Science of the Total Environment 409 (2011) 1406–1417

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Why and how to combine evidence in environmental assessments: Weighing evidence and building cases

Glenn W. Suter II *, Susan M. Cormier

National Center for Environmental Assessment, U.S. Environmental Protection Agency, 26 W. Martin L. King Drive, Cincinnati, OH, 45268, USA

ARTICLE INFO

Article history:
Received 13 September 2010
Received in revised form 15 December 2010
Accepted 21 December 2010
Available online 31 January 2011

Keywords:
Weight of evidence
Risk assessment
Ecological risk assessment
Environmental epidemiology
Kit fox
Lines of evidence

ABSTRACT

All types of environmental decisions benefit from assessments that assemble and analyze diverse evidence. The diversity of that evidence creates complexities that can be managed using an explicit, well-designed process. We suggest two adaptations from the legal lexicon, weight of evidence and building a case. When weighing evidence, weights are assigned to each piece of evidence, and then the body of evidence is weighed in favor of each hypothesis by amassing the weights. Finally, the total weights of evidence for the alternative hypotheses are compared to determine which alternative has the preponderance of evidence in its favor. When building a case, pieces of evidence are organized to show relationships among multiple hypotheses or complex interactions among agents, events, or processes. We provide processes for weighing evidence and building a case and illustrate both approaches in a case study involving the decline of a kit fox population. The general approach presented here is flexible, transparent, and defensible. During its development, it has been applied to risk assessments for contaminated sites and to causal assessments in aquatic and terrestrial systems. It is intended to balance the need for rigor and discipline with the need for sufficient flexibility to accept all relevant evidence and generate creative solutions to difficult environmental problems.

Published by Elsevier B.V.

1. Introduction

Environmental assessors regularly find that multiple pieces of evidence are relevant to an inference. These may include multiple estimates of a parameter, multiple models of a relationship, or multiple types of evidence (e.g., a laboratory test, a field experiment, and an observational field study). Assessors may simply choose one piece of evidence and ignore the others, or they may consider all relevant evidence in a "weight-of-evidence process." Arguments for and against combining evidence are presented in Appendices A and B.

The phrase, weight of evidence (WoE), is used commonly but inconsistently and often vaguely. Dale et al. (2008) concluded that "An approach to interpreting lines of evidence and weight of evidence is critically needed for complex assessments, and it would be useful to develop case studies and/or standards of practice for interpreting lines of evidence." Similarly, Stahl (1998) complained that the U.S. Environmental Protection Agency's guidelines for ecological risk assessment lack guidance on weight-of-evidence approaches. Existing reviews (most notably, Weed (2005); Krimsky (2005), and Linkov et al. (2009)) have described existing practices but have not provided methodological guidance.

We believe that the lack of consensus concerning WoE is largely due to a lack of agreement about what it is and what it does. The term is used for approaches that range from genuine weighing of commensurable pieces of evidence to a process of interrelating heterogeneous evidence that we call "building a case." In this paper, we attempt to provide a useful theory and practice of combining evidence for environmental assessment. We do this by presenting the two metaphors (weighing and building), and then provide a simple general framework, alternative methods for implementing it, and an approach that we find to be generally useful. However, we do not intend to prescribe a particular methodology. We believe that it would be impractical for two reasons. First, the diversity of applications is too great. Second, most assessors develop their own methods that fit their preferences and those of their stakeholders and decision makers. Rather, our intent is to provide assistance to assessors as they decide how to combine evidence to solve problems.

We present alternative methods for combining evidence and recognize that there are potentially more methods. More than one method may be applied in a case to different evidence or in different stages of the assessment process. However, the most important advice is to use a formal method and to be explicit about what method you are using (Suter and Cormier, 2010).

1.1. Two legal metaphors

The concepts of weight of evidence (WoE) and building a case (BaC) can both be traced to jurisprudence. WoE is represented by the scales of justice that balance the weight of the evidence for guilt against that for innocence or for one party against another. Pieces of

* Corresponding author. Tel.: +1 513 569 7808; fax: +1 513 569 7475.
E-mail addresses: suter.glenn@epa.gov (G.W. Suter), cormier.susan@epa.gov (S.M. Cormier).

