Challenges and recommendations on the conduct of systematic reviews of observational epidemiologic studies in environmental and occupational health

Whitney D. Arroyave¹, Suril S. Mehta², Neela Guha³,⁶, Pam Schwingl¹, Kyla W. Taylor², Barbara Glenn⁴, Elizabeth G. Radke⁴, Nadia Vilahur³, Tania Carreón⁵, Rebecca M. Nachman⁴, Ruth M. Lunn²,*

¹Integrated Laboratory Systems, Morrisville, North Carolina, USA
²Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA
³International Agency for Research on Cancer, World Health Organization, Lyon, France (former affiliation)
⁴Center for Public Health and Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC, USA
⁵World Trade Center Health Program, National Institute for Occupational Safety and Health, Cincinnati, Ohio, USA
⁶Office of Environmental Health Hazard Assessment, California Environmental Protection Agency, Oakland, California, USA

Abstract

Systematic reviews are powerful tools for drawing causal inference for evidence-based decision-making. Published systematic reviews and meta-analyses of environmental and occupational epidemiology studies have increased dramatically in recent years; however, the quality and utility of published reviews are variable. Most methodologies were adapted from clinical epidemiology and have not been adequately modified to evaluate and integrate evidence from observational epidemiology studies assessing environmental and occupational hazards, especially in evaluating the quality of exposure assessments. Although many reviews conduct a systematic and transparent assessment for the potential for bias, they are often deficient in subsequently integrating across a body of evidence. A cohesive review considers the impact of the direction and magnitude of
potential biases on the results, systematically evaluates important scientific issues such as study sensitivity and effect modifiers, identifies how different studies complement each other, and assesses other potential sources of heterogeneity. Given these challenges of conducting informative systematic reviews of observational studies, we provide a series of specific recommendations based on practical examples for cohesive evidence integration to reach an overall conclusion on a body of evidence to better support policy making in public health.

Introduction

When integrating scientific evidence to reach environmental and occupational health-related conclusions, the systematic review is a powerful tool for drawing causal inference for evidence-based decision-making. Systematic reviews and evidence integration have broad and important policy implications, and are utilized by governmental agencies and authoritative bodies to identify and evaluate environmental and occupational hazards (e.g., to set acceptable levels of exposure to chemicals in consumer products and the environment), develop evidence-based guidelines (e.g., to protect the health of the general public and workers), and are used by many national and international agencies to inform critical public health decisions (e.g., [1–3]).

Based on an evolving understanding of what constitutes an informative review, the scientific community has developed multiple systematic review methods over the past decade, leading to a proliferation of published systematic reviews and meta-analyses. However, not all reviews are informative and transparent because of flaws in methodology or inadequate integration of the evidence, which compromise the acceptability and use of the review’s conclusions. The primary methods in this area were developed to conduct reviews of experimental evidence in clinical epidemiology, often for the purpose of producing guidelines (e.g., GRADE-Grading of recommendations assessment, development, and evaluation [4], and Cochrane Review [5]). Many of these methodologies have subsequently been applied to the review of observational studies, but have not been adequately adapted in evaluating potential biases and evidence integration, especially in environmental and occupational epidemiology studies. Moreover, consideration of the impact of the biases on the results and evidence integration is often superficial at best. As a result, the quality and utility of published systematic reviews vary considerably.

There is a crucial need for systematic review methods adapted to assess observational studies in environmental and occupational health. As a group of epidemiologists specializing in the conduct of systematic reviews and meta-analyses of environmental and occupational health, we offer a discussion of the methodological challenges faced when creating cohesive systematic reviews. We provide recommendations on ways to address these challenges, supported by examples from the published literature, with a focus on the systematic integration of evidence across studies. We also reflect on automation, an emerging trend in systematic reviews with the potential to impact the review process.
**Discussion**

### Develop the systematic review framework

Steps common to well-conducted systematic reviews are shown in Figure 1. The initial step identifies the research question and develops the systematic review framework and methods. This consists of a series of iterative activities including scoping and problem formulation, systematic literature searches, developing the framework for the type of studies and evidence included in the review (i.e., the Population, Exposure, Comparator, and Outcome, referred to as a “PECO” statement [6]), and protocol development, which identifies important issues and concerns specific to the topic under study, defines clear and specific inclusion/exclusion criteria, and provides guidelines for evaluating studies and evidence integration. Articulating key scientific issues, which are defined below, early in the process is needed to evaluate potential sources of selection bias, measurement error of exposures and outcomes, key potential confounders (critical risk-of-bias domains, Box 1), and study sensitivity (i.e., the ability of the study to detect a true effect [7], Box 1; Figure 1) and should be systematically evaluated during evidence integration. Examples of key scientific issues include definitions of the exposure of interest and exposure proxy used to measure that exposure, effect modifiers to be considered, a relevant biological understanding of the exposure-outcome relationship that inform the causal determination (Figure 1), and other factors that may explain heterogeneity across studies. All steps of the review process should be clearly and transparently documented (8). A team of reviewers with combined methodological and subject matter expertise in the exposure and outcome of interest is essential during protocol development to produce a meaningful systematic review.

