IMM report 1/2026

INSTITUTE OF ENVIRONMENTAL MEDICINE, IMM

Systematic reviews and meta-analyses of epidemiological studies in environmental research

Emilie Helte
Fredrik Söderlund
Ana Rodrigues
Agneta Åkesson

IMM
Institute of Environmental Medicine
Institutet för Miljömedicin

Karolinska Institutet

Systematic reviews and meta-analyses of epidemiological studies in environmental research

Institute of Environmental Medicine

Emilie Helte

Fredrik Söderlund

Ana Rodrigues

Agneta Åkesson

Karolinska Institutet

Table of contents

Published by the Institute of Environmental Medicine, Karolinska Institutet.

ISSN 1101-2803

IMM Report 1/2026

Preface

In 1993, the Cochrane Collaboration was founded, marking a pivotal moment in the development of systematic reviews. This organization formalized methodological standards for conducting reviews in the medical field, building on the legacy of Archie Cochrane, a medical researcher who had passed away just five years earlier. Cochrane had been a vocal critic of the medical practices of his time arguing that they were not sufficiently grounded in scientific evidence. Throughout his career, he championed the use of randomized controlled trials (RCTs) and the critical summarization of relevant evidence to inform clinical practice. Over the years, the field of systematic reviewing has continued to evolve. A growing body of guidelines and structured methodologies has been developed to reduce bias and improve reproducibility. Among the most influential are the PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), first published in 2009 and updated in 2020, which provides a standardized framework for reporting systematic reviews. Additionally, tools like GRADE (Grading of Recommendations Assessment, Development and Evaluation) have been introduced to assess the quality of evidence. The use of systematic reviews has expanded beyond medicine into fields such as education, psychology, social sciences, and environmental science, each adapting the methodology to suit its unique research questions and evidence types. Today, we are witnessing a new phase in the evolution of systematic reviews, characterized by the integration of artificial intelligence and machine learning. These technologies are increasingly used to assist researchers with tasks such as literature screening, data extraction, and even evidence synthesis, making the review process more efficient and scalable. Despite these advancements, it remains essential for researchers to understand the core steps of conducting a systematic review and to remain vigilant about potential sources of bias. Methodological rigor and critical thinking continue to be the cornerstones of trustworthy evidence synthesis.

Svensk sammanfattning

Systematiska översikter syftar till att sammanställa all befintlig vetenskaplig evidens inom en avgränsad forskningsfråga på ett strukturerat, transparent och objektivt sätt. Metaanalys används där det är lämpligt för att statistiskt kombinera resultat från flera studier. Denna rapport ämnas utgöra en metodologisk vägledning för genomförandet av systematiska översikter och metaanalyser inom epidemiologisk miljömedicinsk forskning, med särskilt fokus på etiologiska frågeställningar.

Följande centrala steg i arbetsprocessen tas upp i rapporten:

- **Formulering av forskningsfråga** enligt PI/ECOTSS-modellen (Population, Intervention/Exponering, Jämförelsemått, Utfall, Tidpunkt, Kontext, Studiedesign).
- **Upprättande av granskningsprotokoll** i enlighet med PRISMAs riktlinjer
- **Systematisk litteratursökning** med hjälp av strukturerade sökstrategier
- **Granskning och selektion av studier** med stöd av digitala verktyg.
- **Dataextraktion** för att säkerställa konsistens och kvalitet.
- **Bedömning av risk för bias** i de ingående studierna med hjälp av etablerade verktyg såsom ROBINS-E, RoB 2 och andra metodspecifika instrument.
- **Databearbetning och metaanalys**, inklusive hantering av heterogenitet och val av statistiska metoder
- **Samlad bedömning av evidensens styrka**, bland annat med hjälp av GRADE

Rapporten understryker vikten av metodologisk stringens, transparens och kritisk granskning för att säkerställa tillförlitliga och användbara slutsatser. Den belyser även vanliga metodologiska utmaningar i observationsstudier, såsom selektionsbias och informationsbias, samt hur dessa kan påverka

evidensens kvalitet i den samlade analysen. Avslutningsvis presenteras internationellt erkända riktlinjer för rapportering av metod och resultat (PRISMA och MOOSE) samt verktyg för kvalitetsgranskning av systematiska översikter (AMSTAR 2).

Abstract

The aim of a systematic review is to summarize the currently available evidence for a given research question. In addition, meta-analysis is often used to statistically combine results from multiple studies. This report provides a methodological guidance for conducting systematic reviews and meta-analyses in environmental epidemiology, with a particular focus on etiological research questions.

The report outlines the following key steps in the systematic review workflow:

- **Formulating the research question** using the PI/ECOTSS framework (Population, Intervention/Exposure, Comparator, Outcome, Timing, Setting, Study design)
- **Developing a review protocol** in accordance with PRISMA-P guidelines
- **Conducting a systematic literature search** using structured and reproducible strategies
- **Screening and selecting studies** for inclusion in the review
- **Extracting data**
- **Assessing risk of bias** in the included studies using validated tools
- **Synthesizing data and performing meta-analysis**, including selecting appropriate statistical models and handling heterogeneity
- **Evaluating the overall strength of evidence** using the GRADE approach as well as alternative frameworks

The report emphasizes the importance of methodological rigor, critical appraisal, and transparent reporting throughout the review process. It also highlights common sources of bias in observational studies and discusses their implications for evidence synthesis. Finally, the report introduces internationally recognized reporting standards (PRISMA and MOOSE) and tools for appraising the quality of systematic reviews (e.g., AMSTAR 2),

providing a robust foundation for researchers aiming to produce high-quality and policy-relevant evidence in environmental health.

Introduction

Systematic reviews are often ranked the highest when grading evidence, because they summarize all published original research relevant to a particular research question. By using a systematic method that is documented in advance, the risk of bias, resulting from e.g. subjective or incomplete inclusion of studies, can be reduced. Systematic reviews may, or may not, include a data synthesis part where data from different studies are combined using meta-analysis techniques.

In environmental research, most epidemiological studies of etiological research questions are observational in design. This is mainly because exposures like air pollution and environmental or dietary contaminants are difficult or unethical to randomize. Moreover, in those rare cases when randomized studies have been conducted, they often suffer from limitations like a short follow-up, which makes it challenging to study diseases with long latency periods.

Examples of observational study designs that are often used in environmental research include cohort studies, case-cohort studies, case-control studies, nested case-control studies, and cross-sectional studies. In addition, *in vitro* and animal studies may provide further understanding of the mechanisms by which environmental exposures contribute to disease. Irrespective of study design, the methodological quality of individual studies should be carefully assessed, which is a key component of the systematic review process.

