Published in final edited form as:
Integr Environ Assess Manag. 2020 September ; 16(5): 718–728. doi:10.1002/ieam.4271.

SYSTEMATIC REVIEW AND WEIGHT OF EVIDENCE ARE INTEGRAL TO ECOLOGICAL AND HUMAN HEALTH ASSESSMENTS: THEY NEED AN INTEGRATED FRAMEWORK

Glenn Suter,
Office of Research and Development, Emeritus, U.S. Environmental Protection Agency, Cincinnati, Ohio, USA

Jennifer Nichols,
Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

Emma Lavoie,
Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC, USA

Susan Cormier
Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, Ohio, USA

Abstract

Scientific assessments synthesize the various results of scientific research for policy and decision making. Synthesizing evidence in environmental assessments can involve either or both of two systems: systematic review (SR) and weight of evidence (WoE). SR was developed to systematically assemble results of clinical trials to be combined by meta-analysis. Weight of evidence (WoE) approaches have evolved from jurisprudence to make inferences from diverse bodies of evidence in various fields. Our objectives are to describe the similarities and differences between SR and WoE and suggest how their best practices can be combined into a general framework that is applicable to human health and ecological assessments. Integrating SR and WoE is based on the recognition that two processes are required, assembling evidence and making an inference. SR is characterized by methodical literature searching, screening, and data extraction, originally for meta-analysis but now for various inferential methods. WoE is characterized by systematically relating heterogeneous evidence to considerations appropriate to the inference and making the inference by weighing the evidence. SR enables the unbiased assembly of evidence from literature, but methods for assembling other information must be considered as well. If only one type of quantitative study estimates the assessment endpoint, meta-analysis is appropriate for inference. Otherwise, the heterogeneous evidence must be weighed. A framework is presented that integrates best practices into a methodical assembly and weighing of evidence. A glossary of terms

Contact Information: Glenn Suter, suterpro@earthlink.net, 3720 Fallentree Lane, Cincinnati, Ohio 45236.

[1)]The authors declare that they have no financial or other conflicts of interest.

Publisher's Disclaimer: Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

for the combined practice and a history of the origins of SR and WoE are provided in supplemental material.

Keywords

Systematic review; Weight of evidence; Hill's criteria; meta-analysis; Evidence integration

INTRODUCTION

Frameworks for risk assessment of environmental contaminants often present a relatively simple process in which it is assumed that, following a problem formulation, a single exposure estimate is brought together with a single exposure-response relationship to characterize the risk. In practice, multiple pieces of evidence are often available to formulate the problem, estimate exposure, and derive exposure-response relationships. Synthesizing the evidence is an essential element of environmental assessment, but it is often not clear how best to perform that synthesis. When assembling information from the literature, you need some sort of Systematic Review (SR) to minimize bias, and when making inferences from a mixture of evidence, you need some sort of Weight of Evidence (WoE). But assessors with a typical training in environmental science, have only a general concept of WoE and little if any awareness of SR. In most environmental assessments, although not explicitly stated, elements of both SR and WoE are used to assemble and make inferences from evidence. To make assessments more transparent and defensible, we recommend that assessors consciously and deliberately integrate SR and WoE, as appropriate for their assessments.

When information is obtained from multiple studies (e.g., multiple rodent carcinogenicity tests) to answer an assessment question (e.g., did a chemical cause an observed cancer cluster?), each study may provide a piece of evidence (e.g., a cancer test that provides evidence of carcinogenicity), multiple pieces of evidence (e.g., a positive cancer test that also provides mechanistic evidence), or information that contributes to a complex piece of evidence (e.g., a positive cancer test in combination with exposure estimates provides evidence of causal sufficiency). Some organizing approach is desirable for selecting relevant information and for analyzing complex bodies of evidence. In human health and ecological assessments of environmental contaminants, two approaches are commonly applied for assembling and drawing inferences from multiple studies: systematic review (SR) and weight of evidence (WoE). SR and WoE have different histories, traditions, and approaches for synthesizing information. The differences are substantial enough that the European Food Safety Authority has separate guidance for WoE and SR (EFSA 2010, 2017).

SR and WoE practices have largely been distinct, and some have left the impression that assessors must choose one or the other. However, in some organizations and contexts, particularly in the USEPA, WoE and SR are evolving and have begun to overlap (NRC 2018; USEPA 2018). The premise of this paper is that the convergence should be encouraged and formalized, based on an understanding of what each practice offers environmental assessors. In this paper, we address three major questions. 1) What are the essential features of WoE

and SR, and what do they contribute? 2) What is the appropriate balance of pragmatism and consistent procedures? 3) How can the best features of WoE and SR be combined in a complimentary manner? Because terminology is not standardized and may be unclear to many readers, a glossary of SR and WoE terms is provided as supplemental material.

BACKGROUND

The WoE tradition derives from the ancient Greek goddess of justice, Themis. Her scales weigh the evidence on each side of an issue, and thereby she reaches a judgment. The metaphor of weighing evidence is so compelling that it is used to this day in jurisprudence, scholarship, business, and science to describe inferences from multiple pieces of evidence.

Inspired by Archie Cochrane (1972), the Cochrane Collaboration formalized SR in 1992 to provide reliable and consistent summaries of multiple randomized clinical trials of medical treatments. Since then, SR has been adapted to the social sciences (Campbell Collaboration), environmental management (Collaboration for Environmental Evidence (CEE)), animal testing (Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES)), health risks of environmental contaminants (Navigation Guide and Office of Health Assessment and Translation (OHAT)), and others. At the USEPA, aspects of SR are now a part of Integrated Risk Information System (IRIS) assessments (NRC 2018) and Toxic Substances Control Act (TSCA) risk evaluations (USEPA 2018).

Both SR and WoE have archetypal systems with their own histories (Supplemental material). The archetype for SR is the Cochrane Collaboration's handbook (Higgins and Green 2011), and adaptations for various fields refer to it. Until the Cochrane Collaboration, meta-analyses of clinical trials had been criticized for using incomplete or biased data sets. The Cochrane Collaboration greatly reduced bias by developing a process for methodically reviewing the potential input studies and documenting the review process. The original SR process consists of methodically searching the literature for relevant studies, screening the studies, assessing the risk of bias, and then performing meta-analysis on data from the retained studies. We refer to the many systems that use this approach (e.g., Cochrane, Campbell, CAMARADES, CEE, and EFSA) as Classic SR (Figure 1). Some attributes of Classic SR and Classic WoE are contrasted in Table 1.

The archetype for scientific WoE is A.B. Hill's (1965) codification of the method used by the U.S. Surgeon General's Commission (USDHEW 1964) to determine that the association of smoking with lung cancer is causal. The Commission and Hill recognized that the issue could not be resolved by statistics, so they weighed the available evidence in terms of a set of causal considerations. Those considerations are still frequently used and adapted when weighing evidence of health and environmental causation. Hill-style WoE, such as those used by the USEPA's Integrated Science Assessments (ISAs), that identify hazards of air pollutants in the U.S., methodically review the literature and present a description of the weight of evidence for each Hill consideration and for the conclusion (Owens et al. 2017; USEPA 2014). This approach is Classic WoE (Figure 1). WoE, however, covers a wide range

of approaches that may use narratives, weight categories, or numerical methods (Linkov et al. 2009; Martin et al. 2018; Rhomberg et al. 2013; USEPA 2016; Weed 2005).