### Evaluate individual studies

The evaluation of study quality (Box 1) includes an assessment of internal validity (i.e., “risk-of-bias”) and can also include an evaluation of study sensitivity (Box 1), which is the ability of the study to detect the true effect (7). Study sensitivity (also called informativeness [1]) includes evaluating whether the size of the exposed population is adequate to provide precise estimates of an effect, adequate length of follow-up to allow sufficient induction time, and whether the level, duration, and time window of exposure in the population at risk is sufficient or adequate to detect an effect of exposure. Several systematic review methods propose a “domain approach”, using a series of questions to arrive at a risk-of-bias judgment for each specific type of potential bias. These include selection bias, measurement error, or information bias (i.e., outcome and exposure misclassification), and potential confounding, although the terminology and grouping of questions related to these bias domains may vary (Figure 1, Box 1). The impact of potential biases or confounders should consider the magnitude and direction (i.e., towards or away from the null) of the bias on the effect estimate. When information is available, a formal, quantitative bias analysis would be desirable (9).

### Establish the framework for assessing potential bias

There are several approaches for establishing a framework for evaluating potential bias in systematic reviews of observational studies. The first defines the risk-of-bias domain ratings by comparing the current study to an ‘ideal’ experimental study (i.e., a randomized
controlled trial (RCT) (5, 10). The second defines the risk-of-bias by comparing the current study to characteristics of the ‘ideal’ observational study, one with optimal selection methods and well-designed to measure exposure and outcome, that can address temporality, confounders, and sensitivity, and produce results undistorted by bias (e.g., U.S. Environmental Protection Agency’s Integrated Risk Information System’s assessments (3), U.S. National Toxicology Program’s (NTP) Report on Carcinogens Handbook) (2). The third evaluates potential for bias by developing precise criteria for each risk-of-bias domain rating (Box 1) relevant to the exposure-outcome relationship under study, as conducted in a systematic review by the NTP Office of Health Assessment of traffic-related air pollution and hypertensive disorders of pregnancy (11). An experimental study design is neither feasible nor ethical when examining environmental and occupational exposures in humans; moreover, risk-of-bias comparison to an experimental study inhibits the ability to differentiate the quality between well done and poorly conducted observational studies. Therefore, we postulate the latter two approaches described above are the most meaningful and informative, offering greater transparency for differentiating the quality of different studies and providing more accurate answers to policy questions of environmental and occupational hazards of concern for public health.

Evaluate information bias: exposure misclassification

All bias domains are essential in any robust review. In general, systematic review methods for observational studies adapted from clinical reviews have developed adequate methods to assess confounding, selection bias, and information bias related to outcome measurement in a comprehensive way. Here we focus on information bias related to exposure measurement, as clinical systematic review tools are typically inadequate to evaluate environmental and occupational exposure studies. Measurement error can lead to misclassification of exposure status. Exposure classification in RCTs is generally well-characterized, easily measurable, and administered in a controlled environment with pre-defined categories. However, when assessing observational studies, particularly the complex, real-world exposures in environmental and occupational studies, the challenge is to develop methods to accurately measure or assess exposure and classify subjects by exposure level or group. A body of literature of environmental and occupational studies evaluating an exposure-outcome relationship may include observational studies which assess the same exposure through an array of different methods such as biomonitoring, personal or environmental monitoring, statistical modeling, environmental sampling, job-exposure matrices, or questionnaires. These studies use both direct and proxy measures to assess the exposure of interest, which may be continuous or categorical, and the methods used to measure exposure (or proxy for exposure) can vary across studies. For example, in a review on environmental exposure to nitrogen dioxide and lung cancer, reviewers considered studies with exposure proxies that included environmental sampling and data modeling (12). An experienced reviewer must determine if the exposure metric used is an acceptable proxy for the true exposure of interest within the relevant population when assessing potential bias. Subsequently, the reviewer must also determine the level of confidence in the exposure or proxy metric, how accurately it was measured and classified, and whether this imposes incorrect assumptions of homogeneity within exposure groups (e.g., ever vs. never exposed, high vs. low exposure).
Considerable scientific expertise in exposure science or industrial hygiene is needed to evaluate the potential for exposure misclassification.

A common source of bias is non-differential exposure misclassification (Box 1) with respect to case status that frequently (although not always) biases findings towards the null (13), underestimating and diluting any real effects of the exposure on the outcome. Furthermore, the sensitivity of study results (Box 1) to answer the research question in a systematic review may be lower due to poor exposure contrast in particular study populations (7), and should be adequately explored.

An example of a review with a challenging exposure assessment and the high potential for measurement error is the NTP’s evaluation of night shift work and cancer. The true exposure of interest, “working conditions that would result in circadian disruption”, was not directly measured in any available studies. Rather, the most relevant proxies for night shift work were well-defined exposure variables related to work during specific hours at night. A complete lifetime work history, including shift work, allowed for the calculation of other relevant proxy metrics such as the average number of nights worked in a given period, cumulative frequency of night shifts worked, years of night shift work, and age at starting night work. Unlikely shift work exposure among the comparison groups, as well as low risk of differential recall bias also increased confidence in these proxy measures of exposure and reduced the likelihood of exposure misclassification (14).