Several handbooks and journal articles providing guidance for planning and conducting systematic reviews have been published. One example is the Cochrane Handbook for Systematic Reviews of Interventions, which focuses on intervention studies such as randomized controlled trials (RCTs) (1).

This report aims to describe the essential elements required for conducting a systematic review and meta-analysis of epidemiological studies addressing etiological questions in environmental research. The key steps will be described (**Figure 1**), and references to relevant sources of information will be provided.

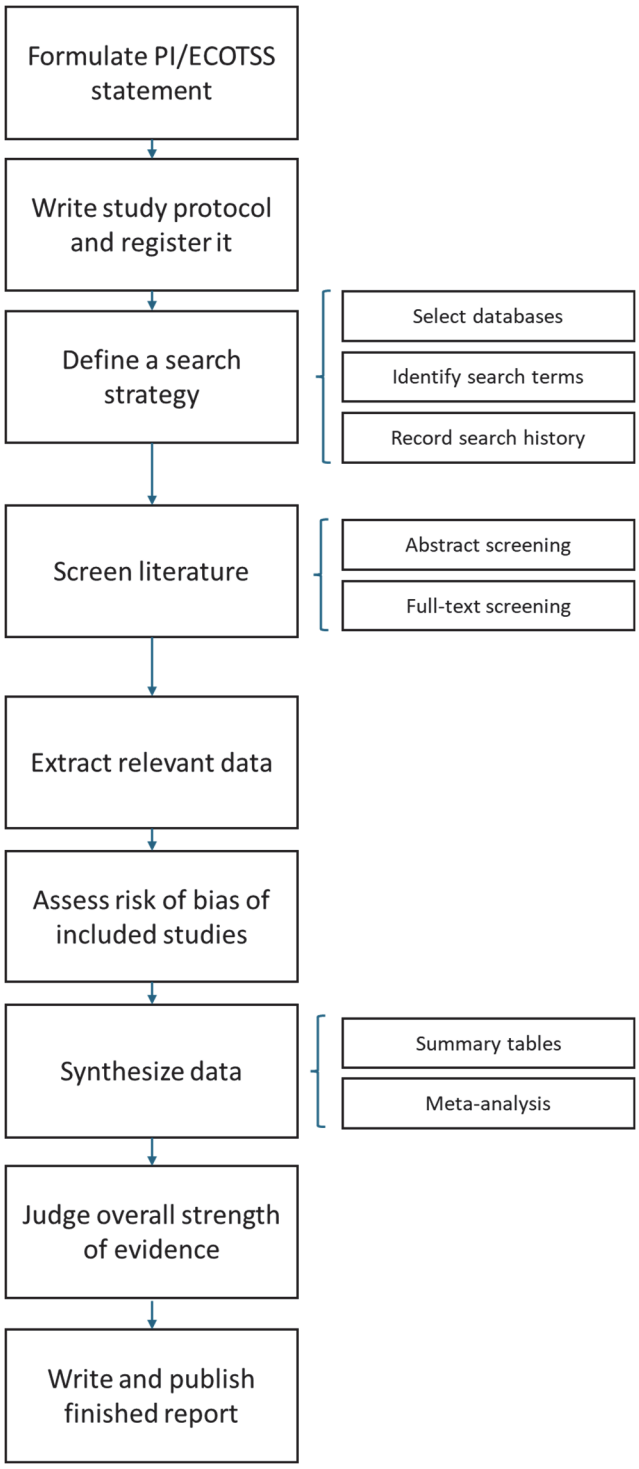


Figure 1. Flow diagram describing the key steps in conducting a systematic review.

Key steps for conducting a systematic review and meta-analysis

Conducting a systematic review involves several key steps. These include building a review team, framing the research question (defining a PI/ECOTSS statement), writing a study protocol, conducting a literature search, record screening, data extraction, risk of bias assessment, data synthesis including statistical methods for meta-analysis, and approaches for judging the overall body of evidence.

Building a review team

Undertaking a systematic review and meta-analysis is a resource-demanding task in terms of work and time. Therefore, it is important to have a well-qualified and collaborative review team engaged. A general requirement is that at least two people work independently on record screening, data extraction, and risk of bias assessment. It is also highly recommended to have a third member to discuss and solve disagreements with. Moreover, the review team should preferably possess all the necessary skills, including methods and content area expertise. Particularly the risk of bias assessment requires that the reviewers have expert knowledge both in methodology and in the research area to ensure reliable appraisals about the validity of the individual studies included in the review.

Framing the research question and developing a PI/ECOTSS statement

Arguably, the most important part of the systematic review process is framing the research question. The research question determines which papers will be included in the review and which will be excluded and therefore it has a major influence on the overall outcome of the review. For etiological research questions, it can be useful to define a Population, Intervention or Exposure, Control, Outcome, Timing, Setting, and Study design (PI/ECOTSS) statement, which is an extension of the well-known PICO framework. An explanation of each PI/ECOTSS domain, with an example, is provided in **Table 1**.

Table 1. PI/ECOTSS statement.

Elements of a Population, Intervention/Exposure, Control, Outcome, Timing, Setting, Study design (PI/ECOTSS) statement. For each domain, an example is provided for a hypothetical systematic review on arsenic exposure and lung cancer.

PI/ECOTSS domain	Explanation
Population	This domain refers to the population you would like to generalize your results to. Are you interested in adults, adolescents, or children? *Example: Healthy adults (≥18 years).*
Intervention/Exposure	This domain is about defining the exposure of interest. What is the exposure? How is it classified? How should it be measured? *Example: Chronic exposure to inorganic arsenic through residential drinking water, classified by measured concentrations in µg/L.*
Comparator	This domain is about defining the standard population against which the exposed population is compared against. It could be a non-exposed population or a population with a lower exposure level. *Example: Participants with higher arsenic concentrations in residential drinking water will be compared with those with lower concentrations.*
Outcome	This domain is about defining any primary and secondary outcomes you want to study. How is the disease defined? How are cases ascertained? *Example: Lung cancer is defined according to ICD-10 codes (C34) collected through national patient registries or national death registries.*
Timing	This domain defines the time required for the intervention/exposure to influence the outcome. Is there an incubation time? How long does it take to develop the outcome? *Example: Prospective cohorts must have >12 months of follow-up.*
Setting	This domain defines the relevant setting of which the results will be applicable to. Are you interested in the general population? Workers? Hospitalized patients? *Example: Relevant for the general population.*
Study design	This domain defines the type of study designs included to answer the research question. Are you only assessing intervention studies? Or specifically addressing the prevalence of a disease? What study design can answer your question? *Example: Prospective cohort studies, nested case-control studies, and case-cohort studies.*

Writing the study protocol

There are several reasons why it is important to write a protocol before starting a systematic review. First, drafting a protocol helps the reviewer plan the work in advance and identify potential issues early on. Second, the protocol documents the planned methods, analyses, and outcomes, which can later be compared with the published review to minimize the risk of selective reporting. Finally, if the protocol is published and publicly available, unnecessary duplication of work can be avoided.

The Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) is an extension of the PRISMA statement and provides helpful guidance on what to include when writing a protocol (2). The check-list, available online (https://www.prisma-statement.org/protocols), includes 17 items that should be described at a minimum, namely: background and rationale for conducting the study, the review question, the PI/ECOTSS statement, tabulation of important confounders, study eligibility criteria, search strategy including information sources for obtaining relevant studies, record screening, selection and data extraction, risk of bias assessment, data synthesis including statistical methods and planned analyses, and a method for judging the overall body of evidence.

The systematic review protocol should be registered so that it is accessible to others before any review tasks are initiated. PROSPERO, the international prospective register of systematic reviews, is an online database where systematic review protocols can be openly published and where users can search for ongoing but not yet published reviews (https://www.crd.york.ac.uk/prospero/). It is freely available to all and currently contains more than 362,000 records. The database was established in 2011 by the Centre for Reviews and Dissemination (CRD) at the University of York in response to calls from experts for improved documentation and public availability of systematic review protocols. It is funded by the UK National Institute for Health and Care Research (NIHR).

Figure 2. The PROSPERO interface (https://www.crd.york.ac.uk/prospero/).

Apart from PROSPERO, a few other organizations also publish protocols, including the Cochrane Collaboration, the Campbell Collaboration, and the Agency for Healthcare Research and Quality. Moreover, BioMed Central's open access journal *Systematic Reviews* specializes in publishing systematic review products including protocols. Nevertheless, before PROSPERO existed, very few completed systematic reviews referred to a published protocol in their methods.

Literature search

An exhaustive, thorough, objective, reproducible, and transparent search is one of the key features that distinguishes a systematic review from a narrative review. Developing and implementing a well-planned search strategy is therefore essential. A systematic review with an inadequate or inappropriate search strategy risks missing critical studies and may consequently introduce bias into the overall findings.

The search strategy should aim to capture all relevant reports while avoiding an overwhelming number of irrelevant ones, that is, achieving high sensitivity while maintaining reasonable precision. Designing an effective search strategy is a complex task, and it is therefore advisable to consult an experienced research librarian early in the process for guidance and support.

Key concepts, search terms, and search blocks

When developing a search strategy, it is generally advisable to return to the research question and the PI/ECOTSS statement. The elements of the statement can be used to identify key concepts and search terms that can be combined into search blocks. Which elements to include depends on the research question, but "Population" and "Intervention or Exposure" are often central, while "Outcome" and "Study design" are also frequently incorporated. However, including too many PI/ECOTSS elements will narrow the search and increase the risk of missing articles relevant to the research question. In general, systematic literature searches are designed to be broad, meaning they achieve high sensitivity but often at the cost of lower precision.

After deciding which key concepts to use, each should be examined carefully to identify synonyms, related terms, and potential subject headings (such as keywords or Medical Subject Headings (MeSH) terms). Thereafter, free-text words and controlled subject headings are used to construct search blocks. The next step is to combine the search terms or search in a way that aligns with the logic of the database being searched. This is typically done using the Boolean operators "AND", "OR", and "NOT" which can be used in almost all medical literature databases. An example of how to combine free-text terms or subject headings representing the same key concept into search blocks and how to combine different blocks is shown in **Figure 3**. To ensure transparency and traceability, the search history should be saved and published alongside the systematic review.

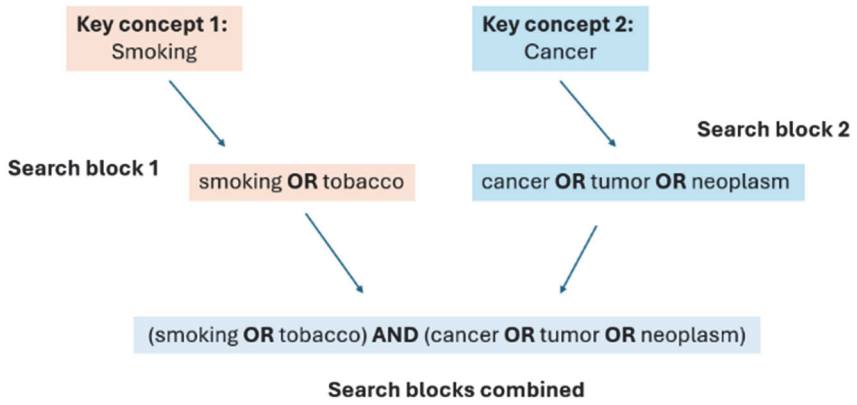


Figure 3. Constructing and combining search terms and search blocks.

Data sources
A systematic review literature search should be conducted in at least two different databases. In medicine, the most frequently used databases are PubMed, MEDLINE, Embase, and the Cochrane Central Register of Controlled Trials. Multidisciplinary databases such as Scopus and Web of Science are also widely used and include research from fields such as technology, social sciences, and the humanities. Depending on the topic, subject-specific databases, such as CINAHL, which indexes literature in nursing, physiotherapy, and occupational therapy, may also be appropriate. It may also be worth contemplating whether it is necessary to search grey literature or other unpublished material, or to examine existing reviews on the same topic and their reference lists, a process often referred to as *snowballing*.

Record screening and study selection

Because a systematic literature search aims to be broad, it is common for only 2–3% of retrieved records to be relevant to the research question. Nevertheless, only records that meet the pre-specified eligibility criteria, as outlined in the PI/ECOTSS statement, should be included in the review. The processes of identifying and selecting these reports are referred to as record screening and study selection.

Record screening can be divided into two steps that differ in the level of detail in the report being assessed: (1) title and abstract screening and (2) full-text screening. Before screening begins, duplicate records, that is those appearing in more than one database, should be removed to ensure that each study is screened only once.

To reduce bias, the process should be carried out by at least two reviewers who are blinded to each other's decisions. Blinding is lifted after each round of screening, and any disagreements should be resolved before proceeding to the next step. If consensus cannot be reached, a third assessor, often a more senior team member, may be consulted.

Title and abstract screening
The purpose of title and abstract screening is to quickly remove records that are clearly irrelevant to the research question, which helps save time and reduce the workload during full-text screening. In this initial step,

assessors review only the title and abstract of each report and decide whether it should proceed to the next stage or be excluded, based solely on this information. Because most retrieved reports are typically excluded at this stage, it is advisable to be over-inclusive to avoid mistakenly discarding potentially relevant studies. The number of excluded studies should be documented, preferably in a flow diagram (see section *Tracking study selection using a flow diagram*).

