A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP6980.

Risk of Bias Assessments and Evidence Syntheses for Observational Epidemiologic Studies of Environmental and Occupational Exposures: Strengths and Limitations

Kyle Steenland,[1] M.K. Schubauer-Berigan,[2] R. Vermeulen,[3] R.M. Lunn,[4] K. Straif,[5,6] S. Zahm,[7] P. Stewart,[8] W.D. Arroyave,[9] S.S. Mehta,[4] and N. Pearce[10]

[1]Rollins School of Public Health, Emory University, Atlanta, Georgia, USA
[2]International Agency for Research on Cancer (IARC), Lyon, France
[3]Institute for Risk Assessment Science, University of Utrecht, Utrecht, Netherlands
[4]Division of the National Toxicology Program (NTP), NIEHS, Research Triangle Park, North Carolina, USA
[5]Global Observatory on Pollution and Health, Boston College, Boston, Massachusetts, USA
[6]ISGlobal, Barcelona, Spain
[7]Shelia Zahm Consulting, Hermon, Maine, USA
[8]Stewart Exposure Assessments, LLC, Arlington, Virginia, USA
[9]Integrated Laboratory Systems, Morrisville, North Carolina, USA
[10]London School of Hygiene and Tropical Medicine, London, UK

Background: Increasingly, risk of bias tools are used to evaluate epidemiologic studies as part of evidence synthesis (evidence integration), often involving meta-analyses. Some of these tools consider hypothetical randomized controlled trials (RCTs) as gold standards.

Methods: We review the strengths and limitations of risk of bias assessments, in particular, for reviews of observational studies of environmental exposures, and we also comment more generally on methods of evidence synthesis.

Results: Although RCTs may provide a useful starting point to think about bias, they do not provide a gold standard for environmental studies. Observational studies should not be considered inherently biased vs. a hypothetical RCT. Rather than a checklist approach when evaluating individual studies using risk of bias tools, we call for identifying and quantifying possible biases, their direction, and their impacts on parameter estimates. As is recognized in many guidelines, evidence synthesis requires a broader approach than simply evaluating risk of bias in individual studies followed by synthesis of studies judged unbiased, or with studies given more weight if judged less biased. It should include the use of classical considerations for judging causality in human studies, as well as triangulation and integration of animal and mechanistic data.

Conclusions: Bias assessments are important in evidence synthesis, but we argue they can and should be improved to address the concerns we raise here. Simplistic, mechanical approaches to risk of bias assessments, which may particularly occur when these tools are used by nonexperts, can result in erroneous conclusions and sometimes may be used to dismiss important evidence. Evidence synthesis requires a broad approach that goes beyond assessing bias in individual human studies and then including a narrow range of human studies judged to be unbiased in evidence synthesis. https://doi.org/10.1289/EHP6980

Introduction

Evidence synthesis (or evidence integration) is widely used to summarize findings of epidemiologic studies of environmental and occupational exposures. Such syntheses are part of systematic reviews of observational epidemiologic study findings.

Systematic reviews are defined by Cochrane guidelines as reviews that "identify, appraise and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a specific research question. They use explicit, systematic methods that are selected with a view aimed at minimizing bias, to produce more reliable findings to inform decision making" (https://www.cochranelibrary.com/about/about-cochrane-reviews). Systematic reviews ideally should include a statement of the goals of the review and a clear description for *a*) determining which studies are relevant to the goals; *b*) how individual studies are evaluated regarding potential biases; and *c*) a method to synthesize evidence across studies (which sometimes includes a meta-analysis). Assessments of biases and their impact play a useful role in both *b*) and *c*). Figure 1 shows a schematic of a systematic review. Boxes 4 and 5 of this figure (evaluate evidence, integrate evidence) depict where risk of bias assessments come into play via evaluations of individual studies and evidence synthesis across studies, and they are the subject of this paper.

Systematic reviews play a similar role today as literature reviews in the past in that both attempt to provide an overview of the literature on a particular topic, either within a discipline (e.g., epidemiology) or across disciplines, and typically assess the evidence for causality for the association between exposure and disease. Systematic reviews are often done in conjunction with a meta-analysis. A meta-analysis yields a quantitative effect estimate, such as the strength of the association between an exposure and an outcome. It also provides an opportunity to explore heterogeneity across studies, e.g., by study design, type of population under study, or other characteristics. Subjectivity (value-based judgment) is inevitably present in the assessments of the quality of the individual studies (including whether they suffer from biases) and in the decisions to include or exclude studies in evidence syntheses and meta-analyses. It is present in the degree to which the authors interpret the reported association to be causal. It is also present in the degree to which the meta-analysis authors account for other evidence that is not considered in the meta-analysis itself, such as studies with exposure effect estimates not compatible with those in the meta-analysis (e.g., prevalence rather than incidence measures), ecological studies, animal data, and mechanistic data. The existence of such subjectivity is generally recognized as inherent to systematic reviews, and the goal is to make such judgments transparent (Whaley et al. 2016; Savitz et al. 2019). There is a tension, however, between the need for expert (necessarily subjective) judgment and consistency and replicability in such reviews.

Address correspondence to Kyle Steenland, Rollins School of Public Health, Emory University, 1518 Clifton Rd., Atlanta, GA 30322 USA. Telephone: (404) 727-0196. Email: nsteenl@emory.edu.

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

Received 26 February 2020; Revised 21 August 2020; Accepted 21 August 2020; Published 14 September 2020.

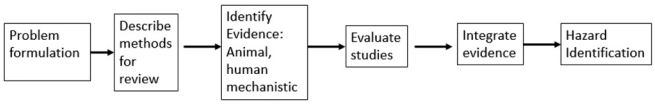

Figure 1. Schematic for systematic review. Adapted from National Research Council (2014).

Risk of bias tools have been developed with the intention of increasing transparency and reducing subjectivity. They are now often used in systematic reviews to evaluate individual studies for bias and to determine which studies should be given more or less weight in evidence synthesis, based on ranking systems to evaluate bias in individual epidemiological studies. Risk of bias tools include ROBINS-I (Sterne et al. 2016), the Newcastle-Ottawa Scale (http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp), the Navigation Guide (Woodruff and Sutton 2014), Office of Health Assessment and Translation (OHAT) (NTP 2019), and a new tool to be used with Grading of Recommendations Assessment, Development and Evaluation (GRADE) (Morgan et al. 2019) (see below) Another tool, ROBINS-E, is under development and not yet available (http://www.bristol.ac.uk/population-health-sciences/centres/cresyda/barr/riskofbias/robins-e/). The risk of bias tool used in the Navigation Guide comes from a combination of methods described by Viswanathan et al. (2008) and Higgins and Green (2011).

GRADE (https://www.gradeworkinggroup.org/) is a method to assess the overall certainty of evidence in a set of studies, developed in the context of making clinical decisions based on human studies. GRADE has advocated risk of bias as one part of this process without, until recently, proposing a specific tool. There has been some discussion about improving the certainty of evidence criteria in GRADE (Norris and Bero 2016).

We recognize that not all risk of bias tools in current use are alike (Losilla et al. 2018; Rooney et al. 2016). Different tools include different bias domains and/or define the same domains differently. Typically, all include consideration of exposure or outcome misclassification/mismeasurement, confounding, and selection bias. Some are accompanied by guidelines for evidence synthesis, and others are not. Furthermore, some partly address the concerns we outline below (see Table 1) for differences between different risk of bias tools.