Biomarker measures are often considered good proxies for the true exposure of interest; however, the potential for exposure misclassification does exist. For example, in a systematic review of health effects of phthalate exposure (15), the majority of evaluated studies measured phthalate metabolite biomarkers as a proxy of exposure. The authors carefully considered multiple criteria to determine their confidence in the exposure proxy, as outlined in their review protocol. They found evidence that repeatability of these measures over time varied by type of phthalate (i.e., long-chain vs. short-chain) and that urine was the preferred matrix for measurement as there was potential for metabolism of contaminants in other tissues (16). Therefore, studies with non-urine biomarker exposures were deemed of low quality in the review.

**Consult with scientific experts and be cautious of checklist or algorithmic approaches**

Checklist tools commonly emphasize the mechanics of the review process. Although checklists support the standardization of methodology across reviews, in our view, they insufficiently emphasize the underlying science. Their use results in a more algorithmic approach to reviews, lacking focus and internal validity, which may not be informative for answering the research question (17, 18). As a consequence, algorithm-based reviews do not meaningfully add to the literature base or inform decision-making (19). Many algorithmic approaches summarize the qualitative aspects of a systematic review into a quantitative score, assigning predefined weights to checklist items, which are difficult to justify and do not account for differences in the impact of a given bias across different studies (5, 18). Furthermore, the validity and reliability of these scoring tools have not been adequately tested (20). A study might ‘score’ poorly in one or more bias domains, but provide key evidence in others, lending valuable support to a review conclusion, while in an algorithmic-
based approach it may have been disregarded. In a systematic review of air pollution and diabetes, Eze et al. (21) found value in study designs typically given a ‘low’ score (e.g., ecologic and cross-sectional) as these studies included evidence on confounding and exposure that the authors found valuable in assessing the relationship.

An example of a checklist approach is the Newcastle-Ottawa Scale (22), one of many numeric scales used to judge risk-of-bias in observational studies. This scale has been criticized in recent years for being unable to distinguish between low- and moderately-biased studies, leading to questionable validity of the conclusions of systematic reviews and meta-analyses (23). Despite these flaws, this scale continues to be widely used in published reviews (24). A recent systematic review and meta-analysis on indoor formaldehyde exposure and asthma utilized the Newcastle-Ottawa Scale for case-control and cohort studies (25). In this study, no information was provided by the authors as to why each study scored the way it did, leading to a lack of transparency in their risk-of-bias findings. In addition, the review also used the Joanna Briggs Institute checklist for cross-sectional studies (26). This basic tool combines an inadequate set of bias domain questions with reporting requirements to ultimately screen studies for inclusion. Combined, poor risk-of-bias reviews can lead to confusion and may produce results that cannot be substantiated by the wider scientific community.

Study quality guidelines and risk-of-bias checklists with overly-rigid criteria, even when not associated with a quantitative score, are also problematic. For example, when evaluating potential selection bias, it may be tempting to include a minimum participation level. However, selection bias would be present and important only if the decision to participate was related to both the exposure and outcome (27), and lower participation could still result in minimal bias.

There are many tools available for assessing risk-of-bias in observational studies, but little consensus on optimal tools or key elements to include when assessing quality. Reviewers who utilize a tool or checklist should be aware of these limitations and choose one best suited to address the exposure-outcome relationship under investigation. Preferably, an expert reviewer would carefully evaluate each study for the potential for bias, and if possible, the magnitude and direction of the bias, using transparent guidelines outlined in the protocol and considering complementary evidence. A team of reviewers with scientific, methodologic, and subject matter expertise (e.g., exposure scientists, occupational hygienists, biologists, medical specialists) are needed to interpret levels of validity and accuracy within and across studies, creating a more comprehensive and informative review.

**Integrate evidence and reach conclusions**

**Conduct an in-depth and cohesive analysis of the evidence across studies**

The strengths of many systematic reviews include defining a clear research question, conducting a comprehensive literature search, transparency in methods to evaluate potential biases, and providing a summary risk estimate in the case of meta-analysis; however, reviews often lack an in-depth and cohesive integration of the whole body of evidence. Some reviews only use one of many results reported in the study. The same rigor that is used to
evaluate risk of bias in individual studies should be used to systematically evaluate sensitivity, effect modification and other scientific issues across studies. Underlying study databases in environmental and occupational health research are often very heterogeneous. This heterogeneity is not always explored in detail, but can inform causal judgments, and allow exploration of important topics in vulnerable populations for targeted policy action. Studies with limitations or biases may provide key information for causal inference and can be instrumental in examining and ruling out alternative hypotheses, contributing to a coherent network of evidence (28). For example, concerns about confounding among individual studies may be minimized or ruled out if consistent results are seen across different populations, study designs, and exposure settings. Likewise, studies lacking statistical significance may be key to revealing patterns in risk estimates across a group of studies. Other factors that contribute to the strength of the evidence include consistency (considering sources of heterogeneity), the magnitude of an effect, and the presence of an exposure-response relationship. Algorithmic approaches frequently discard evidence from individual studies deemed flawed or biased. This reductionist approach does not allow for the integration across studies, where supporting evidence, even from studies with some methodological deficiencies, can add to an integrated network of evidence.