Full-text screening

Once title and abstract screening is complete, all records that move on to full-text screening should be sought for retrieval. Any records that cannot be obtained must be documented and clearly reported. Next, the reviewers read the entire study and assess it against the detailed eligibility criteria outlined in the PI/ECOTSS statement. The reasons for exclusion should be documented.

It is common for multiple reports to originate from the same study or trial, and the choice of which report to include can influence the findings of the meta-analysis. Therefore, it is recommended to establish in advance how such cases will be handled, for example by prioritizing the report with the longest follow-up or the most recent publication. Another option is to extract and combine relevant information from all related reports to maximize data completeness. Regardless of the approach, it is essential to ensure that the same study population is not included more than once in a meta-analysis, as this would bias the results.

Finally, all studies that meet the eligibility criteria after full-text assessment are included in the review.

Tools to facilitate the screening and selection process

Screening and selecting studies for inclusion in a systematic review is a labor-intensive task. To improve efficiency and accuracy, it is therefore advisable to use a dedicated tool that facilitates the process. Several such tools are available, each with its own set of features. Below is a summary of some of the most widely used ones.

Rayyan (https://www.rayyan.ai/) is a web-based tool designed to accelerate the systematic review process and is particularly useful for title and

abstract screening. It offers AI-assisted features that help identify duplicate records and highlight studies that are likely to be relevant to the research question. Rayyan supports collaboration by allowing multiple reviewers to work on the same project while remaining blinded to each other's decisions until blinding is lifted. Most core features are available for free, while some advanced options require a subscription. The tool is easy to use, quick to set up, and is best suited for quick title and abstract screening.

Covidence (https://www.covidence.org/) is a web-based platform developed to streamline the entire systematic review process. It is more advanced than Rayyan and offers more features, including data extraction to customized forms, risk of bias assessment, PRISMA flow charts, etc. Multiple reviewers can work together on the same project blinded towards each other. There is no free version of Covidence, but KI staff, researchers, and students have access to an institutional license. Covidence is best suited for researchers who wish to receive structured end-to-end assistance with their systematic review.

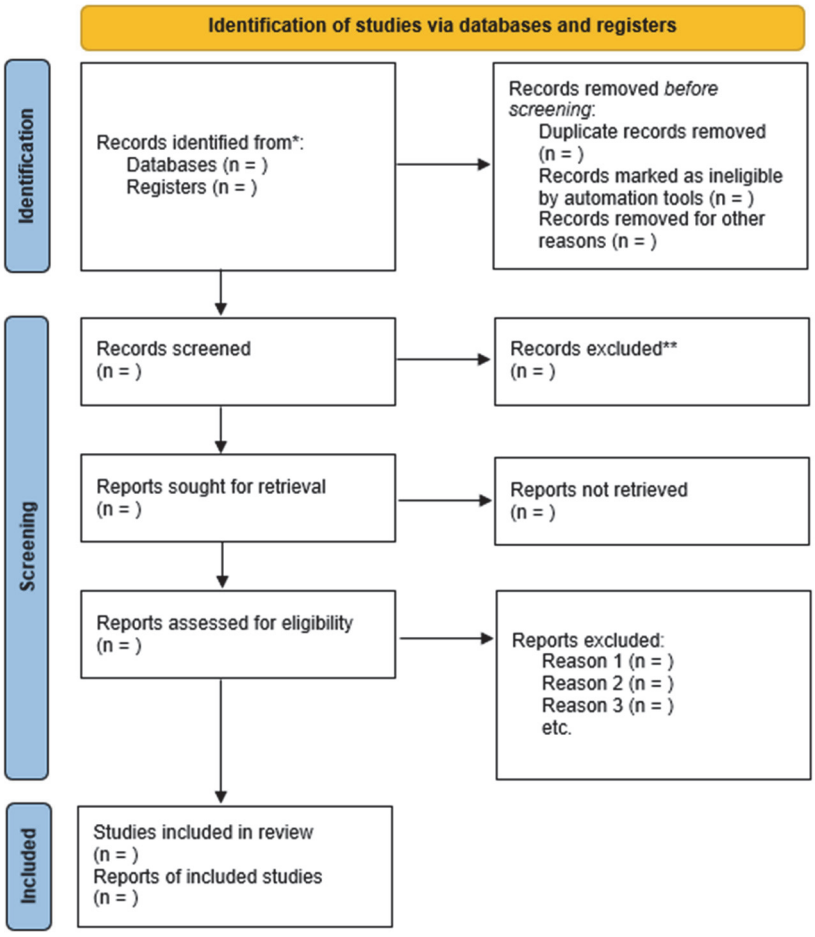
Distiller (DistillerSR | Systematic Review Software | Literature Review Software) is a web-based tool that supports the entire systematic review process, from literature search to data collection and reporting. It uses AI and intelligent workflows to streamline and automate tasks, reducing the workload for the entire review team. Distiller is only available to paying subscribers.

EPPI-reviewer (EPPI-Reviewer: systematic review software (ioe.ac.uk)) was developed by the EPPI-Centre at the University College London, and the tool recommended by Cochrane for systematic review authors. Like Covidence and Distiller, it assists in all steps of the systematic review process, but also in conducting meta-analyses. EPPI-reviewer is subscription-based but offers free trial periods.

Tracking study selection using a flow diagram

The entire screening process, including the number of records excluded at each step and the reasons for exclusion, should be documented in a flow diagram. Templates developed by PRISMA are available online (see **Figure 3**) and are recommended for use. The flow diagram should, at minimum, include the number of records identified in the search, the number of

records screened and excluded at each stage, the reasons for exclusion after full-text screening, and the number of records finally included in the systematic review.



*Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers).

**If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.

Figure 4. PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only. Retrieved from https://www.prisma-statement.org/prisma-2020-flow-diagram

Data extraction

Thorough and accurate data extraction is essential when conducting a systematic review, as it directly influences the review's findings. To minimize the risk of introducing incorrect information, data should be extracted independently by at least two assessors, alternatively by one assessor with a second assessor performing a thorough check.

The data extracted in a systematic review typically includes general study information (authors, year of publication), study design, study population (participation rate, age ranges, proportions of women and men enrolled, number of cases and controls), intervention and setting (exposure of interest, exposure levels and their units, model adjustments), and outcome data or results (outcome of interest, effect size measure, effect estimates, confidence intervals). The specific data to be extracted can be adapted to the needs of individual projects and may include both fewer and more categories (e.g. exposure/outcome measurement methods, limits of detection, type and characteristics of control groups, other quantitative results, outcome definitions, etc.).