0048-9697/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.scitotenv.2010.12.029

evidence of different weights are placed in the appropriate pan of Justice's scales. The side that is lowest in the end prevails.

If one party bears the burden of proof, a standard weight (i.e., sufficient WoE) is placed in one pan. Evidence exceeding the sufficient weight must be loaded in the other pan to confirm that party's position. There may be multiple standard weights if there are multiple possible outcomes (e.g., human carcinogen, probable human carcinogen …not a carcinogen).

In adapting this judicial model to environmental assessment, we weight each piece of evidence, then we weigh the body of evidence in favor of each hypothesis by amassing the weights, and finally we compare the combined WoE for the alternative hypotheses to determine which has a preponderance of evidence in its favor. In environmental assessment, this weighing is typically done implicitly by professional judgment (Weed, 2005).

This judicial model fits cases that are genuinely analogous to a trial in that the assessment process must reach a dichotomous decision. Is the chemical a carcinogen, did this effluent cause the impairment, will the pesticide cause bird kills, etc.? However, many assessment questions involve magnitude, probability, or frequency, so the weighing of evidence must combine estimates as well as weights. What is the slope of the dose–response model, what is the probability that cadmium was the cause, how frequently will this pesticide cause bird kills, etc.? Hence, the weighing of environmental evidence must include processes for combining information as well as weights.

In sum, weighing evidence is a synthetic process that combines the information content of multiple weighted pieces of evidence. The information may be dichotomous (supports or not), quantitative values (e.g., an exposure or risk estimate), qualitative properties (e.g., large, medium or small), or a model. The weights that are applied to the information may express various properties that affect its credibility or importance and the weights themselves may be qualitative or quantitative. The combining of evidence may be a simple quantitative operation (e.g., weighted averages of concentration estimates) but more often involves difficult qualitative judgments.

The metaphor of building a case implies a very different process. Pieces of evidence are not simply equivalent discrete masses, but rather multiple parts of a structure or device. The constructed arguments may show relationships among multiple hypotheses. Fans of courtroom dramas are familiar with instances of this metaphor in which a case appears weak until a few critical pieces of evidence are provided that make the case fit together and explain how the crime occurred. Hence, under this metaphor, an assessor should be concerned about how the evidence might be logically combined rather than with which hypothesis has the weightiest body of evidence.

We believe that both metaphors are potentially useful. Some instances of combining evidence are simply a matter of weighing, some of building a case, and many require both.

2. Background to weighing evidence

2.1. Reasons for weighting and weighing evidence

Many assessment methods that are called weight of evidence combine evidence without explicitly assigning weights to pieces of evidence. This practice implies that all evidence is equally strong and of equal quality. That presumption is improbable. Even if all evidence was generated using high quality methods by people who never make mistakes, it is unlikely that all pieces and types of evidence provide equally strong or clear information. For example, observational and experimental data differ inherently. If you do not explicitly weight the evidence, you must either ignore those differences or consider them implicitly. Implicit weighting is not transparent to reviewers and stakeholders and may be subject to unconscious biases or incomplete logic.

2.2. Weighing evidence in different types of assessments

Although frameworks for performing "Weight of Evidence Assessments" have been proposed (Burton et al., 2002), we suggest that weighing evidence is not a particular type of assessment, but rather a method for planning, analyzing, or synthesizing information in various types of assessments. Specifically, WoE can be applied to each type of environmental assessment in the fully integrated framework (Fig. 1) (Cormier and Suter, 2008).

Condition assessments analyze monitoring data to determine whether environmental goals are being achieved that protect human health and ecosystems. WoE is applied when more than one measure of condition is available. For example, sport fishing records and data from electrofishing, seining, or snorkeling may be combined to determine whether a trout fishery is impaired (Wiseman et al., 2010).

Causal assessments use different pieces and types of evidence to determine whether an apparent association of cause and effect is actually causal (Suter et al., 2002) (http://www.epa.gov/caddis).

Weighing evidence has been a standard approach to causal assessment since the U.S. Surgeon General's Commission and A.B. Hill used it to demonstrate that smoking causes lung cancer.