We highlight two qualitative reviews (one of which also had a related meta-analysis) and two meta-analyses as examples of systematic reviews that address heterogeneity, evaluate key issues, and integrate evidence across studies in insightful ways.

An example of a systematic review which provided a cohesive analysis that was informed by problem formulation is the International Agency for Research on Cancer (IARC) review, which newly classified welding fumes as ‘carcinogenic to humans’ in 2017. During problem formulation, key issues were identified which directed the systematic review process. These included a clear definition of exposure (i.e., exposure to welding fumes as measured by proxy of occupation as a welder or directly using equipment; furthermore, studies with broadly defined occupational categories for which cancer could not be attributed to welding were excluded), the quality of the exposure assessment and factors that could affect exposure (i.e., type of material welded and welding process), potential confounders, and focusing on cancer outcomes with a literature database that was adequate to conduct an assessment (Box 1) (29, 30). The evidence was carefully interpreted by an expert panel, and the direction and magnitude of biases were systematically considered for the assessment of individual studies. During the evidence integration step, the key issues (e.g., type of welding material, welding process, studies controlling for potential confounders) were systematically evaluated and the association between lung cancer exposure remained robust. The IARC Working Group subsequently published a meta-analysis following their systematic review (29). The association between welding fumes and lung cancer remained robust after an in-depth exploration of factors that contributed to heterogeneity in this body of literature (e.g., type of welding material, welding process, studies controlling for potential confounders). The rigor and informativeness of the IARC review led to evidence-based policy changes in less than a year since its publication: the Health and Safety Executive of Britain has acted to strengthen its enforcement expectations for fume control for welding activities (31, 32).
The NTP cancer hazard evaluation of trichloroethylene (NTP [33]) found that study utility (Box 1) and estimated exposure levels explained the heterogeneity seen across studies. Specifically, an increased risk of kidney cancer was observed in most moderate- and high-quality studies and in studies with the highest estimated exposure to trichloroethylene; the latter emphasizing the importance of identifying studies with higher sensitivity to detect an effect and likely reflecting some underlying biological mechanisms of toxicity of this particular exposure.

A meta-analysis by Huss et al. (34) found that heterogeneity was explained by quality of the exposure assessment; a positive association between occupational exposure to extremely low-frequency magnetic fields and risk of amyotrophic lateral sclerosis was observed among studies assessing exposure using complete occupational histories, but not among those assessing exposure via self-report or from death certificates. This example underscores the importance of considering the direction of bias when integrating study results and illustrates the tendency of non-differential exposure misclassification to bias results towards the null, with potential implications for public health decision making. Another meta-analysis exemplifies the importance of understanding the biology of an exposure-outcome relationship. Steenland et al. (35) conducted a bias sensitivity analysis and found that an inverse association between serum perfluorooctanoic acid (PFOA) and birthweight was limited to studies measuring PFOA in samples collected later in pregnancy, which may be susceptible to confounding, and little to no association in studies measuring PFOA in early pregnancy. Low glomerular filtration occurs late in pregnancy and is related to both increased serum PFOA and decreased birthweight (36). Thus, studies of PFOA samples taken during late pregnancy may lead to a biased overestimation of the true effect. Previous meta-analyses, which found an inverse association between PFOA and birthweight, had not considered the timing of exposure (37, 38). They also excluded a large, relatively null study of a population highly exposed to PFOA because it used a modeled PFOA exposure based on estimates before pregnancy (which would not be susceptible to potential confounding) rather than measured data and would have high sensitivity to detect an effect if it existed. The differences in how the same studies were considered across multiple meta-analyses demonstrate the importance of study sensitivity in evaluating study quality, and argues against using rigid criteria in assessing risk-of-bias, and the exclusion of potentially informative individual studies.

**Explore the use of triangulation in evidence integration**

One powerful approach that considers the overall literature base and requires a more inclusive view of evidence synthesis is triangulation, which integrates data from different methods, designs, and theoretical approaches, as well as data with different and unrelated sources of potential bias, to determine if findings converge on one conclusion (28, 39). Studies are not automatically downgraded due to the presence of bias; rather, the risk-of-bias in each study is recognized, and the potential effect the bias would have on a risk estimate (i.e., the direction and magnitude of bias) is identified. If a review recognizes different sources of biases across studies and data results are consistent given the possible biases, then triangulation can assist the reviewer in reaching a more certain conclusion. For example, in a review of occupational exposure to benzene and lymphoma (40), studies were stratified by
study quality (based on the quality of the exposure and outcome measurements in each study). Meta-analyses then calculated the relative risks stratified by three dimensions of study quality identified that could bias the effect estimates: year of start of follow-up, the strength of the exposure-outcome association, and the quality of the exposure assessment. Combined relative risks were then estimated within each stratum. The authors used these domains to triangulate their conclusions, ultimately concluding there was an overall association. Triangulation is an inclusive approach that recognizes that many different types of studies, even those with potential biases, contribute to the evidence integration and is preferable to more algorithmic approaches, or to systematic review approaches that exclude studies or downgrade evidence upfront on criteria such as study design.