As SR and WoE have been adapted to questions in environmental assessment, each has adopted some features of the other. Our exemplar of post-Classical practices in SR for health effects is OHAT (2015). SR practices such as OHAT can incorporate animal tests, human observational studies, mechanistic studies (to a limited extent), and, in rare cases, human tests. Meta-analysis may be applied to each type of study, but a WoE technique must be used to evaluate qualitative properties and to make the inference concerning the occurrence of effects. For example, OHAT derives a level of evidence (equivalent to a weight of evidence) for each type of evidence and then integrates them to assign a low, moderate or high hazard conclusion. On the WoE side, systems are beginning to recommend more methodical literature reviews and greater process transparency. Our exemplar of such post-Classical WoE practices is the guidelines for weight of evidence in ecological assessments developed by the USEPA's Risk Assessment Forum (hereafter referred to as Eco WoE) (USEPA 2016).

Key message: To understand the utility of a method, one must understand the often-narrow purpose for which it was developed. Knowing that original purpose, the method can be adapted to a new purpose.

PROCEDURAL PRACTICALITY VERSUS CONSISTENCY

The detail with which methods are prescribed varies among assessment approaches. SR, in general, is highly prescribed, and that feature has been considered a major strength. The detailed methods (e.g., Figure 2) facilitate consistency, transparency and reproducibility. The degree to which WoE methods are prescribed is variable but is seldom as detailed. Prior to recent changes, the EPA's application of Classic WoE in IRIS assessments was criticized for being ill-defined and inconsistent (NRC 2011). This critique led to greater structure and consistency in some Classic WoE applications, including adoption of elements of SR (Owens et al. 2017, NRC 2018). However, the appropriate degree to which methods are prescribed is controversial. For example, separate reviews of the ISAs recommended different directions. An industry-funded review recommended greater specification of the procedures for study selection, study quality evaluation, etc. (Goodman et al. 2013). In contrast, a Clean Air Science Advisory Committee, which provides advice on the USEPA National Ambient Air Quality Standards, expressed "concern about applying strict evaluation criteria to various studies" (CASAC 2015).

The degree to which procedural guidance is prescriptive can be informed by the following considerations.

1. *Time and resource constraints*. If sufficient time and staff are available, it is possible to apply and document detailed procedures for identifying, obtaining and evaluating information, performing analyses, and reporting results. If information must be assembled and conclusions drawn in an emergency situation, expert consensus workshops or even semi-structured phone surveys may be employed (Donnelly 2018, Cormier 2008). Most assessment approaches will fall in-between.

2. *Scope of the assessments*. Literature-based reviews of a single study type, like Cochrane reviews of clinical trials, follow detailed guidance. At the other extreme, the range of assessments in agencies like the USEPA is very broad, so Agency-wide guidelines like the Eco WoE are less detailed to allow for a greater scope of questions to be answered.

3. *Technical Feasibility.* A prescribed method may or may not be feasible in practice. For example, a prescribed method may use techniques that require types of data and information that are not routinely available, uncommon expertise or training, or the use of tools that are not generally available or difficult to implement. If so, it may be presented as an ideal, but other methods should be allowed.

4. *Need to adhere to procedural consistency.* Highly inconsistent procedures may lead to inconsistent quality of results and to charges that they are arbitrary and capricious. On the other hand, requirements to adhere to a procedure that does not fit the situation, can lead to errors or inability to inform a decision. Procedures must be flexible enough to address the range of questions, information, time, and resources.

Because SR procedures are detailed, we have found that they tend to provide consistency, transparency, and defined process quality (see, however, a review by Ioannidis (2016) critiquing the quality of SRs). In contrast, Hill provided no guidance on how his considerations were to be applied, leaving WoE procedurally diverse and often in Classic WoE, rather informal. Although some WoE applications have been clearly structured, some other WoE applications have been criticized for inconsistency and lack of transparency (NRC 2011). In our opinion, when a narrowly defined type of assessment is repeatedly encountered, guidance for WoE can benefit from methods that are as prescribed and explicit as Classic SR methods. One example of such optimization is the ISAs that have methodical literature searching and screening methods but employ WoE considerations in drawing conclusions on the causal nature of reported air pollutant-induced effects (Owens et al. 2017; USEPA 2014).

Key message: If the question to be answered and the relevant information are consistent from one assessment to another, a detailed procedure can be useful. Inconsistent cases, however, call for less procedural detail and more expert judgment.

THE BASIC STEPS

Systems for answering questions by combining evidence require at least two steps, assembling information that provides evidence and making inferences from the evidence. The assembly of information should generate a reasonably complete set of relevant and reliable information that provides evidence concerning the alternative hypotheses. The inference step should apply appropriate qualitative and quantitative methods to derive a conclusion with associated expressions of confidence. Classic SR and WoE differ in both steps, and an integrated system must include both to assure the appropriateness of the information that has been obtained and the inferences that are performed.

Although SR and WoE have each been described as complete and unitary assessment practices, information assembly and inference have been recognized as distinct. A recent review of evidence synthesis approaches distinguished "the process of initial identification, assembly, and abstraction of relevant data—which we refer to as systematic review—from the process of evaluating the support they may or may not give to causal inference—which we refer to as "the integration of and weighing of the evidence" and the whole process from scoping to conclusions as a "WoE framework"" (Rhomberg et al. 2013). Similarly, a National Research Council (NRC) committee defined "evidence integration" (their alternative term for WoE, which they consider vague) as "the process that occurs after the completion of systematic reviews" (NRC 2014). Distinguishing these steps highlights the fact that SR is inherently a method for assembling information that can support inference, and WoE is inherently an inferential approach that evaluates information from any source.

Key message: Combining evidence to answer a question requires two broad steps, assembling the evidence and making the inference.

INTEGRATING WOE AND SR PRACTICES

The popularity of WoE appears to have arisen from two features: 1) the ubiquity of the need to combine different pieces and types of evidence to answer a question and 2) the intuitive appeal of the idea that the number of pieces of evidence in favor of each hypothesis and the weightiness (determined by the relevance, reliability, or other properties) of each piece should determine the result. The use of Hill's approach defines a third feature of Classic WoE, the use of considerations as a basis for interpreting bodies of evidence. The continuing appeal of the WoE concept in assessing risks to health and the environment is illustrated by the fact that WoE is mandated in the recently revised U.S. law regulating the marketing of industrial chemicals (USC 2016) and four of nine recent pieces of European environmental and food safety legislation (Agerstrand and Beronium 2016).

The growing popularity of SR appears to have arisen from the success of Cochrane SRs in making the practice of medicine more evidence based and thus more effective. A good Cochrane review is a thing of inferential beauty, and other fields have tried to emulate its consistent logical structure, clarity, defensibility, and transparency. Simply put, a Classic SR is meta-analysis of data obtained by systematically searching and screening published studies. In any assessment that relies on a literature review, a SR can, by minimizing bias and increasing transparency, forestall the criticism that the data selection processes are biased. Classic SRs are appropriate when multiple quantitative studies of the same type (e.g., same agent, tested taxon, and endpoint) are applied to a well-defined question. "If the question structure can be specified in such a way that a particular primary research study design can be envisaged that would answer the question, then it is likely that a systematic review would be appropriate" (EFSA 2010). Because the studies are of the same type, meta-analysis or at least vote counting (e.g., 8 of 10 studies showed the effect) can be used to judge hypotheses. However, evidence for effects of pollutants on health rarely meet the standard set by Cochrane SRs of clinical trials. As a result, SRs for environmental toxic effects such as OHAT include elements of qualitative WoE along with meta-analysis.

