Cochrane intervention reviews. (MECIR). [place unknown]: Cochrane.

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. 2019. Cochrane handbook for systematic reviews of interventions version. 2nd ed. Chichester (UK): John Wiley & Sons.

Hoffmann S, de Vries RBM, Stephens ML, Beck NB, Dirven HAAM, Fowle Higgins JPT, Lasserson T, Chandler J, Tovey D, Thomas J, Flemyng E, Hoffmann S, Hartung T. 2006. Toward an evidence-based toxicology.

Ioannidis JP. 2016. The mass production of redundant, misleading, and IBM. 2020. SPSS statistics for windows, version 27.0. Armonk (NY): IBM Corp.

[IOM] Institute of Medicine. 2011. Finding what works in health care: standards for systematic reviews. Washington (DC): National Academies Press.

Mann HB, Whitney DR. 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist. 18(1):50–60.

McKieran EC, Schimanski LA, Muñoz Nieves C, Matthias L, Niles MT, Alperin JP. 2019. Use of the journal impact factor in academic review, promotion, and tenure evaluations. Elife. 8:e47338.

Menon J, Struijs F, Whaley P. 2020. A survey of the basic scientific quality of systematic reviews in environmental health. Open Science Framework Registration. https://doi.org/10.17605/OSF.IO/MNPV5.

Meissner M, Ward JM, Kopp-Schneider A, McNamee JP, Wood AW, Rivero TM, Thayer K, Straif K. 2022. Effects of radiofrequency electromagnetic fields (RF EMF) on cancer in laboratory animal studies. Environ Int. 161:107106.

Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 62(10):1006–1012.

Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Morgan RL, Whaley P, Thayer KA, Schubauer-Berigan MK, Verbeek JH, Holloway AC, Santesso N, Schünemann HJ. 2016. GRADE: Assessing the quality of evidence in environmental and occupational health. Environ Int. 92–93:611–616.

Morgan RL, Whaley P, Thayer KA, Schünemann HJ. 2018. Identifying the PECO: a framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. Environ Int. 121(7):1027–1025.

[NASEM] National Academies of Sciences, Engineering, and Medicine. 2017. Application of systematic review methods in an overall strategy for evaluating low-dose toxicity from endocrine active chemicals. [place unknown]: National Academies Press.

[NASEM] National Academies of Sciences, Engineering, and Medicine. 2021. The Use of systematic review in EPA’s toxic substances control act risk evaluations. [place unknown]: National Academies Press.

[NASEM] National Academies of Sciences, Engineering, and Medicine. 2022. Review of U. S. EPA’s ORD staff handbook for developing IRIS assessments: 2020 version. [place unknown]: National Academies Press.

Nawijn F, Ham WHW, Houwerm HW, Groenwold RHH, Hietbrink F, Smeeking DPJ. 2019. Quality of reporting of systematic reviews and meta-analyses in emergency medicine based on the PRISMA statement. BMC Emerg Med. 19(1):19.

NHMRC. 2019. Guidelines for guidelines: assessing risk of bias. [accessed Nov 2021]. https://nhmrc.gov.au/guidelinesforguidelines/develop/assessing-risk-bias.

NTP-OHAT. 2019. Handbook for conducting systematic reviews for health effects evaluations. Research Triangle Park, NC: U.S. Department of Health and Human Services. [accessed 2022 Jan 10]. https://ntp.niehs.nih.gov/whatwedo/assessment/noncancer/handbook/index.html.

Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, Catalá-López F, Li L, Reid EK, Sarkis-Onofre R, et al. 2016. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med. 13(5):e1002028.

Pega F, Momen NC, Ujita Y, Driscoll T, Whaley P. 2021. Systematic reviews and meta-analyses for the WHO/ILO joint estimates of the work-related burden of disease and injury. Environ Int. 155:106605.