To ensure consistency across studies, it is often helpful to use a standardized extraction form. Such forms should be developed before data extraction begins, and their performance should be pilot tested with a few studies. Several tools commonly used to facilitate screening also provide functionalities that support data extraction and allow users to create customized extraction forms (see section *Tools to facilitate the screening and selection process*).

Risk of bias assessment

The purpose of a risk of bias assessment is to evaluate the quality of the individual studies included in a systematic review and thereby inform the reader about the trustworthiness and validity of the synthesized evidence. Because the process relies heavily on expert judgment, it is inevitably subjective. It is therefore essential to use a structured risk of bias tool to bring consistency and objectivity to the assessment (see section *Tools for assessing risk of bias*). To ensure that assessments are as accurate as possible, the topic of the research question should fall within the reviewer's

area of expertise. The risk of bias assessment should be conducted in duplicate.

Common types of bias in epidemiological studies

Bias in epidemiological studies refers to systematic errors in the design, conduct, analysis, or reporting of a study that result in incorrect estimations of associations between the exposures and outcomes. Presence of bias in individual studies will distort the validity of the systematic review and may result in misleading conclusions. Multiple sources of bias exist, but the most common types in epidemiological research are selection bias, information bias (misclassification of the exposure or the outcome), and confounding. A brief summary of these types of bias is provided below. For further reading on bias in epidemiological studies, *Modern Epidemiology* by Rothman and colleagues is recommended (3).

Selection bias occurs when the study population is not representative of the source population, meaning that the observed association in a study differs from what it would have been had the entire source population been observed. Selection bias can arise for several reasons. It is most commonly a concern in case-control studies, where it may occur due to unrepresentative sampling of controls. However, it can also arise in cohort studies, for example when loss to follow-up is related to the exposure or outcome of interest, or as a result of collider bias.

Information bias arises from errors in the measurement of exposures, outcomes, or other relevant variables. When these errors are related to the exposure or outcome of interest, the resulting bias is referred to as differential misclassification, which may lead to under- or overestimation of the "true" effect. Non-differential misclassification occurs when the misclassification is similar across comparison groups and is generally less concerning, as it tends to bias associations toward the null. In case-control studies, differential misclassification of the exposure can occur due to recall bias, as the disease itself may influence participants' ability to remember past exposures, causing cases and controls to report the exposure differently.

Confounding occurs when the exposed and non-exposed groups differ with respect to a third variable that also influences the outcome (**Figure 4**).

Confounding can be addressed either through study design, such as randomization, or, more commonly in observational studies, through statistical adjustment. In practice, it is difficult to fully eliminate confounding, and therefore the risk of unmeasured or residual confounding must always be carefully considered when evaluating the validity of observational studies.

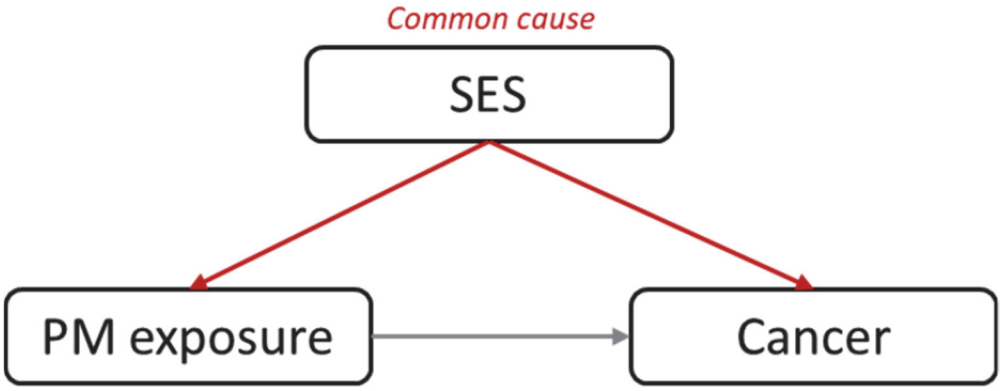


Figure 5. Simple Directed acyclic graph (DAG) depicting how a common cause/confounder is related to both the exposure and outcome of interest simultaneously. With SES = Socioeconomic status and PM = Particulate matter.

Tools for assessing risk of bias

Several tools are available for assessing risk of bias. Although they differ in structure and detail, they share some core features. First, they aim to increase transparency and replicability by providing assessors with structured frameworks and workflows. Second, nearly all tools evaluate bias related to exposure or outcome misclassification (information bias), confounding, and selection bias. Finally, most tools use checklists and/or signalling questions, often in combination with scoring or rating systems, to classify individual studies both within specific domains and in terms of their overall risk of bias.

The main differences between tools lie in how they define and address the various domains of bias. As a result, a tool may be more or less suitable depending on factors such as the research question, the study design(s) of

the included studies, and the type of exposure or outcome examined. Choosing an appropriate tool therefore requires careful consideration of these factors. Some of the most widely used tools for reviews of epidemiological studies are presented in **Table 2** along with their main characteristics. The list of available tools is, however, not limited to these and appropriate alternatives can be found elsewhere. New tools which are specific to certain types of research questions or disciplines are also continuously being developed. One recent example is the ROBINS-E tool (risk of bias assessment in non-randomized studies of exposures), which is a Cochrane tool particularly well suited for assessing bias in observational epidemiological studies with a focus on exposures.

Once the risk of bias assessment is completed, blinding should be lifted and any inconsistencies between assessors' judgments resolved. The harmonized results should then be presented transparently in a summary table. A heatmap can provide a clear visual overview of how individual studies performed in the risk of bias assessment (**Figure 6**). *Robvis* is an online tool for creating visualizations of risk-of-bias assessments (https://mcguinlu.shinyapps.io/robvis/).