The distinction between Classic WoE and SR is the basis for EFSA's provision of guidance for both SR and WoE (EFSA 2010, 2017). Although no guidance document explicitly integrates WoE and SR, some such as OHAT and Eco WoE have elements of both. OHAT includes an evidence integration step that serves the function of WoE, and Eco WoE recommends SR for literature reviews. Our goal is to make the integration of SR and WoE explicit and define a general framework for making reliable inferences from systematically derived sets of information. Achieving that goal requires careful consideration of the evidence assembly and inference tasks. In the following sections, we consider best practices from both WoE and SR, constraints on their use, and opportunities for integration.

Key message: WoE is a conventional inferential process that is commonly demanded by decision makers. SR is an approach for methodically generating an unbiased body of evidence from a literature review to support any inferential process.

Collecting and selecting information

Once the question to be answered is defined and it is determined that multiple sources of information must be combined, the information is assembled from the available sources. SR procedures and tools are broadly applicable for literature searching, screening the search results, extracting data, and documenting the process (Figure 2). Although WoE traditionally has not been particularly concerned with methods for literature reviews, many WoE analyses have been methodical in reviewing and screening the literature without reference to SR procedures (e.g., USEPA 2012). Recent WoE guidance has addressed the need for more rigorous and transparent literature reviews (EFSA 2017; USEPA 2016). A challenge to SR is the time and effort that is often required when the literature is large and varied and the review process is complex and must be documented (e.g., documenting which studies were screened out from the search results and on what basis). The tools developed for SR that improve efficiency and incorporate automation of searching and screening (e.g., Distiller and Swift tools) can make SR more attractive to WoE practitioners. An alternative approach to achieve efficiency is rapid review, which simplifies or omits steps in the SR process in the interest of providing timely input to decision making (Dobbins 2017).

WoE analyses often use information sources and types of studies not found in the literature such as unpublished data sets (e.g., state water quality records), model results (e.g., quantitative structure-activity relationships), or studies performed for the assessment (e.g., contaminated site characterizations). ECHA (2018) guidance for WoE lists 5 information sources other than published literature to be consulted for human health assessments. An evidence assembly process called an SR might go beyond systematic literature review and, like WoE, obtain information in other ways, but that makes the SR concept less clear. Although approaches to obtaining information such as contacting experts for unpublished data are never as systematic as a conventional SR, they can be done in a planned and transparent manner.

The method for assembling evidence should be fit for purpose. We recommend that WoE practitioners be methodical when assembling information from the literature in order to be more transparent and unbiased, and that SR practitioners consider whether other types and sources of information should be used to better address the question of concern. We also

suggest that additional guidance on how to identify, obtain, and assemble heterogeneous information in a transparent and unbiased manner could be useful.

Key message: Classic SR shows how to methodically search the literature and extract information for a well-defined question that is addressed in the literature resulting in a body of evidence that should be defensibly complete and unbiased. Classic WoE obtains evidence from various sources using less methodical means than SR but can be improved by explicitly stating how information is selected.

Inference and implications

Inference from evidence determines what should be believed and with how much confidence given the relevance and reliability of the assembled information. Inference in Classic SR is conceptually straight-forward. The assembled data sets are used to perform meta-analyses to test the hypothesis that an intervention is efficacious or that an exposure poses a particular hazard. The mean and confidence interval of the combined studies can be interpreted as best estimates of whether a hypothesis is believable and, in statistical terms, how strongly it should be believed, given a reasonably similar and unbiased set of studies (Borenstein et al. 2009). Although conventional frequentist meta-analyses are common, meta-regression, Bayesian analysis, and other statistics are also options.

Because, in Classic SR, the assessment question is answered by statistically combining the study results, only one type of study is used (EFSA 2010). One should consider if this is appropriate given the question and the available evidence. The clinical trials evaluated by Cochrane SRs are ideal for meta-analysis, because they directly answer the question of efficacy in humans and are inherently causal because they are experiments. The studies used for assessment of human environmental risks seldom achieve this ideal. However, ecological studies may. For example, pond mesocosms represent ponds in agricultural areas, and multiple mesocosm studies of an agrochemical may be conducted, so meta-analysis can be used for the inference (USEPA 2016, Moore et al. 2017, Giddings et al. 2018). To use meta-analysis to answer a human or ecological assessment question, four issues should be considered.

1. The test endpoint or other measurement endpoint of the studies should be equivalent to the assessment endpoint. For example, a rat carcinogenicity test endpoint may be considered to ascertain, with more or less confidence, whether a chemical is a rodent carcinogen, a mammalian carcinogen, or a human carcinogen.
2. A causal relationship must be explicitly defined and either demonstrated or stated as being assumed. Experiments are inherently causal, but it is very seldom possible to dose humans with non-therapeutic chemicals, and laboratory-based ecotoxicological studies seldom test all the species and responses of interest. In addition, meta-analyses of epidemiological and other observational studies are associational and require additional inference. Therefore, any SR using observational information depends on WoE to establish causation in addition to any application of meta-analysis.

3. The plausibility of the studies should be considered. For example, mechanisms of action are often uncertain or incompletely known, but one or more mechanisms may be plausible given a general understanding of chemistry, physiology, and natural history. In the extreme, no plausible mechanism is known, or general theory indicates an implausible mechanism. For example, SRs of clinical trials of homeopathic therapies are unconvincing in the absence of a plausible mechanism (Mathie et al. 2017).

4. The reliability of the studies should be assessed as well as their potential to answer the assessment question. Studies may pass screening criteria but be so weak or biased that they contribute little to answering the question. Also, the body of studies may be small or may poorly comply with statistical assumptions.

Each of these four issues with respect to using meta-analysis may be resolved by either a policy judgment or inference by weighing evidence.

Because heterogeneous bodies of evidence cannot be combined statistically, the inference ultimately must evaluate the body of evidence qualitatively, guided by prescribed considerations. Even if the individual types of evidence (e.g., fish acute lethality tests) are evaluated by meta-analysis or equivalent statistics, when types are combined, statistical integration is inappropriate. Hill (1965) wrote that his considerations are “aspects” of potentially causal associations that we should “especially consider” and that provide “nine different viewpoints, from all of which we should study associations.” They provide multiple viewpoints by including characteristics of causation (e.g., temporality), types of evidence (e.g., experiment), and properties of evidence (e.g., strength) (Cormier et al. 2010, USEPA 2016). He did not, however, distinguish characteristics of causation, types of evidence, or properties of evidence or acknowledge that different considerations have logically and functionally distinct roles in inference. Types of evidence tell us how to organize the evidence (e.g., it is an acute lethality test). Properties tell us how much weight we should give the evidence (e.g., it is moderately relevant and highly reliable). And, the causal considerations or other implications tell us what the evidence means with respect to the hypotheses (e.g., dying fish in the stream exhibit the same symptoms as fish in the acute test and therefore, the cause could be the same). Although Hill’s nine considerations were able to demonstrate that the association of lung cancer with cigarette smoking was almost certainly causal, WoE for more complex or difficult cases requires that practitioners use more complete and logically organized sets of considerations than Hill suggested. This more rigorous approach to WoE (USEPA 2016) is analogous to the rigorous approach to literature searching and screening provided by SR.