**Avoid assigning higher bias ratings based on study design**

During evidence integration, systematic review tools adapted from clinical epidemiology often downgrade or exclude the quality of evidence from observational designs (4). Depending on the exposure-outcome relationship under investigation, a cross-sectional, case-control, or cohort study may be the most - if not the only - practical and feasible design to provide the highest quality evidence (41). Moreover, a single systematic review will likely evaluate multiple observational study designs. Downgrading evidence based solely on study design limits the full potential and unique strengths of observational evidence in causal analysis and in resolving critical public health gaps.

The consequences of automatically downgrading observational studies have been recognized and some efforts to correct this limitation have been made (42). In a systematic review of epidemiology studies of environmental noise and annoyance conducted to inform the current World Health Organization’s Environmental Noise Guidelines for the European Region (43), scientists adapted GRADE to identify study designs best able to capture the exposure-outcome relationship between sleep disturbance and noise annoyance. In this context, cross-sectional studies were assigned a high evidence ranking, as the impact of noise annoyance on sleep disturbance was considered an immediate effect and reverse causality was unlikely (43). Ecologic studies have proven informative to a review in exceptional circumstances. For example, when investigating arsenic in drinking water (44), variations of arsenic by time periods and regions allowed investigators to use ecologic data as a natural experiment to examine health effects (45). Likewise, new statistical methods such as the difference-in-difference method have been applied in ecologic studies to model the variability of exposure and health effects over time (46). The key scientific issues characterizing the particular exposure-outcome relationship under investigation should dictate the strengths and limitations of each study, independent of the study’s design, and should inform the utility of that study during evidence integration.

**Conduct transparent meta-analysis, when data permit, in addition to the qualitative evaluation**

When properly conducted, the meta-analysis is a valuable tool to quantify an exposure-outcome relationship explored in a systematic review. We offer comments on the role they play in hazard evaluation, considerations on when to conduct them, and how they should be
conducted. A meta-analysis should be conducted in the context of the broader qualitative evaluation and not as the sole determinant for decision-making.

In observational studies, particularly studies of environmental and occupational health, varying study designs, exposure measurements or levels, and populations make combining data difficult, and sometimes even inappropriate, as the exposure proxy, level, duration, and method of assessment can vary greatly across studies. When this occurs, combining data to produce a meta-analysis would not lead to an informative result. Night shift work and breast cancer is an example where a meta-analysis may not be considered informative for evaluating potential causality because of the significant heterogeneity in the definition of night shift work (the exposure) across study populations (14). It is important to note that while a summary estimate produced by a meta-analysis may be inappropriate, a forest-plot or other visual representations of the data (which can be stratified by key issues or variables) can be useful when integrating evidence in a review, as was done in the NTP night shift work (e.g., type of breast cancer, age starting night shift work) and trichloroethylene reviews (e.g., estimated exposure level).

Meta-analyses, ideally conducted in the context of a qualitative review, such as the IARC Monographs evaluation of benzene and cancer (47), should present in detail the methods for combining exposure assessments, statistical analyses, and data manipulation, so that the overall process is transparent and useful for evaluating sources of heterogeneity. A lack of transparency in how and why meta-analyses are conducted leads to confusion in the literature, especially when differing conclusions are reached. For example, one meta-analysis of chromium VI exposure found an increased risk of stomach cancer (48), while another did not (49). Differences in study selection and selection of available exposure measures may have led to these differences, as discussed by Welling et al. (48).

The increased proliferation in a very short time of published meta-analyses has resulted in analyses with incorrectly extracted data, effect estimates from multiple studies within overlapping participant population, or with studies that do not evaluate the endpoint of interest. An example is a recent retraction of a meta-analysis on night shift work and cancer in women (retraction [50]). Thus, there is a need for increased rigor by systematic review authors (a team including both subject matter experts and methodologists), and a need for quality assurance in the peer review and editorial stages to prevent the perpetuation of incorrect findings.

**Reporting the systematic review findings in a transparent and comprehensive manner**

Data presentation in systematic reviews is an important factor influencing the utility of the conclusions. Many scientific journals require a standardized reporting checklist, aimed at increasing the transparency and scientific quality of the review (e.g., the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) [51] and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement [52]), which also encourage the publication and registration of a systematic review protocol where the outcomes of interest are stated upfront. These statements are sometimes erroneously used as checklists for performing systematic reviews and meta-analyses, but are valuable tools offering guidance on reporting. Recent reviews have included more visualizations and less
text, providing an advantage in readability and in meeting journal length restrictions; however, some narrative study description is necessary to fully describe a complex study database, be it in the main article or provided as supplementary text. We recommend systematic reviews of environmental and occupational epidemiology clearly present the following crucial information (some of which are standard in reporting guidelines, e.g. PRISMA and STROBE):

- A discussion of the key scientific issues (as defined above) for the specific exposure and outcome relationship;
- Transparent documentation of the literature search;
- Clear and specific inclusion and exclusion criteria;
- An overview of study characteristics (e.g., population characteristics, exposure assessment methods, outcomes) included in the review, even if not included in the evidence integration for bias, quality, or other reasons;
- A discussion of biases and limitations for each bias domain across studies, in addition to the rationale for the risk-of-bias at a study level;
- A scientific narrative of the interpretation of study findings including a discussion of heterogeneity across studies (not limited to potential for biases) and the rationale for the conclusion (e.g., consideration of dose-response relationships, consistency, ruling out chance bias and confounding);
- When possible, include a discussion on whether the epidemiologic findings of the review are consistent with the currently understood mechanisms of action or biological pathways involved based on a review or comprehensive analysis of mechanistic data, if available and relevant.