Predictive assessments include risk and management assessments. Risk assessments estimate the nature, magnitude, and probability of effects for alternative policies or management actions. Management assessments may identify a preferred management action by weighing multiple types of evidence concerning benefits, costs, risks, public preferences, technical feasibility, and other considerations. The weighing is usually performed informally, but multi-criteria decision analysis, cost-benefit analysis, net benefit analysis, or other formal methods may be employed to weigh the evidence.

Outcome assessments determine whether a management action has succeeded. Although they seldom weigh evidence, outcome assessments could benefit from multiple categories of evidence, particularly in difficult or controversial cases. For example, to determine the outcome of a remedial action for contaminated sediment, one might apply all three components of the sediment quality triad (Chapman, 1990).

2.3. Steps in the assessment process involving weighing evidence

Environmental assessments of all types have three steps: planning, analysis, and synthesis (Fig. 2). The steps have different names in different types of assessments and different contexts. For example, for

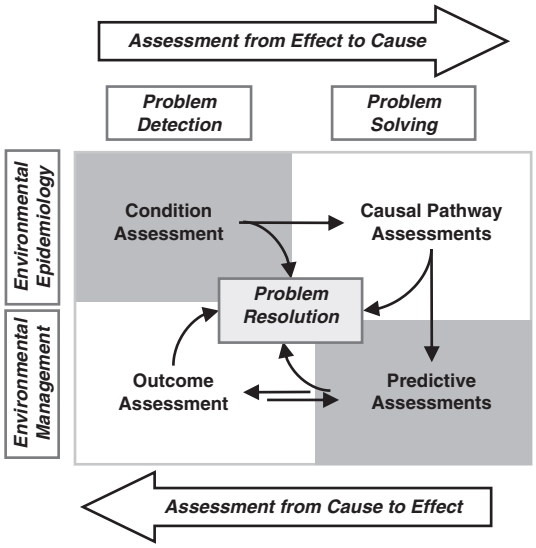

Fig. 1. The basic structure of an integrated framework for environmental assessment (Cormier and Suter, 2008).

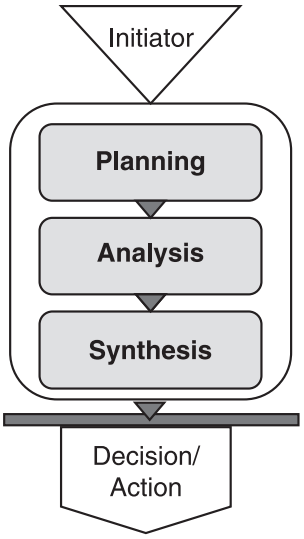

Fig. 2. A common process for performing environmental assessments (Cormier and Suter, 2008).

ecological risk assessments in the U.S. EPA (1998), the three steps are called problem formulation, analysis, and risk characterization.

In the planning step of risk assessments (e.g., problem formulation or hazard identification), evidence is weighed primarily to determine the mode of action of a chemical (i.e., its hazard). In particular, WoE is used to determine whether a chemical is a carcinogen and, if so, what type of carcinogen (U.S. EPA, 2005). However, evidence can be weighed for other planning purposes such as determining whether an exposure scenario is credible or determining how a mixture of chemicals should be assessed (Mumtaz and Furkin, 1992).

The analysis phase includes deriving exposure–response models and exposure levels or effects levels to implement the model. Hence, WoE may be used to select a model or a parameter value. The process of determining which mathematical function should be chosen as a model can be performed by determining the WoE provided by the data for the alternative functions (Good, 1961). This general approach to model selection may be based on relative likelihoods or relative information content and is now common in environmental sciences (Anderson, 2008; Good, 1983; Hilborn and Mangel, 1997). Second, multiple estimates of a parameter value may be weighted. The simplest approach is to derive a weighted mean of the alternative values, where the weights may be based on the quality of the methods used to derive each value. If multiple credible models or estimates are generated, they may all be carried forward into the synthesis stage.