The implications of evidence are perhaps the most important and least acknowledged aspect of evidence. The implication of a piece of evidence is a statement of how the evidence can influence a hypothesis. Information without an implication is not evidence. One useful formulation of implications is that the evidence does or does not exhibit a characteristic of causation or of another hypothesized attribute (e.g., protection, contaminant of concern, impairment, and remediation) (USEPA 2016). The intended and appropriate use of the implications is to aid scientific interpretation by inference to the best explanation of the

evidence (Lipton 2004). The importance of implications to the future of environmental toxicology is discussed in Text Box 1.

An important feature of implications of evidence is that they are not independent. The characteristics of a system being assessed are indicated by the implications interacting in a logically concordant manner (Rhomberg 2015; USEPA 2016). Each implication influences others and together they may create a coherent body of evidence that is more than the evaluation of individual considerations. Inference from multiple pieces of evidence is akin to assembling a jigsaw puzzle; all the pieces should fit together to form the picture (Rosenbaum 2017). For example, in the sediment quality triad, chemical analyses can imply that one or more chemicals are present in potentially toxic amounts, toxicity tests can imply that the mixture of chemicals is toxic, and biological surveys can imply that the community is impaired (Dagnino et al. 2008). If no chemical occurs at a potentially toxic concentration, but the toxicity test is positive, the community is impaired, and the evidence is all relevant and reliable, we might infer that the chemical causing toxic effects was not measured or the mixture is toxic even though no constituent alone is toxic. Only interpretation can address unexpected or discordant results in a body of technically reliable information. Thus, analysis of alternative interpretations can be a powerful inferential tool when evidence with interacting implications and with different properties (e.g., relevance, reliability and strength) give the evidence different weights. As the U.S. Army Corps of Engineers stated concerning weighing evidence in sediment assessments, we must consider “the conclusions that can be directly drawn from it either supporting or opposing some hypothesis (“what the evidence says”) and as a result of meta-data about the evidence that suggest how much/little we should let it influence our overall conclusions (“how strongly the evidence says it” or “how much we believe it”)” (Bates et al. 2018).

Key message: Inference from multiple studies may be performed by meta-analysis if the requirements are met. WoE can be applied to any set of evidence, but it requires logical analysis. In particular, while a Classic SR analyzes a single type of study that answers the question, any type of WoE addresses diverse pieces of evidence and the implications of the evidence for each hypothesis.

Organizing the body of evidence for WoE

Evidence in environmental assessments is usually organized by types, but we have observed that, except in simple cases, it is useful to also organize evidence in terms of explanatory implications and to weigh the categories of evidence. This approach has developed in ecological assessment, because the diversity of evidence in ecological assessments is often greater than in human health assessments (i.e., you cannot sample and manipulate people the way you can nonhuman populations and communities). Relevant evidence of effects is also more readily generated in ecological cases, because the effects of concern are more often large enough to be readily observed and measured, nonhuman organisms are more exposed to contaminated environments, and some non-human taxa will be more sensitive (Suter 2007). As a simple but illustrative example of how to organize diverse evidence, Table 2 summarizes a causal assessment of mass mortalities of tundra swans that weights the evidence using qualitative ratings for six causal characteristics and three collective properties

of the body of evidence and then combines the weights into an overall weight of evidence for the hypothesis that lead toxicity is the cause. (This example also illustrates how some assessments such as contaminated site assessments rely on evidence generated for the case rather than from literature reviews, so SR may not be needed.)

We have found that the implications of evidence for human health hazards and their use to weigh evidence are not explicitly formalized. However, in the USEPA, there is a move toward more formalization using causal determination summary tables (ISAs) and evidence profile tables (IRIS) based on adaptations of Hill's considerations.

Weighing evidence for sets of explanatory implications, such as the causal characteristics or the sediment quality triad, is a powerful interpretation technique but not the only one. If the assessment involves a causal chain or network rather than a direct cause-effect relationship, the evidence may be organized in terms of a conceptual model of the system (Figure 3). Associating evidence with links in a network model can illustrate mechanistic causal relationships within the body of evidence. When causation is unclear because the evidence is not concordant, reorganizing the model to illustrate a better understanding of causal relationships can make the evidence concordant or show what additional evidence is needed (Norton and Schofield 2018; Suter and O'Farrell 2008; USEPA 2010). A related approach is the weighing of evidence for adverse outcome pathways which are conceptual models of chemically agnostic mechanisms of toxic action (OECD 2016). Graph theory uses directed acyclic graphs (i.e., network models with unidirectional arrows and no feedback loops) to analyze causal relationships, but the analyses depend on knowing the actual structure of the dependencies (Pearl and Mackenzie 2018) which requires WoE. Cox provides a WoE approach to assessing the causality of dependencies in epidemiological causal networks, based on "updated Hill's considerations" (Cox 2018).

Interpretation of evidence (as opposed to only amassing weights) has the advantages of directly using the assessors' knowledge, but it is susceptible to the narrative fallacy. That is, people tend to force information into a story and to accept explanations that provide a good story. To minimize the narrative fallacy, it helps to avoid traditional narrative WoE and to provide an a priori method based on the implications of the evidence, explanatory structures such as causal networks, and weighting of the properties of the evidence. However, it is wise to avoid formal weighting systems that provide ratings but no understanding or that could falsely imply mathematical rigor by adding numerical ratings.

Key message: Diverse bodies of evidence should be organized with respect to two characteristics, types of evidence and implication of the evidence. Standard implications such as the characteristics or causation are desirable to provide consistency, but association of evidence with links in a causal network is also useful.

Weighting and rating evidence

The WoE concept implies assigning weights to the pieces of evidence and then weighing the accumulated body of evidence for each hypothesis. In practice, the Classic weighing of evidence has expressed weighting in narratives for each consideration, but explicit weighting, by evaluating the evidence and assigning ratings, has become common

(Rhomberg et al. 2013; Susser 1986; USEPA 2016). WoE ratings capture relevance, reliability, and strength of the evidence as a basis for inferring which hypothesis is best supported by the evidence. Classic SR systems typically rate the quality of the studies, as in the use of the GRADE system to determine the appropriate confidence in results of the meta-analysis or other inference (Higgins and Green 2011). A separate step is not needed for WoE, because the weight of a body of evidence generated by the inference also expresses confidence in a hypothesis (USEPA 2016).

Weighing evidence can serve two purposes: supporting inference and expressing confidence. That is, one might accept the hypothesis with the weightiest body of evidence, or one might make the inference by another method and use evidential weight to express the degree of confidence in the result. WoE systems have used weighting for both inference and determining confidence, but SR systems typically use rating only to derive a confidence rating for the collection of studies. An example of the latter is the use of the GRADE system in Cochrane SRs. In either case, weighting is done first for individual pieces of evidence or studies and then for collective properties of bodies of evidence (Table 2). The overall weight for a body of evidence expresses the confidence in the result from a meta-analysis or an inference from weighing evidence (NRC 2018, OHAT 2015, USEPA 2016).