Assigning actual scores to individual studies based on risk of bias is not done in most of these tools, has been shown to not be effective, and is discouraged in reviews by Jüni et al. (1999) and Stang (2010) and on the Cochrane website [although scoring was recently resurrected in a new systematic review method being implemented for the U.S. Environmental Protection Agency (EPA)'s Toxic Substances Control Act (TSCA) program; see Singla et al. (2019)]. All of the above-cited risk of bias tools evaluate individual studies by level of bias (e.g., low, moderate, serious, and critical) in different domains (e.g., confounding, selection bias, and information bias), and the evaluations may potentially result in exclusion of studies deemed too biased across one or more domains from evidence synthesis. However, they do not consistently assess the direction, magnitude, or overall importance (on the effect estimate) of the various types of bias, and they bring these considerations directly into risk of bias tools. Note also that risk of bias does not mean the actual study is biased. The Navigation Guide (Woodruff and Sutton 2014) and OHAT (NTP 2019) both suggest using the direction of confounding and result of control of confounding to upgrade or downgrade estimates of confounding bias but do not formally build it into a tool. The *Report on Carcinogens Handbook* incorporated direction and magnitude of bias in their guidelines for study quality assessment guidance and evidence integration steps (NTP 2015). Other risk of bias tools also mention this issue but do not tackle it directly. For example, in ROBINS-I, the authors note, "It would be highly desirable to know the magnitude and direction of any potential biases

Table 1. Comparing risk of bias tools.

	Study name				
RoB within individual studies	ROBINS-I[a]	Newcastle Ottawa scale[b]	Morgan (GRADE)[c]	Navigation guide[d]	OHAT[e]
RCT/target experiment as ideal study design	Yes	No	Yes	No	No
Consider direction or magnitude of bias, and importance for effect estimate	Optional, but not formally incorporated into tool	No	Optional[f]	No[g]	Optional, but not formally incorporated into tool
Assign highest domain risk of bias to entire study	Yes	No (but commonly done when used by summing stars/scores across domains)	Yes	No study-level bias summary	No, but used to assign to tiers in study synthesis
Consider statistical methodology as a separate domain	No	No	No	No	Optional
Evidence synthesis					
Rank observational studies as inherently suffering from bias	Not applicable (no formal presentation of evidence synthesis)	Not applicable (no formal presentation of evidence synthesis)	Yes, indirectly because of RCT comparison, but under development	Yes, start at moderate certainty	Yes, start at low to moderate certainty
Possibly reject some studies based on bias	Not applicable (no formal presentation of evidence of synthesis)	Not applicable (no formal presentation of evidence of synthesis)	Yes, although may be allowed in sensitivity analysis	Yes, although may be included in sensitivity analysis	Yes, although may be included in sensitivity analyses

Note: Tools included in this table are risk of bias tools for individual studies with an algorithm-based component. GRADE, Grading of Recommendations Assessment, Development and Evaluation; OHAT, Office of Health Assessment and Translation; RCT, randomized controlled trial; RoB, risk of bias.
[a]Sterne et al. 2016.
[b]http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.
[c]Morgan et al. 2019.
[d]Woodruff and Sutton 2014. The risk of bias tool used in Navigation Guide comes from a combination of methods described by Viswanathan et al. (2008) and Higgins and Green (2011).
[e]NTP 2019.
[f]Direction of bias considered, but not magnitude or eventual impact on effect estimate.
[g]Not mentioned in five published case studies (https://prhe.ucsf.edu/navigation-guide), nor in original paper by Woodruff and Sutton 2014.

identified, but this is considerably more challenging than judging the risk of bias" (Sterne et al. 2016).

Assessing individual study quality is an essential part of systematic review, and risk of bias tools are one way to do this that may increase transparency and replicability in reviews. These tools differ between one another, and we do not here discuss in detail each tool individually but, rather, comment more generally on the limitations of their current use both in the evaluation of individual studies and in evidence synthesis. Although we agree that if risk of bias tools are to be used, they must have a list of domains and some overall evaluation system regarding potential bias, we note that there is no consensus on which domains are to be analyzed and how risk of bias is to be ranked.

In this paper, we first critically review the benefits and pitfalls when using risk of bias assessments for individual studies. We argue, along with other authors (Savitz et al. 2019; Stang 2010; Arroyave et al. 2020), that while the use of risk of bias assessments in evidence synthesis can be a useful tool to improve transparency and limit *a priori* value judgments, they can also potentially be used as a mechanical exercise that leads to erroneous conclusions because the assessments may consider individual studies out of context, may poorly discriminate between studies with minimal and substantial potential bias (i.e., may not evaluate the magnitude and direction of bias and its eventual possible impact on a study's effects estimates), and may have other potential shortcomings as detailed below. Second, we consider broad types of evidence synthesis, such as those proposed by Bradford Hill (Hill 1965) and programs such as the International Agency for Research on Cancer (IARC) Monographs, and then discuss the use of triangulation (Lawlor et al. 2016). Finally, we reflect on some recent evidence syntheses and their risk of bias assessments.

Risk of Bias Assessments for Individual Studies

As previously noted, a risk of bias assessment provides a formal mechanism to systematically evaluate study quality regarding potential biases using the same approach across all studies and hence can add to transparency in systematic reviews. Here, we discuss risk of bias assessments in more detail and also make some recommendations to improve them (Table 2).

Randomized controlled trials as the ideal when assessing bias vs. observational studies. Some currently available risk of bias tools propose using a hypothetical randomized controlled trial (RCT) as a thought experiment to help judge potential biases in individual observational studies (e.g., Sterne et al. 2016; Morgan et al. 2019). We recognize that the RCT model, coupled with thinking about confounding based on counterfactuals and the use of directed acyclic graphs (DAGs) to depict causal relations, has helped advance causal inference in many instances in observational epidemiology. The relative strengths and weaknesses of RCTs vs. observational studies have been an ongoing discussion in the literature (e.g., Eden et al. 2008; Sørensen et al. 2006) but is worth reemphasizing here with respect to environmental epidemiologic studies. RCTs, if properly conducted, can, in theory, avoid or minimize some of the main potential limitations of observational studies (e.g., selection bias, confounding, and differential information bias). However, comparing studies to an RCT gold standard inevitably begins by classifying observational studies as of lower quality, as these studies have the potential to suffer from biases theoretically avoided by RCTs. GRADE, for example, states that "Evidence from randomized controlled trials starts at high quality and, because of residual confounding, evidence that includes observational data starts at low quality" (https://bestpractice.bmj.com/info/us/toolkit/learn-ebm/what-is-grade/). The Navigation Guide and OHAT consider observational studies to provide evidence of moderate quality (Woodruff and Sutton 2014; NTP 2019).

Table 2. Some common practices and suggested improvements to risk of bias assessments for individual environmental epidemiologic studies and evidence synthesis.

Current practice	Suggested improvement
Individual studies	
Compare to RCTs as ideal study	Do not consider RCTs as ideal study
Evaluate bias in different domains (e.g., confounding, selection bias, measurement error)	Consider the magnitude and direction of different biases and evaluate the net likely effect
Rank potential biases (e.g., low, moderate, high)	Rank biases considering the suggestions in rows above
No evaluation of statistical methods	Add a domain for statistical methodology similar to IARC's, i.e., assess the ability to obtain unbiased estimates of exposure–outcome associations, confidence intervals, and test statistics. Appropriateness of methods used to investigate and control confounding
Evidence synthesis	
In some instances, downgrade all observational studies as weak or moderate quality	Assume observational studies are high quality unless important biases are likely
Reject some studies from evidence synthesis based on ranking of bias across their domains. Often make overall judgment based on meta-analyses after rejection of those studies	Retain most studies in evidence synthesis. Use methods such as sensitivity analyses and triangulation to consider net effect of possible biases. Consider evidence from other studies that were not included in meta-analysis because of different designs or parameters

Note: IARC, International Agency for Research on Cancer; RCT, randomized controlled trial.

The RCT gold-standard assumption can lead to extremes in which observational studies are dismissed in their entirety. For example, the current chair of the EPA Clean Air Scientific External Advisory Committee had argued that *a*) all observational studies quantifying an exposure–response relationship are subject to a critical level of bias (Cox 2017, 2018); and *b*) all air pollution epidemiology studies lack adequate control for confounding and are, therefore, subject to high risk for potential bias [see review by Goldman and Dominici (2019) and commentary by Balmes (2019)]. However, the EPA has said they will maintain their traditional approach of considering all observational studies without prejudice when evaluating scientific evidence for hazard identification of criteria air pollutants to meet their mandate for clean air (Parker 2019).