**Emerging trend - Automation**

Systematic reviews are time-consuming and logistically intensive. Ways to increase efficiency without compromising scientific rigor should be explored. Machine learning algorithms (i.e., automation) are a potentially powerful tool, particularly in reducing the amount of time spent screening literature for topics with large study databases (53). However, this technology should be used with caution and not as a replacement for subject matter expertise. Human oversight should be employed for quality assurance at every step. Using automation for data extraction and risk-of-bias evaluations has been proposed for clinical trial evaluations (54). Automated data extraction may be challenging and would require validation by a scientist. Because of the complexity, diverse study designs, and characteristics of observational studies, considerable scientific judgment is required to assess study validity and integrate evidence to reach sound conclusions. Additionally, information from related publications is frequently needed to evaluate risk of bias. Similar to the checklist and algorithmic approaches discussed above, automation should be avoided for risk-of-bias assessments and evidence integration, as it will most likely result in poorly conducted and incorrect assessments. A robust discussion of the positive and negative aspects of automation at each step of the systematic review process is imperative to ensure it
is used as a tool that assists experts, reduces human error and time, but without sacrificing quality and usefulness of the final product.

**Current application and conclusions**

Several U.S. and international agencies specializing in systematic reviews for hazard assessment or policy development have published, or are planning to publish approaches to better evaluate the unique aspects of observational studies of environmental and occupational health (55), using many of the methods described above. Some have modified the GRADE approach (e.g., 6, 56, 57), while others have independently developed structured expert reviews with guidelines that incorporate evaluation of potential bias and study sensitivity into a structured evidence integration process (e.g., 1–3).

Systematic reviews acknowledging the unique strengths of observational studies in identifying the hazards and health consequences of environmental and occupational exposures are crucial for evidence-based decision-making and in identifying causal exposure-outcome relationships. It is important to note that the degree of structure can vary and still produce a high-quality systematic review. The transparency of the systematic review process is a significant advantage and best practices should be followed when evaluating environmental and occupational studies. The rigor and transparency often applied to risk of bias should be applied to all steps of the systematic review by a team of reviewers with scientific, methodologic, and subject matter expertise. These steps include problem formulation, identification of key issues, the scientific narrative, the impact of study limitations on the study’s informativeness, and the cohesive integration of evidence across studies. These critical components are necessary to fully realize the potential of systematic reviews in environmental and occupational health.

**ACKNOWLEDGMENTS:**

The authors would like to thank the following people for their thoughtful reviews and comments on this manuscript, which have improved it greatly: Tara Hartley, Kathleen MacMahon, Robert Daniels, Kris Thayer, Tom Luben, Craig Steinmaus, Andrew Rooney, Christina Parks, and Abee Boyles.

**REFERENCES**

1. IARC. IARC Monographs on the Identification of Carcinogenic Hazards to Humans: Preamble. Lyon, France: International Agency for Research on Cancer; 2019.
2. NTP. Handbook for Preparing the Report on Carcinogens Monographs. Research Triangle Park, NC: National Toxicology Program; 2015.
3. Radke EG, Glenn B, Galizia A, Persad A, Nachman R, Bateson T, et al. Development of outcome-specific criteria for study evaluation in systematic reviews of epidemiology studies. Environ Int. 2019;130:104884. [PubMed: 31299560]
4. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2011;64(4):380–2. [PubMed: 21185693]
5. Higgins JP, Altman DG, Gøtzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928. [PubMed: 22008217]
6. Woodruff TJ, Sutton P. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environ Health Perspect. 2014;122(10):1007–14. [PubMed: 24968373]

7. Cooper GS, Lunn RM, Ågerstrand M, Glenn BS, Kraft AD, Luke AM, et al. Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures. Environ Int. 2016;92–93:605–10.

8. Howard J, Piacentino J, MacMahon K, Schulte P. Using systematic review in occupational safety and health. Am J Ind Med. 2017;60(11):921–9. [PubMed: 28944489]

9. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014;43(6):1969–85. [PubMed: 25080530]

10. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. [PubMed: 27733354]

11. NTP. NTP Monograph on the Systematic Review of Traffic-related Air Pollution and Hypertensive Disorders of Pregnancy. Research Triangle Park, NC: National Toxicology Program; 2019 Report No.: NTP Monograph 7.

12. Hamra GB, Laden F, Cohen AJ, Raaschou-Nielsen O, Brauer M, Loomis D. Lung cancer and exposure to nitrogen dioxide and traffic: A systematic review and meta-analysis. Environ Health Perspect. 2015;123(11):1107–12. [PubMed: 25870974]

13. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med. 1998;55(10):651–6. [PubMed: 9930084]

14. NTP. Draft Report on Carcinogens Monograph on Night Shift Work and Light at Night. Peer Review Draft. Research Triangle Park, NC: National Toxicology Program; 2018.