Key message: Classic WoE provides narrative weighting organized in terms of considerations. However, it is advantageous to identify properties of evidence to be weighted and provide the weights as an explicit rating system.

A FRAMEWORK FOR INTEGRATING WOE AND SR

To create an integrated practice of evidence synthesis, by combining SR and WoE based on the emphases implied by their names. SR is primarily about reviewing the literature (i.e., searching, screening, and data extraction), and WoE is about inference from evidence (Table 1). Because they have originated in different domains, there is no inherent conflict between SR and WoE, and they can be formally combined. We have integrated the various practices into a framework (diagramed in Figure 4) to methodically search the literature, consult other sources of information, and generate needed evidence. Then, if a single type of study can answer the question without considering other information, the question would be answered by performing a meta-analysis. Otherwise, as is almost always the case in environmental assessments, all relevant types of evidence should be weighed. As part of evidence derivation, meta-analysis may also be used to combine studies within a type, but WoE must still be used to make inferences across types. In either case, the results would be presented along with an expression of confidence. In our experience, many WoE practitioners often review the published literature methodically, although perhaps not with the strict procedures of an SR. SR practitioners in environmental health, to the extent that they are combining multiple types of studies, are weighing evidence, often under the term “evidence integration” (OHAT 2015, US EPA 2018). However, they may not perform a formal weighting and weighing of the evidence which could improve consistent treatment of information and transparency.

Classic SR is adequate when a single type of study, which is found in the literature, can answer the question via meta-analysis. WoE without a systematic literature review is adequate when published studies are not a significant source of information. For example, targeted laboratory and field studies are conducted specifically to meet the needs of a contaminated site assessment, an effluent permit, or an investigation of a disease cluster or wildlife mass mortality (Table 2). Otherwise, a process that can meet the needs of various assessments, like the framework in Figure 4, should be considered. A descriptive name for the merged practice is Methodical Assembly and Weighing of Evidence (MAWE). This process is similar to the USEPA (2016) approach for WoE to derive quantities. It can derive either a quality or a quantity, because meta-analysis can derive estimates or test hypotheses and weighing evidence can identify the best-supported value or the best supported hypothesis.

Assessors need not decide whether they practice SR or WoE. As has been highlighted, practices of SR and WoE overlap, and often the decision to name an assessment process SR or WoE results from differences in experience. The methods for assembling evidence and making inferences from the evidence should be chosen to suit to the purpose and should be performed in ways that are appropriately organized, clear, and defensible. We suggest that, whatever the method, and whatever you call it, assessors should follow scientific standards of transparency concerning the reasons for choosing and implementing their methods. By focusing on the needs of the assessment rather than traditional practices, they can defensibly support evidence-based decision making.

Key message: Assessors have been confronted by an apparent choice between SR and WoE when synthesizing information to answer assessment questions. It’s a false choice; they can be combined. We propose a framework and best practices that integrate WoE and SR.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

2) Financial support was limited to the authors’ salaries. 3) This manuscript has benefitted from comments and suggestions by Tina Bahadori, David Bussard, Annette Gatchett, Jason Lambert, Kate Schofield, and Kris Thayer.

Data Accessibility Statement:

No data or calculation tools were used in this paper.

REFERENCES

Agerstrand M, Beronium A. 2016 Weight of evidence evaluation and systematic review in EU chemical risk assessment: Foundation is laid but guidance is needed. Environ Int 92–93:590–596.

Bates ME, Massey OC, Wood MD. 2018 Weight-of evidence concepts: Introduction and application to sediment management. Washington (D.C.): U.S. Army Corps of Engineers ERDC/EL SR-18-1.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2009 Introduction to meta-analysis. Hoboken, NJ: Wiley Pub.

[CASAC] Clean Air Act Science Advisory Committee. 2015 CASAC review of the EPA's integrated science assessment for oxides of nitrogen – health criteria (second external review draft – January 2015). Washington (DC): U.S. Environmental Protection Agency.

Cochrane AL. 1972 Effectiveness and efficiency: Random reflections on health services. UK: The Newfield Provincial Hospitals Trust.

Cormier SM 2008 A synopsis of immediate and deliberate environmental assessments In: Linkov I, Ferguson EA, Magar VS, editors. Real-time and deliberative decision making. Dordrecht (NL): Springer p. 21–29.

Cormier SM, Suter GW II, Norton SB. 2010 Causal characteristics for ecoepidemiology. Hum Ecol Risk Assess 16:53–73.

Cox LA Jr. 2018 Modernizing the Bradford Hill criteria for assessing causal relationships in observational data. Crit Rev Toxicol 48: 682–712. [PubMed: 30433840]

Dagnino A, Sforzini S, Dondero F, Fenoglio S, Bona E, Jensen J, et al. 2008 A "weight of evidence" approach for the integration of environmental "triad" data to assess ecological risk and biological vulnerability. Integr Environ Assess Manag 4:314–326. [PubMed: 18393577]

Dobbins M 2017 Rapid review guidebook, steps for conducting a rapid review. Hamilton (Ontario): National Collaborating Centre for Methods and Tools.

Donnelly CA, Boyd I, Campbell P, Craig C, Vallance P, Walport M, Whitty CJM, Woods E, Wormald C. 2018 Four principles for synthesizing evidence. Nature 558: 361–364. [PubMed: 29925978]

[ECHA] European Chemicals Agency. 2018 Weight of evidence. Accessed 10/6/2019 https://echa.europa.eu/support/registration/how-to-avoid-unnecessary-testing-on-animals/weight-of-evidence

[EFSA] European Food Safety Authority. 2010 Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA J 8:1–90.

[EFSA] European Food Safety Authority. 2017 Guidance on the use of the weight of evidence approach in scientific assessments. EFSA J 15:1–69.

Giddings JM, Campana D, Nair S, Brain R. 2018 Data quality scoring system for microcosm and mesocosm studies used to derive a level of concern for Atrazine. Integr Environ Assess Manag 14:489–497. [PubMed: 29663627]

Goodman JE, Prueitt RL, Sax SN, Bailey LA, Rhomberg LR. 2013 Evaluation of the causal framework used for setting national ambient air quality standards. Crit Rev Toxicol 43:829–849. [PubMed: 24090029]

Higgins J, Green S. 2011 Cochrane handbook for systematic reviews of interventions. Version 5.1.0. Cambridge (UK):The Cochrane Collaboration.

Hill AB. 1965 The environment and disease: Association or causation? Proc R Soc Med 58:295–300. [PubMed: 14283879]

Ioannidis JPA. 2016 The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 94:485–514. [PubMed: 27620683]

Linkov I, Loney D, Cormier S, Satterstrom FK, Bridges T. 2009 Weight-of-evidence evaluation in environmental assessment: Review of qualitative and quantitative approaches. Sci Total Environ 407:5199–5205. [PubMed: 19619890]

Lipton P 2004 Inference to the best explanation. New York (NY): Rutledge.