Pussegoda K, Turner L, Garritty C, Mayhew A, Skidmore B, Stevens A, Bouton I, Sarkis-Onofre R, Bjerve LM, Hrubíjartsson A, et al. 2017. Identifying approaches for assessing methodological and reporting quality of systematic reviews: a descriptive study. Syst Rev. 6(1):1–12.

Radke EG, Yost EE, Roth N, Sathyanarayanaya S, Whaley P. 2020. Application of US EPA IRIS systematic review methods to the health effects of phthalates: lessons learned and path forward. Environ Int. 145:105820.

Ritskes-Hoitinga M, Leenaars M, Avey M, Rovers M, Scholten R. 2014. Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. Cochrane Database Syst Rev. (3):ED000078.

Ritskes-Hoitinga M, van Luijk J. 2019. How can systematic reviews teach us more about the implementation of the 3Rs and animal welfare? Animals. 9(12):1163.

Robinson C, Portier CJ, Čavoski A, Mesnage R, Roger A, Claudiing P, Whaley P, Muilerman H, Lyssimachou A. 2020. Achieving a high level of protection from pesticides in Europe: problems with the current risk assessment procedure and solutions. Eur J Risk Regul. 11(3):450–480.

Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. 2014. Systematic review and evidence integration for literature-based environmental health science assessments. Environ Health Perspect.122(7):711–718.

Samet JM, Chiu WA, Cogliano V, Jinot J, Kriebel D, Lunn RM, Beland FA, Bero L, Browne P, Fritschl L, et al. 2020. Achieving a high level of protection from pesticides in Europe: problems with the current risk assessment procedure and solutions. Eur J Risk Regul. 11(3):450–480.

Schaefer HR, Myers JL. 2017. Guidelines for performing systematic reviews in the development of toxicity factors. Regul Toxicol Pharmacol. 91:124–141.

Schunemann H, Hill S, Guyatt G, Akp EA, Ahmed F. 2011. The GRADE approach and Bradford Hill’s criteria for causation. J Epidemiology Community Health. 65(5):392–395.

Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples).1. Biometrika. 52(3-4):591–611.

Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E, 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.

Sheehan MC, Lam J. 2015. Use of systematic review and meta-analysis in environmental health epidemiology: a systematic review and comparison with guidelines. Curr Environ Health Rep. 2(3):272–283.

Sheehan MC, Lam J, Navas-Acien A, Chang HH. 2016. Ambient air pollution epidemiology systematic review and meta-analysis: a review of reporting and methods practice. Environ Int. 92-93:647–656.

Sheenland K, Schubauer-Berigan MK, Vermeulen R, Lunn RM, Straif K, Zahm S, Stewart P, Arroyave WD, Mehta SS, Pearce N. 2020. Risk of bias assessments and evidence syntheses for observational epidemiologic studies of environmental and occupational exposures: strengths and limitations. Environ Health Perspect. 128(9):95002.

Sterne JAC, Hernán MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, et al. 2016. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 355:i4919.

Sutton P, Chartres N, Rayasam SDG, Daniels N, Lam J, Maghrbi E, Woodruff TJ. 2021. Reviews in environmental health: how systematic are they? Environ Int. 152:106473.