We argue, in contrast, that RCTs are not the gold standard for judging observational studies, particularly occupational and environmental studies. RCTs of most environmental and occupational exposures are, by definition, not possible, as one cannot ethically randomize people to potentially harmful exposures with no perceived benefit. Beyond that, RCTs typically involve limited sample sizes and short follow-up times, which are often inadequate for observing chronic disease or rare outcomes. RCTs deliver the exposure (e.g., medication) at the beginning of follow-up, typically in a limited number of dose levels, which does not mimic the real-life circumstances of environmental observational studies. The RCT may involve highly selective study groups meeting particular criteria, which may have little generalizability to other populations.

In contrast, in real life, and thus in observational studies, uncontrolled exposures are often present before follow-up begins, occur at many different exposure levels, and may vary by intensity, time of first exposure, and duration of exposure. Observational studies often involve outcomes (e.g., cancer and neurodegeneration) with long latencies following exposure, necessitating long follow-up periods with evaluation of lagged exposures and latency periods.

They often also include long exposure histories (which are important for assessing cumulative exposure), require retrospective exposure assessment, and include people who change exposure categories over time. A proportion of the population is likely to have other concomitant exposures, some of which may have similar effects. Observational studies often focus on exposure–response relationships, rather than simple comparisons of an outcome among the exposed and nonexposed population. As a result, exposure–response models have been developed that address complex issues such as control for confounders, consideration of the importance of measurement error in parameter estimation, model misspecification, and the possible use of Bayesian methods to incorporate prior beliefs.

We believe there should be no *a priori* assumption that observational studies are weaker than RCTs for studying occupational and environmental exposures, and it should be acknowledged that they generally represent the best available evidence to assess causality. Others have concluded the same. For example, the Institute of Medicine (Eden et al. 2008) report concluded, "Randomized controlled trials can best answer questions about the efficacy of screening, preventive, and therapeutic interventions while observational studies are generally the most appropriate for answering questions related to prognosis, diagnostic accuracy, incidence, prevalence, and etiology." Thus, in our view, observational studies should be considered as the norm and then assigned a lower quality if significant substantial biases are likely that would affect the parameter estimates.

Identifying, describing, and ranking biases. Absent an RCT, reviewers need to ask, "Among observational studies, what are the possible biases, how likely are they, what direction are they, and how much are they likely to impact the parameter estimate?" Well-conducted assessments of biases and their potential impact are essential in evaluating the contribution of individual studies to evidence on causality, but their implementation is sometimes problematic. In some cases, it may be difficult to estimate the magnitude of the bias or to appropriately assign direction or weights for the various biases or the impact of the biases on the outcome estimates. Often, the original study authors do not provide methods complete enough to evaluate bias issues. Thus, this process necessarily involves subjectivity, which should be informed by expert judgment.

The quantification and relative importance of possible biases is a critical emerging field of epidemiology (Lash et al. 2014), but detailed quantification may be beyond the purview of most current systematic reviews. However, we believe it is important to incorporate these methods to the extent possible in risk of bias tools and specify them (or the lack thereof) in the methods section of systematic reviews.

For example, we know historically that many cohort studies of occupational exposures were criticized for not having smoking data when considering smoking-related diseases like lung cancer. However, relatively early in modern epidemiology, Cornfield et al. (1959) explained how to quantitatively assess the likely importance of confounding factors that differ between exposed and nonexposed populations. Later, it was shown both theoretically (Axelson 1980) and empirically (Siemiatycki et al. 1988) that confounding by smoking is unlikely to explain relative risks (RRs) for lung cancer that exceed 1.2 to 1.4 in the occupational setting. This estimation was based on analyses comparing differences in smoking status between workers and general population referents (with workers often smoking more) and accounting for the large RR for smoking and lung cancer. Hence, although smoking is a very strong lung cancer risk factor and a strong potential confounder, it is not likely to account for the large RRs found for classic occupational lung carcinogens. The impact of lack of control for smoking is expected to be even smaller for outcomes to which it is more weakly related or, in general, for any confounder only weakly related to disease.

The work by Cornfield (1959) and others has been followed by methods to estimate the strength of association that a hypothetical unmeasured confounder, about which the investigator has no prior knowledge, must have with both the outcome and the exposure to have an impact equal to the observed effect (VanderWeele and Ding 2017; Ding and VanderWeele 2016; Lubin et al. 2018). This estimation enables us to say, for example, that for an RR of about 2.0 to be explained by an unmeasured confounder, the unmeasured confounder would have to have a minimal RR of 3.5 with both the exposure and outcome to reduce the observed association to the null value of 1.0. For a factor to have such a strong effect on the outcome and a strong association with the exposure, but be unknown, is generally unlikely in most settings. Yet, even today, smoking, other identified confounders, and other unknown confounders continue to be raised as possible sources of bias to explain positive findings (VanderWeele and Ding 2017). Thus, as a minimum, estimating the likely maximum extent of (unmeasured) confounding and discussing its likely impact on the observed effect estimate can and should be done.

These same considerations of magnitude and direction of bias apply to other potential biases beyond confounding, e.g., selection bias and measurement error. For example, classical nondifferential measurement error will generally bias effect measures to the null so that if an elevated risk is found, it is not likely because of this source of error (although other biases might bias against the null). In contrast, Berkson measurement error generally affects the precision of the findings but not the actual point estimates (Armstrong 1998). Moreover, mismeasurement may have little consequence on exposure–response parameters when there are large exposure contrasts in the population (Avanasi et al. 2016). Sensitivity and specificity of classification of the outcome may also play a role. Selection bias, occurring during recruitment, can only affect estimates to the degree that the estimated parameter of interest (i.e., the association between exposure and outcome) among those people not included in a study differs substantially from that parameter in the population studied. Furthermore, in general, if the direction of two sources of bias is different, they may approximately cancel each other out. A reviewer needs to consider multiple sources of possible bias and their relative importance and whether they are likely to have a net effect of bias away or toward the null in the effect estimate. Thus, taking such considerations into account requires that we go beyond some current ranking schemes (Table 2). We note that these same issues are at the heart of triangulation, which is discussed in more detail below. We believe it is possible to improve risk of bias tools to formally incorporate magnitude and direction of bias, but it will take considerable work.

Other possible domains in risk of bias tools. Another possible bias domain is conflict of interest, which can create a potential bias and is not always assessed in risk of bias tools. There is strong evidence that studies authored by those with vested interest are generally favorable to those interests, hence the need to disclose potential conflict of interests. The effects of conflicts of interest are well documented in clinical medicine (Angell 2008; Krauth et al. 2014; Lundh et al. 2012), and biased results from similar conflicts of interest have been documented in occupational and environmental epidemiology (Michaels 2009). Evaluation of risk of bias regarding conflict of interest may also assess issues of selective reporting of study results (e.g., only reported results that are significant). In our view, however, a potential conflict of interest does not define a specific bias in and of itself, and if specific biases are present, reviewers should be able to detect them in evaluating studies. Hence, we do not argue

to include conflict of interest as a separate domain for risk of bias tools, although such potential conflicts must be clearly acknowledged by authors.

Another domain, generally not included in current risk of bias tools, is potential bias because of problems in statistical methodology. Concerns include choice of an inappropriate and badly fitting model, failure to model exposure–response or to evaluate different exposure–response models, incorrect use of mixed models, incorrect use of Bayesian techniques, violation of statistical assumptions (e.g., normal residuals in linear regression), overadjustment for covariates related to exposure but not to outcome, adjusting for causal intermediates, etc. It may be possible to infer whether the resulting biases can be toward or away from the null; otherwise, it simply may be necessary to indicate that the effect estimate is likely to be incorrect without knowing the direction of bias. We are in favor of adding this domain to risk of bias tools.