Martin P, Bladier C, Meek B, Bruyere O, Feinblatt E, Touvier M, et al. 2018 Weight of evidence for hazard identification: A critical review of the literature. Environ Health Perspect 126:076001-1–076001-15. [PubMed: 30024384]

Mathie RT, Ramparsad N, Legg LA, Clausen J, Moss S, Davidson JRT, Messow C-M, McConnachie A. 2017 Randomized, double-blind, placebo-controlled trials of non-individualised homeopathic treatment: systematic review and meta-analysis. Sys Rev 6:63 10.1186/s13643-017-0445-3.

Meek M, Palermo C, Bachman A, North C, Lewis R. 2014 Mode of action human relevance (species concordance) framework: evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. J Appl Toxicol 34:595–606. [PubMed: 24777878]

Moore D, Greer C, Manning G, Wooding K, Beckett K, Brain R, Marshall G. 2017 A weight-of-evidence approach for deriving a level of concern for atrazine that is protective of aquatic plant communities. Integr Environ Assess Manag 13:686–701. [PubMed: 27862949]

Norton SB, Schofield CL. 2018 Conceptual model diagrams as evidence scaffolds for environmental assessment and management. Freshwater Sci 36:231–239.

[NRC] National Research Council. 2011 Review of the Environmental Protection Agency's draft IRIS assessment of formaldehyde. Washington (DC): National Academies Press.

[NRC] National Research Council. 2014 Review of EPA's integrated risk information system (IRIS) process. Washington (DC): National Academies Press.

[NRC] National Research Council. 2017 Using 21st century science to improve risk-related evaluations. Washington (DC): National Academies Press.

[NRC] National Research Council. 2018 Progress toward transforming the integrated risk information system (IRIS) program: a 2018 evaluation. Washington (DC): National Academies Press.

[OECD] Organization for Economic Cooperation and Development. 2016 Users' handbook supplement to the guidance document for developing and assessing adverse outcome pathways. Paris (France): OECD.

[OHAT] Office of Health Assessment and Translation. 2015 Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. Research Triangle Park (NC): National Toxicology Program.

Owens EO, Patel MM, Kirrane E, Long TC, Brown JH, Cote I, et al. 2017 Framework for assessing causality of air pollution-related health effects for reviews of the national ambient air quality standards. Reg Toxicol Pharmacol 88:332–337.

Pearl J, Mackenzie D. 2018 The book of why: The new science of cause and effect. New York (NY): Basic Books.

Rhomberg L 2015 Hypothesis-based weight of evidence: an approach to assessing causation and its application to regulatory toxicology. Risk Anal 35:1114–1124. [PubMed: 24724710]

Rhomberg LR, Goodman JE, Bailey LA, Prueitt RL, Beck NB, Bevan C, et al. 2013 A survey of frameworks for best practices in weight-of-evidence analyses. Crit Rev Toxicol 43:753–784. [PubMed: 24040995]

Rosenbaum PR. 2017 Observation and experiment: An introduction to causal inference. Cambridge (MA): Harvard U Press.

Smith MT, Guyton KZ, Gibbons CF, Fritz JM, Portier CJ, Rusyn I, et al. 2016 Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. Environ Health Perspect 124:713–721. [PubMed: 26600562]

Susser M 1986 Rules of inference in epidemiology. Regul Toxicol Pharmacol 6:116–186. [PubMed: 2941827]

Suter GW II. 2007 Ecological risk assessment. Boca Raton (FL): CRC Press.

Suter GW II, O'Farrell TP. 2008 Analysis of the causes of a decline in the San Joaquin kit fox population on the Elk Hills, Naval Petroleum Reserve #1, California. Cincinnati (OH): U.S. Environmental Protection Agency EPA/600/R-08/130.

URS Greiner Inc, CH2M Hill. 2011 Remedial investigation report for the Coeur d'Alane Basin remedial investigation/feasibility study. Seattle (WA): U.S. Environmental Protection Agency, Region 10.

[USC] United States Congress. 2016 Frank R. Lautenberg chemical safety for the 21st century act. Washington (DC): Congress of the United States.

[USDHEW] US Department of Health Education and Welfare. 1964 Smoking and health: report of the advisory committee to the surgeon general. Washington (DC).

[USEPA] United States Environmental Protection Agency. 2010 An iterative approach for identifying the causes of reduced benthic macroinvertebrate diversity in the Willimantic River, Connecticut. Cincinnati (OH): USEPA EPA/600/R-08/144.

[USEPA] United States Environmental Protection Agency. 2012 EPA's reanalysis of key issues related to dioxin toxicity and response to NAS comments, volume 1 Washington (DC): USEPA EPA/600/R-10/038F.

[USEPA] United States Environmental Protection Agency. 2014 Welfare risk and exposure assessment for ozone, final. Washington (DC): USEPA Office of Air and Radiation.

[USEPA] United States Environmental Protection Agency. 2016 Weight of evidence in ecological assessment. Washington (DC): USEPA Risk Assessment Forum EPA/100/R-16/001.

[USEPA] United States Environmental Protection Agency. 2018 Application of systematic review in TSCA risk evaluations. Washington (DC): USEPA EPA Document 740-P1-8001.

Weed DL. 2005 Weight of evidence: A review of concept and methods. Risk Anal 25:1545–1557. [PubMed: 16506981]

Text Box 1.

Interpreting the implications of suborganismal data

The trend away from whole-animal testing and toward alternative methods such as modeling, in vitro testing, and omics, increases the need to weigh the evidence, because individual in vitro studies do not estimate conventional assessment endpoints (NRC 2017). They can be used to screen mechanisms, but a full risk assessment based on in vitro tests requires results of multiple different tests that cannot be assessed independently. Sub-organismal tests constitute evidence of endpoint effects such as health outcomes only when they are shown to have mechanistic implications. For example, an adverse outcome pathway may be convincing if it is supported by evidence of biomarkers, symptoms, and endpoint effects from animal or human data for the same or similar chemicals. Each of these types of evidence have independent implications as to exposure, intermediate events and an adverse outcome. These types of evidence have been incorporated in causal WoE based on Hill's considerations (Meek et al. 2014). However, the characteristics of carcinogenicity provide a more promising model of bases for mechanistic WoE (Smith et al. 2016). They provide a consistent set of implications of evidence for carcinogenicity (e.g., is genotoxic, induces epigenetic alterations, or induces chronic inflammation).