Vandenberg LN, Ågerstrand M, Beronius A, Beausoleil C, Bergman Å, Bero LA, Bornemark C-G, Boyer CS, Cooper GS, Cotgreave I, et al. 2016. A proposed framework for the systematic review and integrated
assessment (SYRINA) of endocrine disrupting chemicals. Environ Health. 15(1):74.
Vandenbroucke JP, von Elm DG, Altman PC, Gotzsche CD, Mulrow SJ, Pocock C, Poole JJ, Schlesseleman M, Egger E, STROBE Initiative. 2007. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLoS Med. 4(10):e297.
Wang X, Welch V, Li M, Yao L, Littell J, Li H, Yang N, Wang J, Shamseer L, Chen Y, et al. 2021. The methodological and reporting characteristics of Campbell reviews: a systematic review. Campbell Syst Rev. 17(1): e1134.
Wells GA, Shea B, O’Connell D, Peterson J, Welch V, Losos M. 2009. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses. [accessed 2022 Jan 10]. http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm.
Whaley P, Aiassa E, Beausoleil C, Beronius A, Blotta G, Boobis A, de Vries R, Hanberg A, Hoffmann S, Hunt N, et al. 2020. Recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER). Environ Int. 143:105926.
Whaley P, Blauboer BJ, Brozek J, Cohen Hubal EA, Hair K, Kacew S, Knudsen TB, Kwiatkowski CF, Mellor DT, Olshan AF, et al. 2021. Improving the quality of toxicology and environmental health systematic reviews: what journal editors can do. ALTEX. 38(3):513–522.
Whaley P, Edwards SW, Kraft A, Nyhan K, Shapiro A, Watford S, Wattam S, Wolfe T, Angrish M. 2020. Knowledge organization systems for systematic chemical assessments. Environ Health Perspect. 128(12):125001.
Whaley P, Piggott T, Morgan RL, Hoffmann S, Tsaioun K, Schwingshackl L, Ansari MT, Thayer KA, Schümemann HJ. 2022. Biological plausibility in environmental health systematic reviews: a GRADE concept paper. Environ Int. 162:107109.
Whaley P, Halsall C. 2014. Literature review in toxicological research and chemical risk assessment: the state of the science. F1000Posters 2014, 5:375 (poster).
Whaley P, Halsall C. 2016. Assuring high-quality evidence reviews for chemical risk assessment: five lessons from guest editing the first environmental health journal special issue dedicated to systematic review. Environ Int. 92-93:553–555.
Whaley P, Isalski M. 2019. CREST_Triage. [accessed 2020 Nov 19]. https://crest-tools.site/.
[WHO] World Health Organization. 2021. Framework for the use of systematic review in chemical risk assessment. Geneva, Switzerland: World Health Organization. [accessed 2022 Jan 10]. https://apps.who.int/iris/handle/10665/347876.
Woodruff TJ, Sutton P. 2014. The navigation guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environ Health Perspect. 122(10):1007–1014.
## Appendix A: Table of tools and frameworks

In this manuscript, we frequently refer to several tools or frameworks in a single sentence. Citing these in each instance quickly becomes difficult to read; we therefore provide a table of abbreviations and citations for the frameworks and tools in question.