Another domain that does not entail bias *per se* is informativeness. Consideration in this domain includes whether the study has a large enough sample size, whether the study has sufficient latency, whether results have been reported selectively, and whether the study has sufficient exposure contrast to see an effect of exposure on outcomes. This domain is sometimes called sensitivity in some evidence syntheses (Cooper et al. 2016). However, some of these concerns may be addressed in other domains, particularly the exposure domain or during evidence synthesis, and we do not argue for including informativeness in risk of bias tools.

Summary of concerns about current risk of bias tools. Table 2 lists some areas of concern for risk of bias tools in current use and some suggested improvements. It should be noted that current risk of bias tools differ, and we note in Table 1 what we consider the pros and cons of current risk of bias tools with regard to our areas of concern. We focus on existing risk of bias tools appropriate for observational studies and that include some kind of algorithm to rank different study domains (e.g., confounding, selection bias, and measurement error), i.e., the risk of bias tools noted above in Table 1. An earlier, somewhat analogous review of risk of bias tools, including systematic reviews that assess risk of bias without a formal risk of bias tool, has been published previously by Rooney et al. (2016).

Broad Types of Evidence Synthesis and Causal Inference

Risk of bias assessments of individual studies are eventually incorporated into a broader evidence synthesis (e.g., within systematic reviews), where they can play an important role (Figure 1). A number of government and public health agencies have formalized evidence synthesis guidelines using a broad approach, including IARC (IARC 2019a), the National Toxicology Program (NTP)'s OHAT (NTP 2019), the EPA (National Research Council 2014), and the NTP's "Report on Carcinogens" (NTP 2015).

Here, we discuss such broad syntheses and suggest some ways that bias assessment might be used in them. Table 2 describes some current practices for use of risk of bias assessment in evidence synthesis and suggests some improvements. Our concerns about evidence synthesis stem from the same concerns above regarding risk of bias assessment for individual studies, i.e., the *a priori* downgrading of observational studies compared with a randomized trial and the potential exclusion of useful studies in synthesis because of being judged overly biased.

Ultimately, every process of causal inference should involve the synthesis of different types of evidence (Broadbent et al. 2016; Vandenbroucke et al. 2016), and, in most cases, no single study is sufficient or definitive, although there are exceptions (Benbrahim-Tallaa et al. 2014; de Boer et al. 2018). Even when a study or a set of studies of similar design appears sufficient to establish an association suggesting causality, other kinds of inference may be informative for evidence synthesis (Cartwright 2007). The use of different pieces of interlocking evidence in an argument has support among philosophers of science, e.g., in the crossword analogy by Haack (1998). This can be formalized in terms of triangulation (Lawlor et al. 2016). Below, we present two well-known examples of a broad synthesis procedure, old and new, followed by a discussion of triangulation.

Bradford Hill, a classic treatise on evidence synthesis. Most epidemiologists are familiar with the classic considerations for assessing causality of Bradford Hill (Figure 2), first elaborated over 50 y ago, e.g., strength of association, consistency, dose–response (biological gradient), biological plausibility, temporality (exposure precedes disease), and specificity.

Less well-known are coherence (concordance between epidemiological and laboratory findings), experiment (e.g., a natural experiment in which exposure ceases and disease is then diminished), and analogy (similarities between the observed association and other associations). These last three considerations go beyond direct evidence from epidemiologic studies and are consistent with both the IARC approach and triangulation (Lawlor et al. 2016), both described below. Hill, who called these considerations merely viewpoints, did not consider any of them as a necessary or sufficient requirement for causality, nor do most epidemiologists so consider them today (with the exception of temporality; see Rothman and Greenland 2005). The spirit with which Bradford Hill considered his viewpoints, however, remains a cautionary guide arguing against too mechanical an application of checklists in risk of bias analyses and their subsequent use in evidence synthesis. On the other hand, bias assessments can and should be used in the spirit of Hill.

IARC Monographs, a modern variation of Bradford Hill. In judging whether a given agent is carcinogenic, IARC assesses three types of evidence: *a*) animal bioassay evidence; *b*) human cancer epidemiologic evidence; and *c*) animal, human, and *in vitro* mechanistic evidence (Samet et al. 2020; IARC 2019a). Judgments are made as to whether the evidence in each area is sufficient (or strong in the case of mechanisms), limited, or inadequate, with the final assessment about human carcinogenicity based on integration of these judgments (Figure 3).

There is no formal risk of bias assessment or semiquantitative ranking scheme in the IARC Monographs, although guidelines are provided in the IARC Preamble on how to evaluate individual study quality, synthesize the body of evidence within a single line of evidence (e.g., epidemiologic studies of cancer), and put the streams of evidence together (IARC 2019a). Regarding bias evaluation in individual studies, the Preamble provides a list of different sources similar to many risk of bias tools (confounding, measurement error, selection bias, etc.), as well as the criteria for evaluating each of these domains with regard to bias, and calls for assessment of their impact. The Preamble adds a domain for statistical methodology and also evaluates informativeness or

1. Consistency
2. Specificity
3. Temporality
4. Biological gradient
5. Plausibility
6. Coherence
7. Experiment
8. Analogy

"What I do not believe – and this has been suggested – is that we can usefully lay down some hard-and-fast rules of evidence that *must* be obeyed before we can accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us make up our minds on the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?"

Figure 2. Bradford Hill's viewpoints (Hill 1965).

Evidence of cancer in humans	Evidence of cancer in experimental animals	Mechanistic evidence	Evaluation
Sufficient			Carcinogenic (Group 1)
	Sufficient	Strong (exposed humans)	
Limited	Sufficient		Probably carcinogenic (Group 2A)
Limited		Strong	
	Sufficient	Strong (human cells or tissues)	
		Strong (mechanistic class)	
Limited			Possibly carcinogenic (Group 2B)
	Sufficient		
		Strong (experimental systems)	
	Sufficient	Strong (does not operate in humans)	Not classifiable (Group 3)
All other situations not listed here			

Figure 3. IARC Monographs criteria for evidence synthesis. Adapted from IARC (2019b).

study sensitivity (i.e., the ability to show a true association), which are not generally part of other evidence syntheses [the NTP's "Report on Carcinogens" (NTP 2015) is an exception].

The Preamble then states, with regard to a list of possible biases, that "these sources of error do not constitute and should not be used as a formal checklist of indicators of study quality. The judgment of experienced experts is critical in determining how much weight to assign to different issues in considering how all of these potential sources of error should be integrated and how to rate the potential for error related to each of these considerations." Expert judgment comes into play, for example, when determining how informative individual epidemiological studies are for assessing the carcinogenic hazard posed by the agent and the relevance of specific animal and mechanistic studies. For a judgment that a substance causes a given cancer, IARC requires that the observed overall association as judged in evidence synthesis cannot be explained by confounding, systematic bias, or chance. This formulation corresponds to a bias impact assessment across studies.

Although IARC guidelines for assessing bias are cautionary notes against checklist approaches, they are not formally applied to individual studies for specific domains in the IARC Monographs, although overall strengths and weaknesses of studies are mentioned, and bias impact (where assessable) is emphasized. We argue that bias assessment tools, which reflect the spirit of the IARC guidelines, can—as noted earlier—increase transparency and replicability in the evaluation of individual studies.

Triangulation. Recently, there has been increased interest in the use of triangulation approaches (Lawlor et al. 2016) for assessing causality in evidence synthesis. If different studies are generally concordant regarding the observed association, especially if their possible biases point in opposing directions, a causal interpretation is supported. Thus, triangulation specifically chooses populations and study designs where the bias is likely to be in different directions. Different types of triangulation include cross-context comparisons, use of different controls (in a case–control study), natural experiments, within-sibling comparisons, instrumental variable analyses, Mendelian randomization, exposure negative controls, and outcome negative controls. Other non-RCT-based methods, different from triangulation, include population comparisons and regression discontinuity studies (Pearce et al. 2019). A form of triangulation may be used in meta-analysis in which studies that have different potential sources of bias are variously included or excluded from an analysis to determine the impact, thereby drawing causal inference from the impact of such inclusions or exclusions (Lee et al. 2009; Honaryar et al. 2019; Hauptmann et al. 2020). This type of triangulation is often called sensitivity analysis.