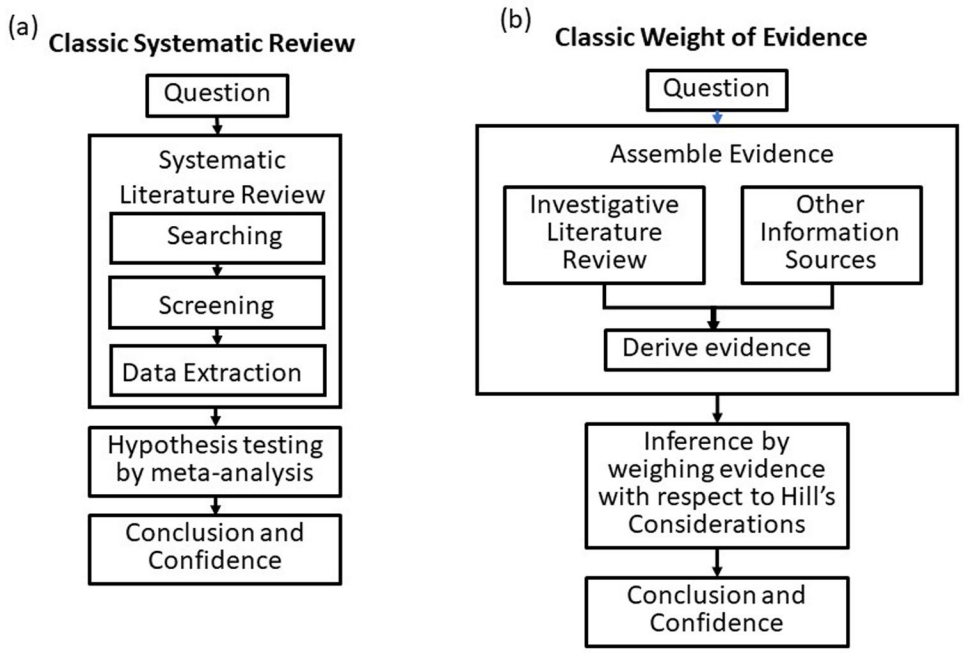

Figure 1.

Flow diagrams illustrating the similarities and differences between Classic Systematic Review (SR) (a) and Classic Weight of Evidence (WoE) (b). Classic SR methodically searches, screens, evaluates bias, and extracts data from relevant studies of the same type and combines them by meta-analysis. Classic WoE uses all relevant study types and inference by weight of evidence using Hill's (1965) or equivalent considerations.

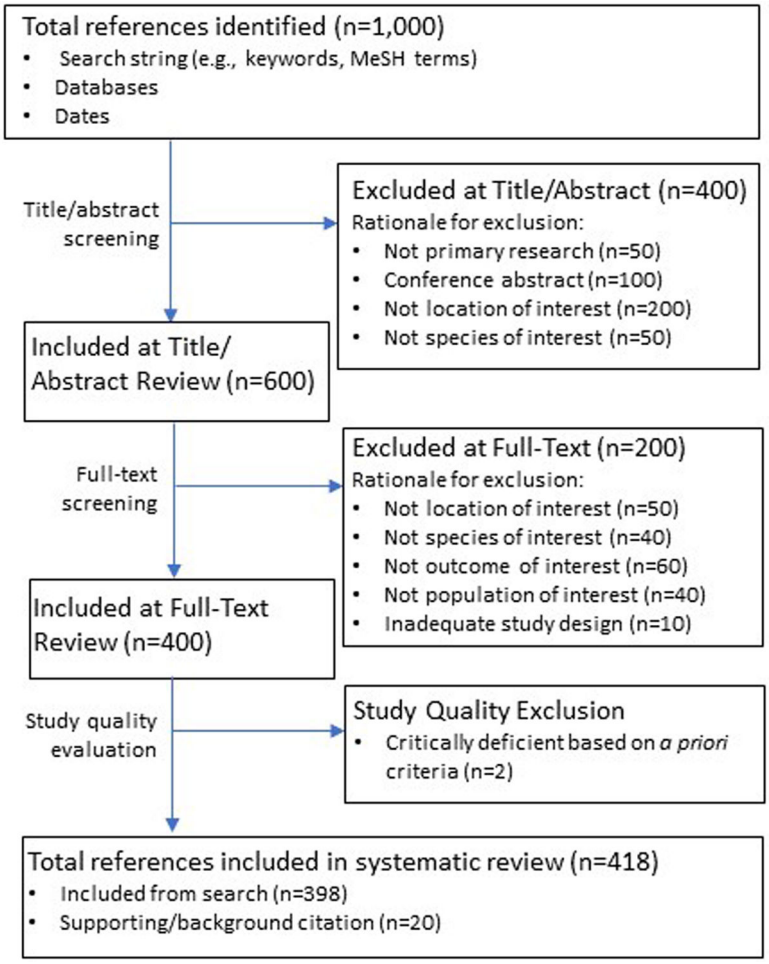

Figure 2.

A hypothetical example of the searching and screening of published studies in a systematic review of the literature.

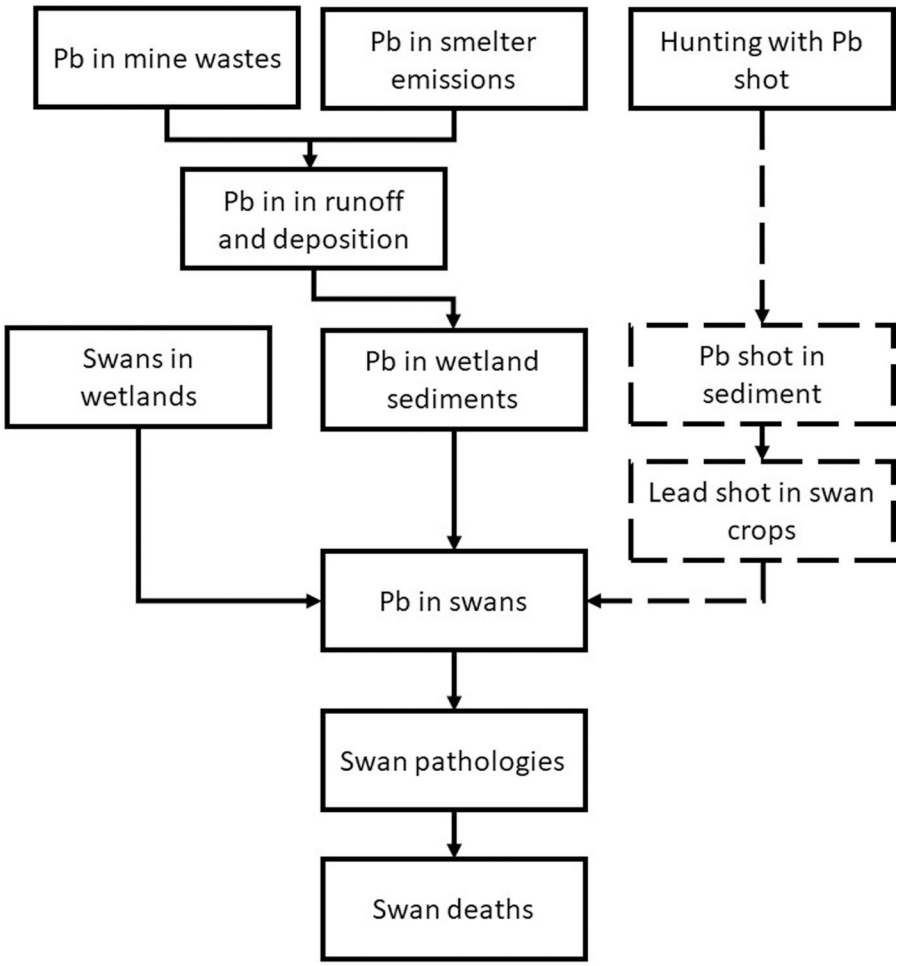

Figure 3.

A simple conceptual model of the induction of mass mortality of tundra swans by mine related Pb in the Coeur d'Alene River watershed, Idaho. The dashed boxes and arrows indicate a hypothesis that was not supported by evidence. As an alternative to organizing an evidence table in terms of causal characteristics for this causal hypothesis (Table 2), evidence could be organized by associating it with the boxes or links in the diagram, which correspond to system states and causal processes. Either approach interprets the body of evidence.