The synthesis step combines the models and estimates from the analysis phase to estimate risks, identify causes, define conditions, etc. When multiple models or estimates are derived, particularly if they involve diverse methods, the evidence must be weighed. For example, in health risk assessments, risk information from epidemiological studies, clinical studies, or animal tests may be weighted and assessed as a body of evidence. In ecological risk assessments, biological surveys, ambient tests, and laboratory tests may be weighted and used in the same way. In causal assessments, various types of evidence are used with causal criteria to weigh the evidence for alternative causes. The weight of the body of evidence, based on the combined weights of individual pieces of evidence, may be used to express confidence or uncertainty in the results. However, the synthesis step often involves moving beyond weighing to logically building a body of evidence.

3. A basic framework for weighing evidence

The fundamental process for weighing evidence consists of three steps: assemble evidence, weight the evidence, and weigh the body of evidence (Fig. 3). This may be done once or iteratively. Iteration is needed if multiple distinct categories of evidence are available and if a category may be represented by multiple pieces of evidence. For example, if assessors have ten laboratory toxicity tests for a chemical and only one field observational study, they would not want to weigh all of those 11 studies in one body of evidence. Rather, the laboratory tests should be assembled, weighted, and combined into a single body of evidence from tests. That body of evidence would then be weighted and combined with the weighted observational study. The joint body of evidence (laboratory tests and field study) could then be weighted for comparison to the bodies of evidence for alternative chemicals causes (Fig. 4) or for comparison to some standard for adequate evidence. Hence, this weighing of evidence may be applied sequentially to individual pieces of evidence, to categories of evidence, and even to the entire body of evidence for a hypothesis.

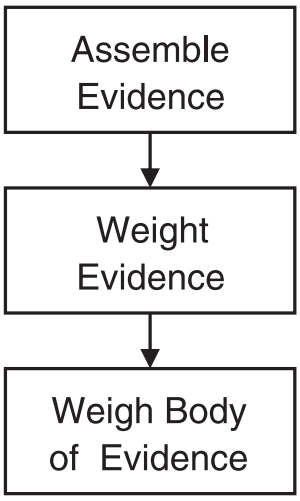

Fig. 3. A framework for the basic process of weighing evidence.

3.1. Assemble pieces of evidence

The WoE process begins by asking what evidence should be included in the set to be weighed. In some cases, such as the sediment quality triad, a standard set of evidence is generated. In others, a set of evidence may be generated to meet the needs of the particular assessment. However, in many assessments, evidence is obtained from the existing literature. Literature searches should be performed using prescribed criteria (e.g., only papers published in the last 20 years in peer-reviewed journals) and then inclusion criteria should be applied to the identified papers. For example, one might include only tests of freshwater fish performed by standard methods.

The assembly of evidence may include its organization into categories that play particular inferential roles and share common qualities in the weighting scheme, as in the top row of Fig. 4. For example, all laboratory toxicity tests provide clear evidence of causation but have questionable relevance to actual field exposures. Most methods for the categorization of evidence for causal assessments begin by referring to Hill's "criteria" (evidence related to consistency, temporality, analogy, etc.). The U.S. EPA's (2000) Stressor Identification guidance and the Causal Analysis Diagnosis Decision Information System (CADDIS—http://www.epa.gov/caddis/) attempted to be clearer and more consistent by defining types of evidence. Cormier et al. (2010) provided further clarification by distinguishing classification based on the source of the evidence from classification based on the characteristic of causation that it supports.

Methods for ecological risk assessment of contaminated sites tend to have similar typologies of evidence. An example is the Oak Ridge National Laboratory (ORNL) scheme for assessing existing contaminant