For example, Pearce et al. (1986) conducted a case–control study of pesticide exposure and non-Hodgkin lymphoma, which involved two control groups: *a*) a general population control group; and *b*) an "other cancers" control group. It was hypothesized that the former control group would produce an upward bias in the estimated odds ratio (differential recall bias if healthy general population controls are less likely to remember previous exposure than the cancer cases), whereas the "other cancers" control group could produce a downward bias in the estimated odds ratio (if any of the other cancers were also caused by the pesticide exposure under study). Both groups yielded similar findings, indicating that neither bias was occurring to any discernible degree. This provided strong evidence that little recall bias or selection bias was occurring. Although a flexible risk of bias tool and expert judgment might have reached this conclusion, a mechanical risk of bias assessment could have led to the rejection of both components of the study as being at high risk of bias (albeit in opposite directions).

Triangulation is also consistent with the approach advocated by Savitz et al. (2019), who argue that risk of bias assessments should focus on identifying a small number of the most likely influential sources of bias, classifying each study on how effectively it has addressed these potential biases (or was likely to have the bias) and determining whether results differ across studies in relation to these hypothesized biases. For example, information bias is unlikely to explain positive findings of studies with nondifferential exposure misclassification if stronger findings are found among studies with better exposure assessments. A good example of triangulation by assessing exposure quality can be found in Lenters et al. (2011), who evaluated the association between asbestos and lung cancer. In this analysis, stratification by exposure assessment characteristics revealed that studies with a well-documented exposure assessment, larger contrast in exposures, greater coverage of the exposure history by the exposure measurement data, and more complete job histories had higher risk estimates per unit dose than did studies without these characteristics. Observing risk estimates that move in the direction of expectation of study quality adds to the strength of evidence. Similar observations have been made for other environmental and occupational exposures (Vlaanderen et al. 2011).

Similarly, if all the likely major sources of confounding have been adjusted for in some studies, these adjustments made little difference to the findings, and the results were similar to those

studies that were not able to adjust for confounding, confounding is less likely to explain the findings. For example, in a meta-analysis of lung cancer and exposure to diesel engine exhaust, 16 of 29 studies controlled for smoking, whereas 13 studies did not. However, meta-risk estimates for smoking-adjusted [RR = 1.35 (95% CI: 1.22–1.49)] and unadjusted [RR = 1.33 (95% CI: 1.25–1.41)] were virtually identical, indicating that not correcting for smoking did not substantially bias the effect estimates (and suggesting that smoking was not a confounder) (Bhatia et al. 1998).

The approach to assessing risk of bias in individual studies as a preliminary step to dismissing specific studies from further evidence synthesis does not fit well with these more nuanced approaches to knowledge synthesis because this approach: *a*) may focus on each study individually, making it impossible to conduct thoughtful triangulation analyses (e.g., cross-context comparisons); *b*) usually does not assess the likely direction and magnitude of the bias and impact on effect estimates; *c*) may not account for important secondary evidence such as exposure or outcome negative controls (Lipsitch et al. 2010); and *d*) may downgrade a specific study based on one flaw while ignoring other strong points. The danger of simplistic risk of bias approaches is that they may result in excluding a great deal of relevant and important evidence. On the other hand, existing risk of bias tools, if flexibly used, can avoid these flaws and can be strengthened by adding guidelines that consider triangulation.

A counterexample where triangulation would have been helpful is provided by a review of diesel exhaust and lung cancer by Möhner and Wendt (2017). These authors did not use a formal risk of bias tool; rather, they adopted a study-by-study approach and dismissed most studies because of their judgment that the studies were biased. They dismissed all 10 case–control studies (all adjusted for smoking) because of alleged residual confounding by smoking stemming from the exclusion of controls with smoking-related diseases. They also argued that the population-based case–control studies suffered from selection biases because controls were more likely to be more highly educated and less likely to be exposed to occupational diesel fumes. However, if this latter point were true, population-based case–control studies would show a higher risk per unit of diesel exposure than hospital-based case–control studies. The large SYNERGY project (http://synergy.iarc.fr), which pooled lung cancer case–control studies of occupation, used a standardized exposure assessment across both population- and hospital-based studies and found no indication of such bias (Olsson et al. 2011). Möhner and Wendt also dismissed virtually all the cohort studies. They argued that the Diesel Exhaust in Miners Study (DEMS) was the only study that could be used for evaluating lung cancer and diesel exhaust quantitatively among the 30 or more studies they reviewed (Silverman et al. 2012; Attfield et al. 2012). They judged that the DEMS study itself was analyzed incorrectly because of the method of controlling for smoking separately for surface vs. underground mining in the same model (smoking strata within each job category stratum), even though separate analyses of surface and underground miners, controlling for smoking, showed little evidence of a diesel–lung cancer effect in the former and a strong largely monotonic effect in the latter, where exposure contrasts were much stronger. Overall, Möhner and Wendt concluded that the evidence did not support a causal link between diesel exhaust and lung cancer. In contrast, a pooled exposure–response analysis (a type of triangulation) of the three prospective studies with quantitative data revealed that the estimates of these three cohorts were largely consistent, even though all could potentially have suffered from different shortcomings (Vermeulen et al. 2014). Similarly, an expert committee convened by IARC in 2012 determined that diesel exhaust is a Group 1 human carcinogen based on sufficient evidence of carcinogenicity from epidemiological studies on lung cancer (IARC 2014).

Examples of Contrasting Approaches to Systematic Reviews and Evidence Synthesis

Below, we consider two examples of systematic reviews with contrasting approaches. The red meat review is an example, in our view, of how not to do a systematic review, and the second example, regarding low-dose ionizing radiation, is an example of how it should be done.

Consumption of red and processed meat. An example of a problematic risk of bias assessment, embedded within a systematic review and meta-analysis, concerns a set of recent papers evaluating the evidence of an association between red meat and processed meat consumption and risk of cancer, among other outcomes (Han et al. 2019; Vernooij et al. 2019; Zeraatkar et al. 2019). Concerns regarding this review include inappropriate use of risk of bias assessments and downgrading of observational studies based solely on using RCTs as a gold standard. We focus here on the paper by Han et al. (2019) regarding cancer, particularly colorectal cancer.

Han et al. (2019) used a risk of bias approach that considered all sources of bias as equally important and arbitrarily assigned studies as having a high risk of bias if two or more elements were rated as having high risk of bias, regardless of the direction or impact of the likely bias [e.g., dietary information bias usually biases toward the null (Freedman et al. 2011)]. In addition, the risk of bias evaluations for exposure downgraded most of the large cohort studies by requiring repeated dietary assessment (e.g., with food frequency questionnaires) be administered every 5 y for the study to receive a low risk of bias rating.

In their evidence synthesis for colorectal cancer incidence and other outcomes, these authors qualitatively weighted evidence from randomized trials more heavily than observational studies despite the small exposure contrasts for meat consumption and short follow-up provided by the trials (Zeraatkar et al. 2019). They universally downgraded the observational studies, which they automatically deemed as providing low or very low certainty despite showing strong and consistent evidence of a dose–response gradient, based on a presumption of their inherent bias, using GRADE as their criteria.

Although their meta-RR per unit of meat intake for colorectal cancer incidence in relation to processed meat was of a similar magnitude to previously reported meta-analyses, the systematic review authors described the overall effect of processed meat as of being of low-to-very low certainty because of the meta-RR's origin in inherently low-certainty observational studies, as well as the small effect observed (which, to some extent, was an artifact of the unit they selected for an RR). They used this interpretation to develop dietary guidelines that red and processed meat consumption need not be reduced (Johnston et al. 2019; Carroll and Doherty 2019), generating considerable opposition by leading subject matter experts. (e.g., Qian et al. 2020; https://www.hsph.harvard.edu/nutritionsource/2019/09/30/flawed-guidelines-red-processed-meat/). The World Cancer Research Fund (https://www.wcrf.org/int/latest/news-updates/red-and-processed-meat-still-pose-cancer-risk-warn-global-health-experts) recommends limiting intake of processed and red meat based on their expert systematic reviews, which concluded that the evidence for an increased risk of colorectal cancer was convincing for processed meat and probable for red meat. IARC reached similar conclusions for processed meat and red meat being probably carcinogenic to humans (IARC 2018).