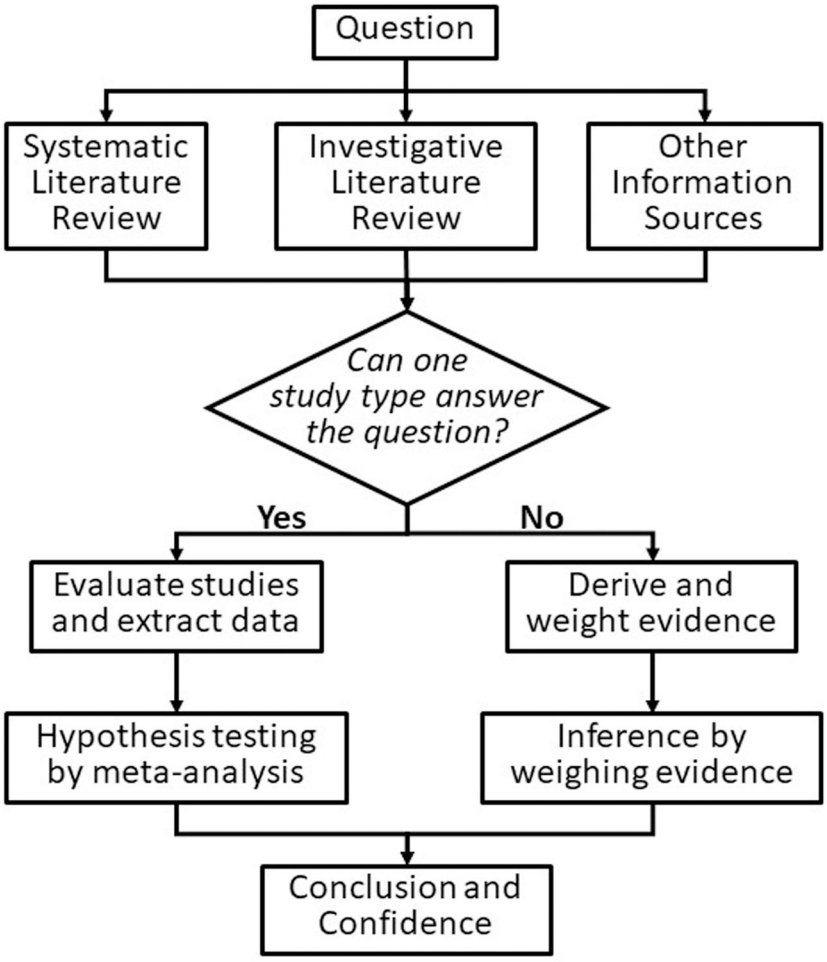

Figure 4.
Diagram of merged SR and WoE approaches to infer a quality from multiple studies or pieces of evidence.

Table 1.

Comparison of attributes of Classic Systematic Review, Classic Weight of Evidence, and an integrated approach.

Attribute	Classic SR	Classic WoE	Integrated SR & WoE
Emphasis	Transparent and unbiased assembly of information	Hypothesis best supported by available information	Scientific rigor while accommodating many situations
Generality of results	General applicability (e.g., the chemical is a carcinogen)	General or case-specific (e.g., the chemical caused this cancer cluster)	General or case specific
Institutions	Institutions specify methods and compile results (e.g., Cochrane)	None; users define methods	Some government agencies recommend for specific applications
Consistency	Consistent methods within fields	Diverse methods even within fields	Consistent framework with diverse options
Sources of information	Published experiments from literature	Literature, purposive studies and models, data bases, etc.	Any type of information
Types of evidence	One per assessment or, if more than one, assessed separately	Usually more than one	Usually more, because most questions cannot be answered with only one type
Implications of evidence	One type of study with one implication	Multiple types of evidence have different implications for hypotheses	Usually multiple types of inferences
Meta-analysis	Standard inferential method	Seldom used	Encouraged when appropriate
Causation	Not an issue because the experiments that answer the question are inherently causal	Causal inference from heterogeneous evidence that seldom experimentally answers the assessment question	Recommends distinct assessment to establish causation and then use the causal relationship to make predictions.
Role of rating	Used to express risk of bias or other qualities, but not for inference	Implied by the concept of weighting but seldom employed	Recommended for transparency of weighting evidence and drawing inferences
Role of expertise	Expertise needed but latitude minimized by detailed methods and statistical inference	Expert knowledge and judgment are essential and explicit	Expert knowledge and judgment are essential and explicit
Tools	Software tools for literature searching, screening search results, and extracting information	No known software tools for automating steps in environmental assessments	Software tools for literature search, screening and extraction.

Table 2.

Summary of evidence for mine-derived lead toxicity as the cause of mass mortality of tundra swans in the Coeur d'Alene River watershed. Based on evidence from (URS Greiner Inc and CH2M Hill 2011). The weight of evidence for the properties of each implication of the evidence is indicated by + for supporting, − for weakening, and 0 for uninformative information.

Causal Characteristic	**Evidence**	**Relevance/Reliability**[a]
Antecedents	• Spills of Pb mine tailings and atmospheric deposition from smelters account for the high sediment Pb levels.	+++/++
	• Hunting occurs elsewhere on flyway, not known locally	+/++
Time order	• No evidence—no pre-mining information on swan mortality	0/0
Co-occurrence	• Swan kills occurred in Pb-contaminated lakes and wetlands and not elsewhere in the region.	+++/++
Sufficiency	• Mortality occurred in laboratory tests of other avian species at Pb doses lower than those estimated for swans in the field.	++/++
	• Mortality occurred in laboratory tests of other avian species at Pb body burdens observed in dead or moribund swans in the field.	++/++
	• Severe sublethal effects occurred in swans fed a diet with 24% contaminated sediment.	++/+++
	• Consistent mortality was observed in the field at blood Pb levels >0.5 μg/g.	+++/++
Interaction	• Pb-contaminated sediments were found in swan guts and excreta.	+++/+++
	• Dead and moribund swans had high blood and liver Pb levels.	++/++
	• Lead shot was NOT found in Swan crops.	− − −
Mechanism	• Pb-contaminated Coeur d'Alene sediments fed to swans caused numerous adverse effects that caused emaciation and weakened the swans.	+++/+++
Specific alteration	• Swans in the field and the sediment feeding study had pathologies characteristic of Pb toxicity, particularly, enlarged gall bladders containing viscous dark green bile.	+++/+++
Collective Properties	**Body of Evidence**	**Properties: Reliability**
Number	• Most types of evidence have only one study.	0
Diversity	• There are several types of field and laboratory studies.	+++
Coherence	• All results support the causal hypothesis and are logically concordant.	+++
Absence of bias	• Critical evidence is from federal contractors or agencies who are unlikely to have conflicts of interest.	0
	• Data quality was rigorously assured and documented.	+++
Integrated WoE		
Finding	Exceptionally consistent, relevant and reliable body of evidence implicates mining derived Pb as the cause of swan kills in the Coeur d'Alene River watershed.	+++
	Pb toxicity from Pb shot refuted by lack of exposure.	− − −

[a]+++ convincingly supports, ++ strongly supports, + somewhat supports, 0 neutral or ambiguous