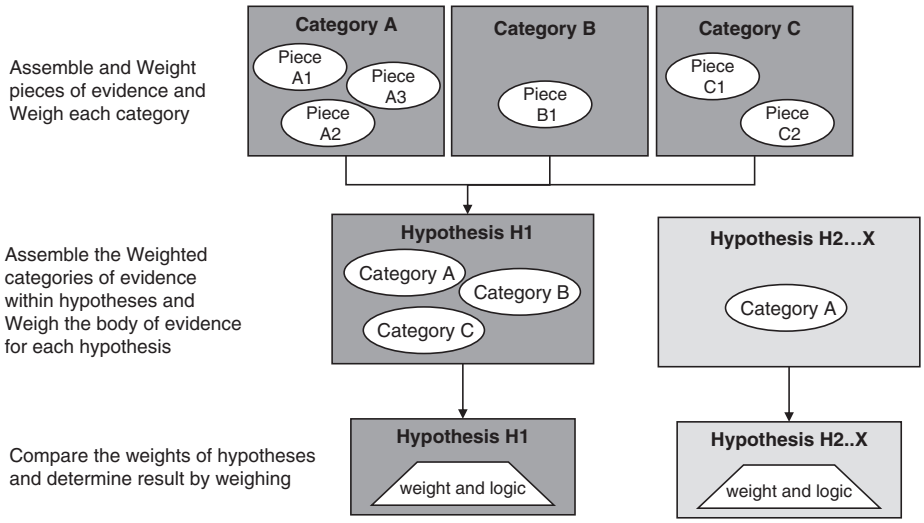

Fig. 4. A framework for tiered weighing of evidence in environmental assessments.

effects in Superfund Remedial Investigations which classifies evidence as 1) single chemical toxicity tests, 2) body burdens, 3) ambient media toxicity tests, 4) biomarkers and pathologies, and 5) biological surveys (Suter, 1996; Suter et al., 2000).

3.2. Weighting evidence

3.2.1. Weighting pieces of evidence

Weighting accounts for the manifest differences in the degree to which evidence should influence a particular conclusion. Weights may be based on the strength and quality of the evidence. However, for discrete hypotheses, the weighting process begins by determining the logical implication (the direction) of the evidence. That is, does it support or weaken an apparent impairment in a condition assessment, a hypothesized cause in a causal assessment, or a hypothesized effect in a risk assessment? Conventionally, this aspect of weight is represented by a + or − symbol.

The strength of a piece of evidence (represented by ovals in the top row of Fig. 4) is the degree to which it is consistent with the hypothesized relationship rather than natural variance. It may be expressed in various ways depending on the type of evidence. For empirical models or other statistical relationships it might be expressed by a correlation coefficient, likelihood, Akaike's (1973) information criterion, or other statistical measure of the degree to which the hypothesized association is supported by the data. For example, an exposure–response model with a high r^2 is stronger evidence than one with low r^2 and should be given more weight. For individual values such as the species richness of a biotic community or the mean body burden of a chemical in a population, the strength may be expressed as the magnitude relative to a reference or benchmark value. For example, a dissolved copper concentration that is five times the LC_{50} of the affected species is strong evidence for copper as a cause of toxicity. The strength of a piece of evidence may depend on whether it supports or weakens the hypothesis. An observed association between a potential cause and the effect of concern is weak supporting evidence, but a clear lack of association is strong evidence that there is no causal relationship. In sum, evidence is strong if it provides a clear distinction from background, reference, control, threshold, or random conditions.

The quality of a piece of evidence is difficult to define and potentially controversial, but it relates, in general, to the way in which the evidence was generated. Reviews with different criteria may rate the same study as high quality or low quality (Agency for Healthcare and Quality, 2002). Definitions of quality depend on the type of evidence and the use to which it is put. General elements of quality may include:

Study design, including control of exposure, replication, and randomization.

Relevance, including similarity of the agent, organisms, conditions, and measured responses to the system and endpoints of concern.

Reporting, including completeness of the description of methods and availability of data for reanalysis.

Performance, including signs of poor technique such as high control mortality, low control growth or fecundity, low-quality reference communities, or loss of treatments.

Statistical analysis, including appropriateness of the analysis to the assessment problem.

Potential for bias, including blinding of the investigators (metaphorically) and sources of funding.

Other elements of quality are specific to particular types of assessments such as risk assessments for contaminated sites (Menzie et al., 1996) or causal assessments of impaired communities (Cormier et al., 2010). It should also be noted that lists of weighting factors often do not discriminate among logical implication, strength, and quality of evidence. We distinguish them here to ensure that assessors consider all three sources of weight when weighting their evidence.

3.2.2. Weighting categories of evidence

Types of evidence or other evidence categories must be weighted before they are combined into a weighted body of evidence (i.e., to move from the top to the second row in Fig. 4). The weight of a category of evidence should be based on three considerations.