Low-dose ionizing radiation. A new systematic review of low-dose ionizing radiation in relation to cancer was recently published. It involved an assessment of biases in the relevant literature and evidence synthesis with a meta-analysis. The rationale and framework for the review are described in Berrington de

Gonzalez et al. (2020). The assessment of biases consisted of a series of four articles that dealt with different types of biases, including dose measurement error (Daniels et al. 2020), confounding and selection bias (Schubauer-Berigan et al. 2020), potential outcome misclassification (Linet et al. 2020), and impacts of different analytical methods (Gilbert et al. 2020). Here, we focus on the assessment of confounding. The article goes over each of the relevant studies and assesses whether potential confounders were controlled and, if they were not, the likely magnitude and direction of bias and the potential impact in the effect estimate. This approach included indirect adjustments for uncontrolled confounders to assess their likely impact on effect estimates, based on likely distributions of the confounder by exposure level and the RR of cancer because of the confounder. The authors summarized the relevant literature (26 studies) that had used cumulative radiation exposure as their metric, had mean cumulative levels <100 milligray (mGy), and used a model that calculated the excess RR per milligray. Their conclusion was that for most studies, substantial confounding was unlikely. Of the 14 studies that were likely to suffer from some degree of confounding bias, 7 were judged potentially biased to the null, 2 away from the null, and in 5, the direction could not be assessed (in some instances because of two biases operating in opposite directions). A similar exercise was done for selection biases.

The subsequent meta-analysis (Hauptmann et al. 2020) contained a number of sensitivity analyses in which studies with likely confounding away from the null, or in which the direction could not be ascertained, were excluded. The authors made note of which studies contributed to heterogeneity and whether the heterogeneity was likely because of uncontrolled biases. The authors concluded, "Traditionally, systematic reviews classify the quality of a study but without formally considering whether the quality of information translates into a bias. . . In addition, the direction of the bias, and, if possible, the magnitude of the potential bias need to be assessed. Without these further considerations, exclusions based on quality or potential bias could result in substantial loss of information. Such an approach has been recently recommended over other approaches, such as the use of a 'risk of bias' checklist" (Savitz et al. 2019).

Conclusions

We have outlined some suggested approaches for evidence synthesis in systematic reviews, particularly for reviews focused on environmental and occupational exposures in observational studies. Broad approaches to synthesis include the use of classical considerations for judging causality, consideration of various streams of evidence including animal and mechanistic data, as well as triangulation. Although we consider that risk of bias assessments can play an important role in evidence synthesis when used appropriately, we also note some potential drawbacks, i.e., their potential for misuse and some of their current limitations. First, we do not consider that RCTs provide a gold-standard model for observational studies such that observational studies should start with an assumption of bias. Second, we argue against simply producing a list of possible biases for individual studies, ranking their different domains by presumed degree of bias, and using these rankings across domains to judge whether individual studies suffer critical bias and should be excluded from evidence synthesis. We call, rather, for more comprehensive identification of the likely direction and magnitude of possible biases and their impact on the parameter estimates. We are concerned that misuse of risk of bias assessment can be used to dismiss important evidence of exposure effects on health, which can have hazardous public health consequences.

Acknowledgments

This work was motivated by the authors' participation in workshops for the development of the ROBINS-E risk of bias tool. There was no grant funding for this work. The authors alone are responsible for the views expressed in this article, and they do not necessarily represent the views, decisions, or policies of the institutions with which they are affiliated.

References

Angell M. 2008. Industry-sponsored clinical research: a broken system. JAMA 300(9):1069–1071, PMID: 18768418, https://doi.org/10.1001/jama.300.9.1069.

Armstrong BG. 1998. Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med 55(10):651–656, PMID: 9930084, https://doi.org/10.1136/oem.55.10.651.

Arroyave WD, Mehta SS, Guha N, Schwingl P, Taylor KW, Glenn B, et al. 2020. Challenges and recommendations on the conduct of systematic reviews of observational epidemiologic studies in environmental and occupational health. J Expo Sci Environ Epidemiol. In press, PMID: 32415298, https://doi.org/10.1038/s41370-020-0228-0.

Attfield MD, Schleiff PL, Lubin JH, Blair A, Stewart PA, Vermeulen R, et al. 2012. The Diesel Exhaust in Miners study: a cohort mortality study with emphasis on lung cancer. J Natl Cancer Inst 104(11):869–883, PMID: 22393207, https://doi.org/10.1093/jnci/djs035.

Avanasi R, Shin HM, Vieira VM, Savitz DA, Bartell SM. 2016. Impact of exposure uncertainty on the association between perfluorooctanoate and preeclampsia in the C8 Health Project population. Environ Health Perspect 124(1):126–132, PMID: 26090912, https://doi.org/10.1289/ehp.1409044.

Axelson O. 1980. Aspects of confounding and effect modification in the assessment of occupational cancer risk. J Toxicol Environ Health 6(5–6):1127–1131, PMID: 7463507, https://doi.org/10.1080/15287398009529933.

Balmes JR. 2019. Do we really need another time-series study of the PM(2.5)-mortality association? N Engl J Med 381(8):774–776, PMID: 31433927, https://doi.org/10.1056/NEJMe1909053.

Benbrahim-Tallaa L, Lauby-Secretan B, Loomis D, Guyton KZ, Grosse Y, El Ghissassi F, et al. 2014. Carcinogenicity of perfluorooctanoic acid, tetrafluoroethylene, dichloromethane, 1,2-dichloropropane, and 1,3-propane sultone. Lancet Oncol 15(9):924–925, PMID: 5225686, https://doi.org/10.1016/S1470-2045(14)70316-X.

Berrington de Gonzalez A, Daniels RD, Cardis E, Cullings HM, Gilbert E, Hauptmann M, et al. 2020. Epidemiological studies of low-dose ionizing radiation and cancer: rationale and framework for the Monograph and overview of eligible studies. J Natl Cancer Inst Monogr 2020(56):97–113, https://doi.org/10.1093/jncimonographs/lgaa009.

Bhatia R, Lopipero P, Smith AH. 1998. Diesel exhaust exposure and lung cancer. Epidemiology 9(1):84–91, PMID: 9430274, https://doi.org/10.1097/00001648-199801000-00017.

Broadbent A, Vandenbroucke J, Pearce N. 2016. Causality and causal inference in epidemiology: we need also to address causes of effects Reply. Int J Epidemiol 45(6):1776–1786, PMID: 26800751, https://doi.org/10.1093/ije/dyv341.

Carroll AE, Doherty TS. 2019. Meat consumption and health: food for thought. Ann Intern Med 171(10):767–768, PMID: 31569212, https://doi.org/10.7326/M19-2620.

Cartwright N. 2007. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge, UK: Cambridge University Press.

Cooper GS, Lunn RM, Ågerstrand M, Glenn BS, Kraft AD, Luke AM, et al. 2016. Study sensitivity: evaluating the ability to detect effects in systematic reviews of chemical exposures. Environ Int 92–93:605–610, PMID: 27156196, https://doi.org/10.1016/j.envint.2016.03.017.

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. JNCI 22(1):173–203, https://doi.org/10.1093/jnci/22.1.173.

Cox LAT Jr. 2017. Do causal concentration-response functions exist? A critical review of associational and causal relations between fine particulate matter and mortality. Crit Rev Toxicol 47(7):603–631, PMID: 28657395, https://doi.org/10.1080/10408444.2017.1311838.