15. Radke E, Glenn B, Braun J, Cooper G. Phthalate exposure and female reproductive and developmental outcomes: a systematic review of the human epidemiological evidence. Environ Int 2019;130:104580. [PubMed: 31351310]

16. Johns LE, Cooper GS, Galizia A, Meeker JD. Exposure assessment issues in epidemiology studies of phthalates. Environ Int. 2015;85:27–39. [PubMed: 26313703]

17. Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol. 2006;59(12):1249–56. [PubMed: 17098567]

18. Savitz DA, Wellenius GA, Trikalinos TA. The problem with mechanistic risk of bias assessments in evidence synthesis of observational studies and a practical alternative: Assess the impact of specific sources of potential bias. Am J Epidemiol. 2019;188(9):1581–5. [PubMed: 31145434]

19. Ioannidis JPA. Meta-analyses in environmental and occupational health. Occup Environ Med. 2018;75(6):443–5. [PubMed: 29574405]

20. National Research Council. Review of EPA’s Integrated Risk Information System (IRIS) Process. Washington, D.C.: National Academies Press; 2014.

21. Eze IC, Hemkens LG, Bucher HC, Hoffmann B, Schindler C, Künzli N, et al. Association between ambient air pollution and diabetes mellitus in Europe and North America: systematic review and meta-analysis. Environ Health Perspect. 2015;123(5):381–9. [PubMed: 25625876]

22. Wells GA, Shea B, O’Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses Ottawa, Canada: The Ottawa Hospital; 2009 [Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

23. Stang A Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol. 2010;25(9):603–5. [PubMed: 20652370]

24. Stang A, Jonas S, Poole C. Case study in major quotation errors: a critical commentary on the Newcastle-Ottawa scale. Eur J Epidemiol. 2018;33(11):1025–31. [PubMed: 30259221]

25. Yu L, Wang B, Cheng M, Yang M, Gan S, Fan L, et al. Association between indoor formaldehyde exposure and asthma: A systematic review and meta-analysis of observational studies. Indoor Air. 2020(ePub Ahead of Print).
26. Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, et al. Chapter 7: Systematic reviews of etiology and risk In: Aromataris E, Munn Z, editors. Joanna Briggs Institute Reviewer’s Manual: The Joanna Briggs Institute; 2017.

27. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615–25. [PubMed: 15308962]

28. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. Int J Epidemiol. 2016;45(6):1776–86. [PubMed: 26800751]

29. Honaryar MK, Lunn RM, Luce D, Ahrens W, ’t Mannetje A, Hansen J, et al. Welding fumes and lung cancer: a meta-analysis of case-control and cohort studies. Occup Environ Med 2019;76(6):422–31. [PubMed: 30948521]

30. IARC. Welding, Molybdenum Trioxide, and Indium Tin Oxide. Lyon, France: International Agency for Research on Cancer; 2018 320 pp.

31. Cherrie JW, Levy L. Managing occupational exposure to welding fume: New evidence suggests a more precautionary approach is needed. Ann Work Expo Health. 2020;64(1):1–4. [PubMed: 31686108]

32. HSE. Change in Enforcement Expectations for Mild Steel Welding Fume: Health and Safety Executive, UK; 2019 [updated Issued on: February 2019 Available from: https://www.hse.gov.uk/safetybulletins/mild-steel-welding-fume.htm.

33. NTP. Report on Carcinogens, 14th Edition Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service, National Toxicology Program; 2016.

34. Huss A, Peters S, Vermeulen R. Occupational exposure to extremely low-frequency magnetic fields and the risk of ALS: A systematic review and meta-analysis. Bioelectromagnetics. 2018;39:156–63. [PubMed: 29350413]

35. Steenland K, Barry V, Savitz D. Serum perfluorooctanoic acid and birthweight: An updated meta-analysis with bias analysis. Epidemiology. 2018;29(6):765–76. [PubMed: 30063543]

36. Verner MA, Loccisano AE, Morken NH, Yoon M, Wu H, McDougall R, et al. Associations of perfluoroalkyl substances (PFAS) with lower birth weight: An evaluation of potential confounding by glomerular filtration rate using a physiologically based pharmacokinetic model (PBPK). Environ Health Perspect. 2015;123(12):1317–24. [PubMed: 26008903]

37. Johnson PI, Sutton P, Atchley DS, Kousta E, Lam J, Sen S, et al. The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth. Environ Health Perspect. 2014;122(10):1028–39. [PubMed: 24968388]

38. Lam J, Kousta E, Sutton P, Johnson PI, Atchley DS, Sen S, et al. The Navigation Guide - evidence-based medicine meets environmental health: integration of animal and human evidence for PFOA effects on fetal growth. Environ Health Perspect. 2014;122(10):1040–51. [PubMed: 24968389]

39. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. Int J Epidemiol. 2016;45(6):1866–86. [PubMed: 28108528]

40. Vlaanderen J, Lan Q, Kromhout H, Rothman N, Vermeulen R. Occupational benzene exposure and the risk of lymphoma subtypes: a meta-analysis of cohort studies incorporating three study quality dimensions. Environ Health Perspect. 2011;119(2):159–67. [PubMed: 20880796]