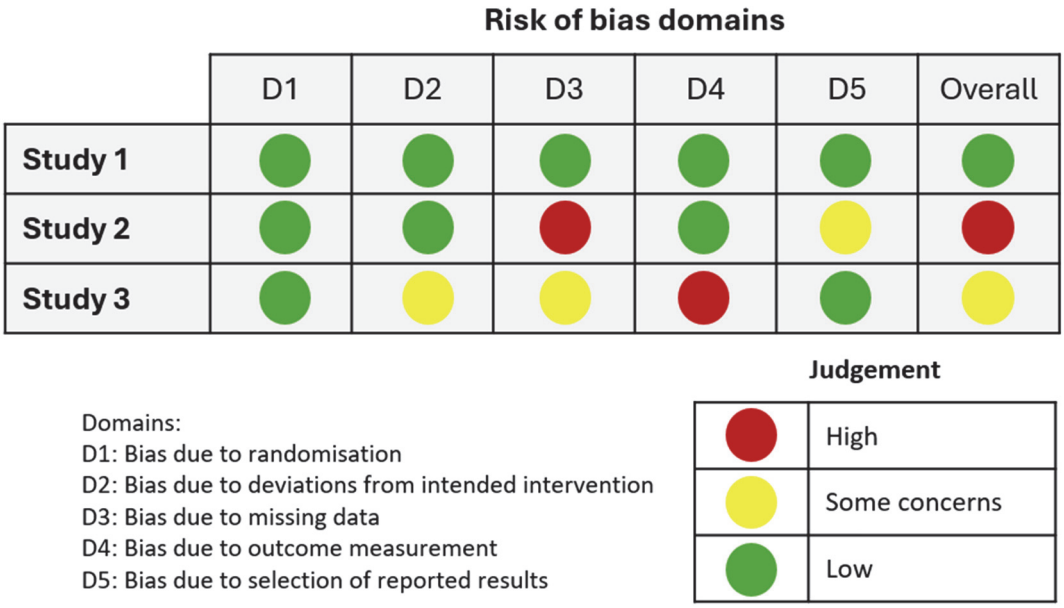


Figure 6. Risk of bias heatmap.

Table 2. Common tools for assessing risk of bias.

Name of tool	Intended use
Cochrane methods	
ROBINS-E	Can be used for risk of bias assessment in non-randomized studies of exposures (Risk of bias tools - ROBINS-E tool).
ROBINS-I V2	Updated and can be used for risk of bias assessment in non-randomized studies of interventions (Risk of bias tools - ROBINS-I V2 tool).
RoB 2	Updated and can be used for risk of bias assessment in randomized controlled trials (Risk of bias tools - RoB 2 tool).
ROB-ME	Can be used to assess bias due to missing data in the individual included studies (Risk of bias tools - ROB-ME tool).
Other methods and tools	
NTP OHAT Risk of Bias Tool	Can be used for risk of bias assessment in human and animal studies that assess the impact of an exposure on an outcome (Risk of Bias Tool).
SYRCLE's Risk of Bias Tool	Largely based on the RoB tool, but with its intended use for risk of bias assessment of animal intervention studies (SYRCLE's risk of bias tool for animal studies | BMC Medical Research Methodology)
Newcastle-Ottawa Scale	A simple but widely used tool for appraising cohort and case-control studies. Less robust and comprehensive than newer domain-based tools.
ToxRTool	Designed for *in vitro* and *in vivo* toxicology studies developed by the European Centre for the Validation of Alternative Methods (ECVAM) of the Joint Research Centre (JRC) of the European Commission (ToxRTool - Toxicological data reliability assessment tool - European Commission).
Navigation Guide	Used to systematically and transparently evaluate environmental health research and to support science-based policy decisions (Navigation Guide | Program on Reproductive Health and the Environment).

Data synthesis

Summary tables

Summary tables are an essential component of a systematic review, providing a transparent and organized presentation of individual study findings. A clear and structured summary table contains one row for each study. The columns should include, but are not limited to, first author and publication year, brief information on the study population, study design, exposure, a brief description of the comparison group, outcome(s), effect estimates, and risk of bias or grading.

Meta-analysis

Meta-analysis is a statistical method used to combine results from multiple individual studies that address a similar research question. This approach increases power and precision and provides an overall estimate of the effect. A meta-analysis should always be preceded by a systematic review, but it is not an essential component of one. Whether it is appropriate to conduct a meta-analysis depends on several factors, including the nature of the research question (for example, whether it is quantitative), the number of studies included in the review, the extent of clinical and methodological heterogeneity, and the quality of the included studies. If the number of studies is small (fewer than 5), study quality is poor, or methodological heterogeneity is substantial, a qualitative synthesis may be more appropriate, since statistical pooling of such evidence could produce misleading conclusions.

Inverse variance approaches

Most methods used to conduct a meta-analysis follow a two-stage process. First, effect sizes and standard errors (SE) are extracted from each study and, when necessary, converted to ensure consistency across studies. An effect size is a value that estimates the magnitude of group differences or the strength of an association between two variables. The type of effect size reported in a study depends on the research question. For example, when data are dichotomous and exposed and unexposed groups are compared with respect to the risk of developing a disease, the effect size is typically a relative risk or an odds ratio. When data are continuous, the effect size of

interest may be a regression coefficient, such as β representing a mean difference.

In the second step, the summary intervention or exposure effect is estimated as a weighted average of the effect estimates from the individual studies. Several different methods exist and while they all estimate the weighted average as

$$\frac{sum\ of (estimate\ \times\ weight)}{sum\ of\ weights}$$

they differ in how the weights are defined and in their underlying assumptions.

The ***fixed effect method*** assumes that a single "true" effect underlies all estimated effects in the included studies, and that any differences between study estimates are due solely to sampling error. In this method, weights are defined as

$$\frac{1}{SE^2}$$

so that the weighted average is

$$\frac{\sum Y_i(1/SE_i^2)}{\sum(1/SE_i^2)}$$

As a result, larger studies with greater precision will be given more weight than smaller studies.

In contrast, the ***random effect method*** assumes that the "true" effect varies across studies because of differences in, for example, study populations, methods, or other factors. This model incorporates an additional source of variability, called *between-study variability* and denoted τ^2, into the calculation of the weights, producing the following equation:

$$\frac{\sum Y_i(1/SE_i^2)}{\sum(1/SE_i^2 + \tau^2)}$$

Although larger studies still tend to receive more weight in a random effect meta-analysis, the weights are more balanced than under the fixed effect approach and reflect uncertainty about the underlying effect across studies. Between-study variability can be estimated using several methods, with

Maximum Likelihood Estimation (MLE), Restricted Maximum Likelihood (REML), and Der Simonian and Laird being among the most commonly used. Results from a meta-analysis are typically presented in a forest plot.

Heterogeneity

When synthesizing evidence from different studies in a meta-analysis, it is expected that observed effect sizes will vary because of sampling variability. However, when this variation exceeds what would be expected by chance, it is referred to as between-study heterogeneity. Heterogeneity may arise from clinical diversity, such as differences in study participants, interventions, or outcomes across studies, or from methodological diversity, which reflects variations in study design and conduct. It may also result from a combination of these factors (4).