| Abbreviation | Full name | Website or manuscript citation |
|--------------|-----------|--------------------------------|
| AHRQ         | Agency for Healthcare Research and Quality | https://www.ahrq.gov/ |
| AMED         | Allied and Complementary Medicine | https://www.ebsco.com/products/research-databases/allied-and-complementary-medicine-database-amed |
| ARRIVE       | Animal Research: Reporting of In Vivo Experiments | https://arriveguidelines.org/arrive-guidelines |
| ASSIA        | Applied Social Sciences Index and Abstracts | https://proquest.libguides.com/assia |
| AXIS         | Appraisal tool for Cross-Sectional Studies | Downes et al. (2016) |
| CASP         | Critical Appraisal Skills Programme | https://casp-uk.net/casp-tools-checklists/ |
| CENTRAL      | The Cochrane Central Register of Controlled Trials | https://www.cochranelibrary.com/central |
| CINAHL       | Cumulative Index to Nursing and Allied Health Literature | https://www.ebsco.com/products/research-databases/cinahl-database |
| CNKI         | China National Knowledge Infrastructure | https://global.cnki.net/index/ |
| EPHPP        | Effective Public Health Practice Project Quality Assessment Tool for Quantitative Studies | https://www.ephpp.ca/PDF/Quality%20Assessment%20Tool_2010_2.pdf |
| EPOC         | Effective Practice and Organisation of Care | https://epoc.cochrane.org/about-us |
| GRADE        | Grading of Recommendations, Assessment, Development and Evaluations | Guyatt et al. (2008) |
| HEB Wales    | Health Evidence Bulletin Wales | |
| ICTRP        | International Clinical Trials Registry Platform | https://www.who.int/clinical-trials-registry-platform |
| JBI critical appraisal tool | Joanna Briggs Institute Critical Appraisal tool | https://jbi.global/critical-appraisal-tools |
| J-STOR       | Journal Storage | https://www.jstor.org/ |
| LilACS       | Latin American and Caribbean Health Sciences Literature | https://lilacs.bvsalud.org/en/ |
| MIDIRS       | Midwives Information & Resource Service | https://www.midirs.org/about-us/ |
| MMAT         | Mixed Methods Appraisal Tool | http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf |
| NHLBI        | National Heart, Lung, and Blood Institute | https://www.nhlbi.nih.gov/ |
| NOS          | Newcastle-Ottawa Scale | Wells et al. (2009) |
| NTP OHAT Framework | National Toxicology Program Office of Health Assessment and Translation Framework for Literature-Based Environmental Health Science Assessments | NTP-OHAT (2019) |
| PEDro        | The Physiotherapy Evidence Database | https://www.pedro.org.au/ |
| Pre-Post     | Quality Assessment Tool for Before–After (Pre-Post) Studies with No Control Group | https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools |
| PRISMA       | Preferred Reporting Items for Systematic reviews and Meta-Analyses | Moher et al. (2009) |
| ROBINS-I     | Risk Of Bias In Non-randomised Studies of Interventions | Sterne et al. (2016) |
| ROBINS-E     | Risk Of Bias In Non-randomised Studies of Exposures | https://www.bristol.ac.uk/population-health-sciences/centres/cresyda/barr/riskofbias/robins-e/ |
| Scielo       | Scientific Electronic Library Online | https://scielo.org/en/ |
| SinoMed      | Systemic review | http://www.sinomed.ac.cn/ |
| SR           | Strengthening the Reporting of Observational Studies in Epidemiology | Vandenbroucke et al. (2007) |
| STROBE       | SYstematic Review Centre for Laboratory animal Experimentation | https://www.syrclce.network/ |
| WoS          | Web of Science | https://www.webofknowledge.com/ |
Appendix B: Index of supplemental materials and study data

All supplemental materials for this study are archived online, in a permanent Open Science Framework registration that can be accessed here: https://osf.io/j2d5v/. The table below presents a guide and direct links to each item of supplemental material.

| Ref. | Description                                                                 | Link                          |
|------|-----------------------------------------------------------------------------|-------------------------------|
| SM00 | Preregistered protocol for conduct of this study, PDF                      | https://osf.io/mnpvs          |
| SM01 | Search strategy for locating eligible SRs, .docx                          | https://osf.io/85c6d/          |
| SM02 | Data collection form, PDF of Google Form                                   | https://osf.io/y7w29/          |
| SM03 | Results from search strategy, PubMed format                                | https://osf.io/j59ku/          |
| SM04 | List of SRs included after screening, .bib format                          | https://osf.io/qtrg6/          |
| SM05 | Raw data from data collection, prior to conflict resolution, .xlsx         | https://osf.io/uv5dg/          |
| SM06 | List of SRs included after screening, .xlsx format                         | https://osf.io/mgxwj/          |
| SM07 | Raw data from data collection, after conflict resolution, .xlsx            | https://osf.io/az4v8/          |
| SM08 | Screenshot showing anonymised version of an included SR                    | https://osf.io/62tpq/          |
| SM09 | Screenshot showing how web text is selected for anonymisation              | https://osf.io/e8s7c/          |
| SM10 | Final, cleaned data set for analysis with charts, .xlsx format             | https://osf.io/6emp8/          |
| SM11 | Hypothesis and normality tests, .xlsx format                              | https://osf.io/hrqya/          |
| SM12 | Contingency tables for exploratory analysis, .xlsx format                  | https://osf.io/h37kq/          |
| SM13 | List of SRs excluded during screening, .csv format                         | https://osf.io/y83tu/          |

Full-size, high-resolution figures are provided in the file “Manuscript Figures.” Captions listed in next section, as per submission instructions.