Cox LA Jr. 2018. Re: "Best practices for gauging evidence of causality in air pollution epidemiology. Am J Epidemiol 187(6):1338–1339, PMID: 29584873, https://doi.org/10.1093/aje/kwy034.

Daniels RD, Kendall GM, Thierry-Chef I, Linet MS, Cullings HM. 2020. Assessment of strengths and weaknesses of dosimetry systems used in epidemiologic studies of low-dose radiation exposure and cancer risk. J Natl Cancer Inst Monogr. In press.

de Boer M, van Leeuwen FE, Hauptmann M, Overbeek LIH, de Boer JP, Hijmering NJ, et al. 2018. Breast implants and the risk of anaplastic large-cell lymphoma in the breast. JAMA Oncol 4(3):335–341, PMID: 29302687, https://doi.org/10.1001/jamaoncol.2017.4510.

Ding P, VanderWeele TJ. 2016. Sensitivity analysis without assumptions. Epidemiology 27(3):368–377, PMID: 26841057, https://doi.org/10.1097/EDE.0000000000000457.
Eden J, Wheatley B, McNeil B, Sox H. 2008. *Knowing What Works in Health Care: A Roadmap for the Nation*. Washington, DC: The National Academies Press.
Freedman LS, Schatzkin A, Midthune D, Kipnis V. 2011. Dealing with dietary measurement error in nutritional cohort studies. J Natl Cancer Inst 103(14):1086–1092, PMID: 21653922, https://doi.org/10.1093/jnci/djr189.
Gilbert ES, Little MP, Preston DL, Stram DO. 2020. Issues in interpreting epidemiologic studies of populations exposed to low-dose, high-energy photon radiation. J Natl Cancer Inst Monogr 2020(56):176–187, PMID: 32657345, https://doi.org/10.1093/jncimonographs/lgaa004.
Goldman GT, Dominici F. 2019. Don't abandon evidence and process on air pollution policy. Science 363(6434):1398–1400, PMID: 30898845, https://doi.org/10.1126/science.aaw9460.
Haack S. 1998. *Manifesto of a Passionate Moderate*. Chicago, IL: Chicago University Press.
Han MA, Zeraatkar D, Guyatt GH, Vernooij RWM, El Dib R, Zhang Y, et al. 2019. Reduction of red and processed meat intake and cancer mortality and incidence: a systematic review and meta-analysis of cohort studies. Ann Intern Med 171(10):711–720, PMID: 31569214, https://doi.org/10.7326/M19-0699.
Hauptmann M, Daniels RD, Cardis E, Cullings Hm, Kendall GM, Laurier D, et al. 2020. Epidemiological studies of low-dose ionizing radiation and cancer: summary bias assessment and meta-analysis. J Natl Cancer Inst Monogr 2020(56):188–200, https://doi.org/10.1093/jncimonographs/lgaa010/.
Higgins JPT, Green S. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons.
Hill AB. 1965. The environment and disease: association or causation? Proc R Soc Med 58(5):295–300, PMID: 14283879, https://doi.org/10.1177/003591576505800503.
Honaryar MK, Lunn RM, Luce D, Ahrens W, 't Mannetje A, Hansen J, et al. 2019. Welding fumes and lung cancer: a meta-analysis of case-control and cohort studies. Occup Environ Med 76(6):422–431, PMID: 30948521, https://doi.org/10.1136/oemed-2018-105447.
IARC (International Agency for Research on Cancer). 2014. *Monograph 105: Diesel and Gasoline Engine Exhausts and Some Nitroarenes*, https://publications.iarc.fr/129 [accessed 3 September 2020].
IARC. 2018. *Monograph 114: Red Meat and Processed Meat*, https://publications.iarc.fr/564 [accessed 3 September 2020].
IARC. 2019a. *IARC Monographs on the Identification of Carcinogenic Hazards to Humans: Preamble*, https://monographs.iarc.fr/wp-content/uploads/2019/07/Preamble-2019.pdf [accessed 3 September 2020].
IARC. 2019b. *Revised Preamble for the IARC Monographs*, https://monographs.iarc.fr/wp-content/uploads/2019/07/2019-SR-001-Revised_Preamble.pdf [accessed 3 September 2020].
Johnston BC, Zeraatkar D, Han MA, Vernooij RWM, Valli C, El Dib R, et al. 2019. Unprocessed red meat and processed meat consumption: dietary guideline recommendations from the Nutritional Recommendations (NutriRECS) Consortium. Ann Intern Med 171(10):756–764, https://doi.org/10.7326/M19-1621.
Jüni P, Witschi A, Bloch R, Egger M. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282(11):1054–1060, PMID: 10493204, https://doi.org/10.1001/jama.282.11.1054.
Krauth D, Anglemyer A, Philipps R, Bero L. 2014. Nonindustry-sponsored preclinical studies on statins yield greater efficacy estimates than industry-sponsored studies: a meta-analysis. PLoS Biol 12(1):e1001770, PMID: 24465178, https://doi.org/10.1371/journal.pbio.1001770.
Lash TL, Fox MP, MacLehose R, Maldonado G, McCandless LC, Greenland S. 2014. Good practices for quantitative bias analysis. Int J Epidemiol 43(6):1969–1985, PMID: 25080530, https://doi.org/10.1093/ije/dyu149.
Lawlor DA, Tilling K, Davey Smith G. 2016. Triangulation in aetiological epidemiology. Int J Epidemiol 45(6):1866–1886, PMID: 28108528, https://doi.org/10.1093/ije/dyw314.
Lee YCA, Cohet C, Yang YC, Stayner L, Hashibe M, Straif K. 2009. Meta-analysis of epidemiologic studies on cigarette smoking and liver cancer. Int J Epidemiol 38(6):1497–1511, PMID: 19720726, https://doi.org/10.1093/ije/dyp280.
Lenters V, Vermeulen R, Dogger S, Stayner L, Portengen L, Burdorf A, et al. 2011. A meta-analysis of asbestos and lung cancer: is better quality exposure assessment associated with steeper slopes of the exposure-response relationships? Environ Health Perspect 119(11):1547–1555, PMID: 21708512, https://doi.org/10.1289/ehp.1002879.
Linet MS, Schubauer-Berigan MK, Berrington de Gonzales A. 2020. Outcome assessment in epidemiologic studies of low-dose radiation exposure and cancer risks: sources, level of ascertainment, and misclassification. J Natl Cancer Inst Monogr 2020(56):154–175, https://doi.org/10.1093/jncimonographs/lgaa007.
Lipsitch M, Tchetgen E, Cohen T. 2010. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 21(3):383–388, PMID: 20335814, https://doi.org/10.1097/EDE.0b013e3181d61eeb.
Losilla JM, Oliveras I, Marin-Garcia JA, Vives J. 2018. Three risk of bias tools lead to opposite conclusions in observational research synthesis. J Clin Epidemiol 101:61–72, PMID: 29864541, https://doi.org/10.1016/j.jclinepi.2018.05.021.
Lubin JH, Hauptmann M, Blair A. 2018. Indirect adjustment of the effects of an exposure with multiple categories for an unmeasured confounder. Ann Epidemiol 28(11):801–807, PMID: 30297163, https://doi.org/10.1016/j.annepidem.2018.09.003.
Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. 2012. Industry sponsorship and research outcome. Cochrane Database Syst Rev 12:MR000033, PMID: 23235689, https://doi.org/10.1002/14651858.MR000033.pub2.
Michaels D. 2009. Addressing conflict in strategic literature reviews: disclosure is not enough. J Epidemiol Community Health 63(8):599–600, PMID: 19596838, https://doi.org/10.1136/jech.2009.089524.
Möhner M, Wendt A. 2017. A critical review of the relationship between occupational exposure to diesel emissions and lung cancer risk. Crit Rev Toxicol 47(3):185–224, PMID: 28322628, https://doi.org/10.1080/10408444.2016.1266598.
Morgan RL, Thayer KA, Santesso N, Holloway AC, Blain R, Eftim SE, et al. 2019. A risk of bias instrument for non-randomized studies of exposures: a users' guide to its application in the context of GRADE. Environ Int 122:168–184, PMID: 30473382, https://doi.org/10.1016/j.envint.2018.11.004.
National Research Council. 2014. *Review of EPA's Integrated Risk Information System (IRIS) Process*. Washington, DC: The National Academies Press.
Norris SL, Bero L. 2016. GRADE methods for guideline development: time to evolve? Ann Intern Med 165(11):810–811, PMID: 27654340, https://doi.org/10.7326/M16-1254.
NTP (National Toxicology Program). 2015. *Handbook for Preparing the Report on Carcinogens Monographs*. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjs3NXIyNLrAhXvgXIEHVJGB64QFjACegQIAxAB&url=https%3A%2F%2Fntp.niehs.nih.gov%2Fntp%2Froc%2Fhandbook%2Froc_handbook_508.pdf&usg=AOvVaw3mnQKMW8Qgtgg4actiEyOA [accessed 3 September 2020].
NTP. 2019. *OHAT (Office of Health Assessment and Translation) Handbook*. https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookdraftmarch2019.pdf [accessed 3 September 2020].
Olsson AC, Gustavsson P, Kromhout H, Peters S, Vermeulen R, Brüske I, et al. 2011. Exposure to diesel motor exhaust and lung cancer risk in a pooled analysis from case-control studies in Europe and Canada. Am J Respir Crit Care Med 183(7):941–948, PMID: 21037020, https://doi.org/10.1164/rccm.201006-0940OC.
Parker S. 2019. EPA staff defend NAAQS 'weight of evidence' from CASAC's criticism. InsideEPA, online edition. 8 November 2019. https://insideepa.com/daily-news/epa-staff-defend-naaqs-%E2%80%98weight-evidence%E2%80%99-casac%E2%80%99s-criticism [accessed 3 September 2020].
Pearce NE, Smith AH, Howard JK, Sheppard RA, Giles HJ, Teague CA. 1986. Non-Hodgkin's lymphoma and exposure to phenoxyherbicides, chlorophenols, fencing work, and meat works employment: a case-control study. Br J Ind Med 43(2):75–83, PMID: 3753879, https://doi.org/10.1136/oem.43.2.75.
Pearce N, Vandenbroucke JP, Lawlor DA. 2019. Causal inference in environmental epidemiology: old and new approaches. Epidemiology 30(3):311–316, PMID: 30789434, https://doi.org/10.1097/EDE.0000000000000987.
Qian F, Riddle MC, Wylie-Rosett J, Hu FB. 2020. Red and processed meats and health risks: how strong is the evidence? Diabetes Care 43(2):265–271, PMID: 31959642, https://doi.org/10.2337/dci19-0063.
Rooney AA, Cooper GS, Jahnke GD, Lam J, Morgan RL, Boyles AL, et al. 2016. How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. Environ Int 92–93:617–629, PMID: 26857180, https://doi.org/10.1016/j.envint.2016.01.005.
Rothman K, Greenland S. 2005. Causation and causal inference in epidemiology. Am J Public Health 95(suppl 1):S144–S150, PMID: 16030331, https://doi.org/10.2105/AJPH.2004.059204.
Samet JM, Chiu WA, Cogliano V, Jinot J, Kriebel D, Lunn RM, et al. 2020. The IARC Monographs: updated procedures for modern and transparent evidence synthesis in cancer hazard identification. J Natl Cancer Inst 112(1):30–37, PMID: 31498409, https://doi.org/10.1093/jnci/djz169.
Savitz DA, Wellenius GA, Trikalinos TA. 2019. 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 188(9):1581–1585, PMID: 31145434, https://doi.org/10.1093/aje/kwz131.
Schubauer-Berigan MK, Berrington de Gonzalez A, Cardis E, Laurier D, Lubin JH, Hauptmann M, et al. 2020. Evaluation of confounding and selection bias in epidemiologic studies of populations exposed to low-dose, high-energy photon radiation. J Natl Cancer Inst Monogr 2020(56):133–153, https://doi.org/10.1093/jncimonographs/lgaa008.
Siemiatycki J, Wacholder S, Dewar R, Cardis E, Greenwood C, Richardson L. 1988. Degree of confounding bias related to smoking, ethnic group, and socioeconomic status in estimates of the associations between occupation and cancer. J Occup Med 30(8):617–625, PMID: 3171718, https://doi.org/10.1097/00043764-198808000-00004.
Silverman DT, Samanic CM, Lubin JH, Blair AE, Stewart PA, Vermeulen R, et al. 2012. The Diesel Exhaust in Miners study: a nested case-control study of lung cancer and diesel exhaust. J Natl Cancer Inst 104(11):855–868, PMID: 22393209, https://doi.org/10.1093/jnci/djs034.