Between-study heterogeneity can be assessed visually by examining forest plots. Limited or no overlap between plotted confidence intervals (CI) from individual studies suggests the presence of heterogeneity. Heterogeneity can also be tested formally, for example using Cochrane's Q statistic, where a low p-value indicates evidence of heterogeneity, or quantified using I^2 statistics. The I^2 statistic describes the proportion of the variability that is due to heterogeneity rather than sampling error. The Cochrane Collaboration provides a general guide for interpreting I^2 values, summarized in **Table 3**.

Table 3. Suggested thresholds for the interpretation of I^2 statistics.

Percentage of variability due to heterogeneity	**Interpretation***
0% to 40%	Might not be important
30% to 60%	May represent moderate heterogeneity
50% to 90%	May represent substantial heterogeneity
75% to 100%	Considerable heterogeneity

**Recommendations according to the Cochrane Handbook for Systematic Reviews of Interventions version (4).*

When substantial heterogeneity is present, subgroup meta-analysis or meta-regression may be used to explore potential sources. These methods both assess the relationship between effect sizes and study level covariates, so called *moderators*, that may modify the effect of the exposure on the outcome.

Subgroup meta-analysis is typically applied when the covariates are categorical, while meta-regression is suitable when at least one covariate is continuous or when the goal is to examine the influence of several factors at the same time. In a subgroup meta-analysis, studies are divided according to the covariate of interest and the pooled estimates for each subgroup are then compared.

Meta-regression can be seen as an extension of subgroup meta-analysis. It is a weighted linear regression in which the studies are the observations, the effect estimate is the dependent variable, and the study level covariates serve as the independent variables. It differs from a traditional linear regression by that it is weighted, meaning the studies with more weight have a greater influence on the relationship than smaller studies with lower weight.

Two types of meta-regression exist: fixed effect meta-regression and random effect meta-regression. The difference between the two is that while fixed effect meta-regression assumes that all heterogeneity is accounted for by the moderators, random effect meta-regression allows for residual heterogeneity that is not explained by the independent variables. The random effect approach is generally recommended because, in practice, moderators rarely account for all between-study heterogeneity.

The output of meta-regression is a regression coefficient with a 95% CI that represents the change in the log risk ratio associated with a one unit increase in the covariate. To ensure reliable estimation, it is typically recommended to include at least 10 studies per moderator.

Other meta-analysis approaches

Dose-response meta-analysis is a specialized form of meta-analysis that estimates the change in response along the range of a continuous exposure by combining findings from multiple studies. Several modelling approaches

exist, but the most commonly applied is the two-stage model. In this approach, the expected shape of the exposure-response relationship is first specified based on content knowledge, and then within-study characteristics are synthesized across studies (5).

Dose-response meta-analysis can be useful when studies report multiple levels of exposure, for example different nutrient intakes or pollutant concentrations. When using standard meta-analysis methods for such data, it is common to only focus on the contrast between the highest and the lowest exposed. Although this approach is not incorrect, it discards valuable information and may introduce heterogeneity, since the magnitude of the association can vary across the full exposure range.

Network meta-analysis (NMA) combines evidence from multiple studies to compare three or more interventions within a single analytical framework. NMAs incorporate both direct comparisons, which come from studies comparing interventions head-to-head, and indirect comparisons, which estimate the relative effects of interventions that have not been directly compared in any single study. By synthesizing all available evidence across the intervention network, an NMA provides estimates for every pairwise comparison and allows for ranking of the interventions.

NMAs rely on several key assumptions: transitivity, consistency, and network connectivity. The assumption of transitivity dictates that all studies included in the network should be on average similar in all important features except for the interventions being compared. When NMAs include observational studies, satisfying this assumption can be more challenging because study populations or confounding structures may differ more substantially than in randomized trials.

The consistency assumption dictates that direct and indirect comparisons should only be integrated within one and the same NMA if their findings are of similar magnitude and direction and, thus, allowing them to be comparable to each other. Consistency can be assessed locally, by comparing specific pairwise estimates, or globally, by evaluating the coherence of the network as a whole. Violations of consistency may suggest that the transitivity assumption might not hold.

When conducting an NMA, it should be decided under which circumstances it is beneficial to extend an NMA using evidence from RCTs with data from observational studies or even whether to rely solely on observational data. One main reason for including observational studies may be the scarcity of evidence from RCTs, where integrating observational data can improve the connectivity of the network. However, one of the pitfalls of using observational studies is the increased risk of bias and, to uphold the assumption of transitivity, that data from both randomized and non-randomized studies need to be sufficiently compatible.

Judgment of the overall strength of evidence

The final step in the systematic review and meta-analysis process is to assess the overall strength of the evidence supporting the estimated effect size. The rationale for doing this is that the usefulness of the estimate depends on our confidence in it. When evidence is considered to be of low quality, it can guide future research by highlighting knowledge gaps and help researchers prioritize areas that need more investigation.

The arguably most popular method for ranking the certainty in the evidence is the *Grades of Recommendation, Assessment, Development and Evaluation (GRADE)* approach. GRADE was developed by the GRADE working group and is the recommended system by several international organizations including the World Health Organization (WHO) and The Cochrane Collaboration (6). The GRADE system classifies evidence into one of four categories of varying quality: high, moderate, low, and very low quality. RCTs always begin as high-quality evidence, while observational studies start as low-quality evidence. Quality can then be downgraded based on study limitations if results are inconsistent, indirect, or imprecise, or if publication bias is suspected. Similarly, quality can be upgraded if the evidence includes a large magnitude of effect, if a dose-response gradient is present, or in a situation when all plausible biases would decrease the magnitude of effect (7).

While widely used and recommended, GRADE is not the only method available. As an example, the World Cancer Research Fund (WCRF), which

focuses on how diet, nutrition, and physical activity influence cancer risk and survival, has developed its own approach to evaluate evidence. Given that most human studies in this field are observational, GRADE would often rate the certainty of evidence as low, even when rigorous research exists. While recognizing the challenges in establishing causality from observational data, WCRF has created a systematic and integrated method based on the Bradford Hill criteria to make well-founded judgments and reliable recommendations based on the best available evidence (8).

Writing the report

After completing all steps in the systematic review and meta-analysis process, the methods used and the findings obtained should be thoroughly documented. As with all publications, the readers' ability to critically appraise a systematic review depends heavily on the quality of the reporting. Ensuring high reporting standards is therefore essential.

The *Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)* statement was first published in 2009 to improve the reporting quality of systematic reviews (9), and it was updated in 2020 to incorporate advances in systematic review methodology (10). Apart from the statement paper, which describes how the reporting guideline was developed, PRISMA also consists of a checklist covering 27 items that should be reported in a systematic review (PRISMA 2020 checklist — PRISMA statement). There is also an expanded checklist that contains detailed recommendations for each reporting item, an abstract checklist (Abstracts — PRISMA statement), and templates for creating PRISMA flow diagrams (PRISMA 2020 flow diagram — PRISMA statement). The PRISMA guidelines have been widely accepted by the scientific community, and many journals require authors to submit a completed PRISMA checklist along with their manuscript when submitting a systematic review.