41. Checkoway H, Pearce N, Kriebel D. Selecting appropriate study designs to address specific research questions in occupational epidemiology. Occup Environ Med. 2007;64(9):633–8. [PubMed: 17704203]

42. Morgan RL, Thayer KA, Santesso N, Holloway AC, Blain R, Eftim SE, et al. A risk of bias instrument for non-randomized studies of exposures: A users’ guide to its application in the context of GRADE. Environ Int. 2019;122:168–84. [PubMed: 30473382]

43. WHO. Environmental Noise Guidelines for the European Region. Copenhagen, Denmark: World Health Organization; 2018.

44. National Research Council. Critical Aspects of EPA’s IRIS Assessment of Inorganic Arsenic: Interim Report. Washington, DC: National Academies Press; 2013.
45. Marshall G, Ferreccio C, Yuan Y, Bates MN, Steinmaus C, Selvin S, et al. Fifty-year study of lung and bladder cancer mortality in Chile related to arsenic in drinking water. J Natl Cancer Inst. 2007;99(12):920–8. [PubMed: 17565158]

46. Leogrande S, Alessandrini ER, Stafoggia M, Morabito A, Nocioni A, Ancona C, et al. Industrial air pollution and mortality in the Taranto area, Southern Italy: A difference-in-differences approach. Environ Int. 2019;132:105030. [PubMed: 31398654]

47. IARC. Benzene. Lyon, France: International Agency for Research on Cancer; 2018 309 pp. p.

48. Welling R, Beaumont JJ, Petersen SJ, Alexeeff GV, Steinmaus C. Chromium VI and stomach cancer: a meta-analysis of the current epidemiological evidence. Occup Environ Med. 2015;72(2):151–9. [PubMed: 25231674]

49. Gatto NM, Kelsh MA, Mai DH, Suh M, Proctor DM. Occupational exposure to hexavalent chromium and cancers of the gastrointestinal tract: a meta-analysis. Cancer Epidemiol. 2010;34(4):388–99. [PubMed: 20430714]

50. Yuan X, Zhu C, Wang M, Mo F, Du W, Ma X. Retraction: Night shift work increases the risks of multiple primary cancers in women: A systematic review and meta-analysis of 61 articles. Cancer Epidemiol Biomarkers Prev. 2019;28(2):423. [PubMed: 30718240]

51. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535. [PubMed: 19622551]

52. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344–9. [PubMed: 18313558]

53. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev. 2014;3:74. [PubMed: 25005128]

54. Marshall JJ, Kuiper J, Wallace BC. Automating risk of bias assessment for clinical trials. IEEE J Biomed Health Inform. 2015;19(4):1406–12. [PubMed: 25966488]

55. Rooney AA, Cooper GS, Jahnke GD, Lam J, Morgan RL, Boyles AL, et al. How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. Environ Int. 2016;92–93:617–29.

56. Hempel S, Xenakis L, Danz M. Systematic Reviews for Occupational Safety and Health Questions: Resources for Evidence Analysis. Santa Monica, CA: RAND Corporation; 2016 102 p.

57. Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. Systematic review and evidence integration for literature-based environmental health science assessments. Environ Health Perspect. 2014;122(7):711–8. [PubMed: 24755067]
**Box 1. Definitions of terms**

| Risk-of-bias | Involves the internal validity of a study and reflects study-design characteristics that can introduce a systematic error or deviation from a true effect that might affect the magnitude and even the direction of the apparent effect. Risk-of-bias domains: Selection bias, information bias (exposure misclassification or outcome misclassification), confounding, analysis/reporting bias |
|---|---|
| Study Sensitivity or Informativeness | The ability of a study to detect a true effect. An insensitive study will fail to detect a difference that truly exists, leading to a false conclusion of no effect. Only a negative result from a highly sensitive study can be interpreted, with confidence, as evidence of no effect. |
| Study Quality | Involves an investigation of the extent to which study authors conducted their research to high standards—for example, by following a well-documented protocol with trained study staff, and with sufficient power to detect effects. Or in how a study is reported—for example, whether the study population is described sufficiently. |
| Study Utility | The ability of the study to inform the hazard (or risk) evaluation; considers study quality (bias) and study sensitivity. (Some methods define this as informativeness.) |
| Study database | The underlying group of published literature from which studies considered for a review are derived. Also called the literature base. |
| Non-differential misclassification | Occurs when the frequency of errors is approximately the same in the groups being compared (i.e. random). Non-differential misclassification of a dichotomous exposure or outcome metric will often bias results towards the null. |
| Differential misclassification | Occurs when the frequency of errors is greater in one group compared to the other (non-random or systematic error). Direction of bias can be difficult to predict. |

1 Reference 57
2 Reference 1, 7
3 Reference 55
Developing the systematic framework includes several key components such as a scoping and review of the literature (which identifies key scientific issues to address); development of the framework of the literature to be included in the review (such as the PECO or other types of evidence); the systematic literature search strategy (including inclusion and exclusion criteria); development of methods and protocol (which directs the review process, provides transparency for evaluating study quality and in-depth and cohesive analysis and integration of the evidence across studies. Input from subject matter experts is critical at all steps of a systematic review. Many steps in the systematic review process are iterative and can inform one another.

*Not all SR do this at this stage*