Singla VI, Sutton PM, Woodruff TJ. 2019. The Environmental Protection Agency Toxic Substances Control Act systematic review method may curtail science used to inform policies, with profound implications for public health. Am J Public Health 109(7):982–984, PMID: 31166745, https://doi.org/10.2105/AJPH.2019.305068.

Sørensen HT, Lash TL, Rothman KJ. 2006. Beyond randomized controlled trials: a critical comparison of trials with nonrandomized studies. Hepatology 44(5):1075–1082, PMID: 17058242, https://doi.org/10.1002/hep.21404.

Stang A. 2010. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 25(9):603–605, PMID: 20652370, https://doi.org/10.1007/s10654-010-9491-z.

Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. 2016. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 355:i4919, PMID: 27733354, https://doi.org/10.1136/bmj.i4919.

Vandenbroucke JP, Broadbent A, Pearce N. 2016. Causality and causal inference in epidemiology: the need for a pluralistic approach. Int J Epidemiol 45(6):1776–1786, PMID: 26800751, https://doi.org/10.1093/ije/dyv341.

VanderWeele TJ, Ding P. 2017. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med 167(4):268–274, PMID: 28693043, https://doi.org/10.7326/M16-2607.

Vermeulen R, Silverman DT, Garshick E, Vlaanderen J, Portengen L, Steenland K. 2014. Exposure-response estimates for diesel engine exhaust and lung cancer mortality based on data from three occupational cohorts. Environ Health Perspect 122(2):172–177, PMID: 24273233, https://doi.org/10.1289/ehp.1306880.

Vernooij RWM, Zeraatkar D, Han MA, El Dib R, Zworth M, Milio K, et al. 2019. Patterns of red and processed meat consumption and risk for cardiometabolic and cancer outcomes: a systematic review and meta-analysis of cohort studies. Ann Intern Med 171(10):732–741, PMID: 31569217, https://doi.org/10.7326/M19-1583.

Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters M, et al. 2008. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. In: *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Rockville, MD: Agency for Healthcare Research and Quality.

Vlaanderen J, Lan Q, Kromhout H, Rothman N, Vermeulen R. 2011. Occupational benzene exposure and the risk of lymphoma subtypes: a meta-analysis of cohort studies incorporating three study quality dimensions. Environ Health Perspect 119(2):159–167, PMID: 20880796, https://doi.org/10.1289/ehp.1002318.

Whaley P, Letcher RJ, Covaci A, Alcock R. 2016. Raising the standard of systematic reviews published in Environment International. Environ Int 97:274–276, PMID: 27567414, https://doi.org/10.1016/j.envint.2016.08.007.

Woodruff TJ, Sutton P. 2014. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environ Health Perspect 122(10):1007–1014, PMID: 24968373, https://doi.org/10.1289/ehp.1307175.

Zeraatkar D, Johnston BC, Bartoszko J, Cheung K, Bala MM, Valli C, et al. 2019. Effect of lower versus higher red meat intake on cardiometabolic and cancer outcomes: a systematic review of randomized trials. Ann Intern Med 171(10):721–731, PMID: 31569236, https://doi.org/10.7326/M19-0622.