In addition to the PRISMA guidelines, specific reporting guidelines for meta-analyses that include observational studies also exist. The *Meta-analyses of Observational Studies in Epidemiology (MOOSE)* guidelines were developed to ensure that study quality and potential sources of bias are adequately addressed when incorporating observational studies in a meta-analysis. By adhering to these guidelines, readers are able to better understand potential

biases and to evaluate the quality of the studies included in the meta-analysis. Like PRISMA, MOOSE consists of a checklist that covers 35 required items (ISSM_MOOSE_Checklist.pdf).

Methodological considerations

Systematic reviews are widely regarded as the highest level of scientific evidence and can provide a comprehensive overview of the current state of knowledge in a given field. However, accepting their conclusions without critical appraisal carries risks. Various methodological factors can influence the validity and reliability of a systematic review which underscores the importance of careful evaluation.

A MeaSurement Tool to Assess systematic Reviews 2 (AMSTAR 2) is an instrument developed for critical appraisal of systematic reviews (12). In their statement paper, the authors mention that there are certain factors that are particularly sensitive to bias and should be considered extra carefully. First, adherence to a well-written protocol reduces the risk of selective reporting bias, that is when reviewers selectively choose what to include in the review based on the direction and significance of findings. Second, that the search methods are sufficiently exhaustive and that exclusions are justified are also important since these factors will have a direct effect on which studies are included in the review and, hence, the overall results obtained. Risk of bias assessment of individual studies and accounting for this when interpreting review findings is another factor that needs some additional attention. If this is insufficiently done, authors risk missing out on potential biases that distort associations in individual studies which in turn will impact the reliability of the review findings. Appropriate meta-analysis techniques are also listed as critically important, since pooled estimates of incompatible data may result in unreliable estimates.

A final factor that is widely recognized as one of the major potential threats to validity in systematic reviews is publication bias. *Publication bias* occurs when studies that show statistically significant findings are more likely to be published as compared to studies presenting null findings and, thus, the null results remain invisible to the reviewers. Bias due to missing evidence may impact the results of meta-analyses as effect estimates may be unavailable from the included studies and lead to the distortion of the “true” effect of a meta-analysis.

Different statistical methods have been developed to investigate publication bias. A *funnel plot* is a scatter plot used to detect small study effects, often indicating publication bias (13). In a funnel plot, each point represents a study, the x-axis the effect sizes, and the y-axis a measure of study precision. In absence of small study effects, the shape of the plot should be a symmetrical inverted funnel. Asymmetrical funnel plots can also arise from between-study heterogeneity. Hence, it is important to address this prior to assessing publication bias.

More formal statistical tests for publication bias also exist. The *Egger's test* is probably the most well-known test and is a regression-based test that performs a linear regression of the effect sizes and the standard errors weighted by their inverse variance. The test for the zero slope is the test for publication bias. The power of the test is, however, low and the capacity of detecting bias when studies are few is limited (14).

Summary and closing words

This report outlined the essential steps for conducting systematic reviews and meta-analyses in environmental epidemiology. It covered the full workflow, from defining the research question and protocol development to literature search, study selection, data extraction, risk of bias assessment, and data synthesis. Tools that can streamline the process together with reporting guidelines and quality rating tools are highlighted to ensure methodological rigor and transparency.

Systematic reviews are powerful instruments for evidence synthesis, but their value depends on careful planning, critical appraisal, and transparent reporting. By following best practices and critically appraising each step, researchers can produce trustworthy syntheses that inform public health and environmental policy.

Further reading

The Cochrane Handbook of Systematic Reviews: Higgins, J. P. T., Green, S., & Cochrane Collaboration. (2008). *Cochrane handbook for systematic reviews of interventions*. Wiley-Blackwell.

Guidance on conducting systematic reviews of observational studies: Dekkers OM, Vandenbroucke JP, Cevallos M, Renehan AG, Altman DG, Egger M. COSMOS-E: Guidance on conducting systematic reviews and meta-analyses of observational studies of etiology. PLoS Med. 2019 Feb 21;16(2):e1002742.

Epidemiological methods: Lash, T. L., VanderWeele, T. J., Haneuse, S., & Rothman, K. J. (Eds.). (2021). *Modern epidemiology* (4th ed.). Wolters Kluwer / Lippincott Williams & Wilkins.
Meta-analysis methods: Egger, M., Higgins, J., & Davey Smith, G. (2022). *Systematic reviews in health research : meta-analysis in context* (Third edition.). Wiley Blackwell/BMJ Books.

GRADE framework: https://www.gradeworkinggroup.org/

References

1. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions. 2nd edition ed. Chichester (UK): John Wiley & Sons; 2019.

2. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews. 2015;4(1):1.

3. Lash TL, VanderWeele TJ, Haneuse S, et al. Modern epidemiology. 4th ed. ed. Phil: Wolters Kluwer / Lippincott Williams & Wilkins; 2021.

4. Deeks J, Higgins J, Altman D, et al. Chapter 10: Analysing data and undertaking meta-analyses. Cochrane Handbook for Systematic Reviews of Interventions version 65. London: The Cochrane Collaboration; 2024.

5. Orsini N, Larsson SC, Salanti G. Chapter 14: Dose-Response Meta-Analysis. Systematic Reviews in Health Research : Meta-Analysis in Context. Newark, UNITED KINGDOM: John Wiley & Sons, Incorporated; 2022.

6. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ : British medical journal. 2008;336(7650):924–6.

7. Guyatt GH, Oxman AD, Kunz R, et al. What is "quality of evidence" and why is it important to clinicians? BMJ : British medical journal. 2008;336(7651):995–8.

8. World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project Expert Report 2018. Judging the Evidence. 2018.

9. Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Annals of Internal Medicine. 2009;151(4):264–9.

10. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. 2021;372:n71.

11. Brooke BS, Schwartz TA, Pawlik TM. MOOSE Reporting Guidelines for Meta-analyses of Observational Studies. JAMA Surgery. 2021;156(8):787–8.

12. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. 2017;358:j4008.

13. Sterne JAC, Harbord RM, Sterne JAC, et al. Funnel Plots in Meta-analysis. The Stata journal. 2004;4(2):127–41.

14. Egger M, Davey Smith G, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. Bmj. 1997;315(7109):629–34.

ki.se/imm

IMM

Institute of Environmental Medicine
Institutet för Miljömedicin

Karolinska Institutet