1. Different categories of evidence may have different inherent weights. In reviews of medical interventions, clinical trials are given greater weight than observational studies and anecdotal evidence is given the least weight (Pope et al., 2007). Similarly, biological surveys have been given more weight in some ecological assessments than toxicity tests or contaminant levels, even when the surveys have "major limitations" (McPherson et al., 2008).
2. The aggregate strength and quality of the pieces of evidence in the category also influence the weight. This component of the weight is an aggregate of the weights of all the pieces of evidence within the category, as discussed in the previous section.
3. The number of pieces of evidence within a category influences the weight.

Hence, an inherently strong category of evidence that has given conclusive results from a large number of high quality studies would have a heavy weight.

3.2.3. Methods for applying weights to evidence

When evidence is explicitly weighted, the process is usually qualitative and categorical. For example, the National Toxicology Program's categories for evidence concerning carcinogenicity and other modes of action in a study are 1) clear evidence, 2) some evidence, 3) equivocal evidence, 4) no evidence, and 5) inadequate study (Weinhold, 2009). The categories may be assigned directly or the evidence may be evaluated with respect to quality criteria or other considerations. The weights may be represented numerically, which permits arithmetically combining the weights (McDonald et al. 2007; Menzie et al., 1996). However, it must be recognized that the weights do not have natural units and cannot be naturally added or averaged. For example, if you assign a weight of 2 for strength of a piece of evidence and 4 for quality, those scores are numerical but not quantitative. If you sum them to obtain an overall weight of six for the evidence, it may impart a sense of rigor, but it is still a qualitative score. Therefore, qualitative weights are better expressed by symbolic systems such as the scale, +++, ++, +, 0, −, −, − − −, commonly used in causal assessments (Fox, 1991; Susser, 1986; Suter et al., 2007), or graphic symbols such as ●, ◘, ○, (see Consumer Reports magazine or McDonald et al., 2007). For communicating results, color coding may be effective (e.g., an evidence table with different shades of red and green for different weights of negative and positive evidence).

Qualitative weights are usually generated by an assessor's professional judgment. However, formal elicitation methods may be used to gain credibility by using outside experts or to gain acceptance by using stakeholders or decision makers. Although weights are usually generated ad hoc, standard weights may be developed (Dagnino et al., 2008; Menzie et al., 1996; U.S. EPA, 2000).

Criteria may be used to guide qualitative weighting. In the simplest case, criteria provide a check list (yes, the evidence has this quality; no, it does not have that one). This approach ensures that all criteria are considered and provides a basis for choosing a weight.

The quantitative weighting of evidence is a matter for statistics. In the simplest case, the weight could be a probability such as the probability that a biotic community with x species is unimpaired, given the distribution of species numbers in designated unimpaired sites. The quantitative definition of WoE is the log of the Bayes factor (Good, 1983). Given that definition, the weighing of evidence for two hypotheses (i.e., models of the same data) is simply the WoE given by a data vector x for hypothesis A (e.g., impaired) relative to hypothesis B (e.g., unimpaired). If the prior distributions for the hypotheses are equal, the WoE is equal to the log of the likelihood ratio. If log base 2 is used, the units of WoE are bits. This suggests that WoE is related to information content, and in fact, if the information loss to parameter estimation is considered, WoE is equivalent to Akaike's Information Criteria for the hypotheses and data vector. As a method for estimating the WoE for impairment, this approach is problematical because it requires estimating a model for impairment even though there are many ways to be impaired (Smith et al. 2002). However, one might identify types of impairments (e.g., low dissolved oxygen impairment, temperature impairment, flashy flow impairment, etc.) and determine the WoE for one cause relative to another given data from a site. The approach is more naturally applied to weighing evidence for one hypothesis and its associated model versus another (i.e., model selection—Anderson, 2008). For example, Garnett and Brook (2007) identified a best model that encompassed 92% of the Akaike WoE concerning causes of the extinction of Austrian birds. However, it must be recognized that all of these statistical approaches weigh only the evidence contained in the variance of a data vector. Other qualities of evidence such as relevance or use of a standard method must be separately weighted.

3.3. Weighing the body of evidence

3.3.1. Definitions

The body of evidence for a hypothesis includes all of the weighted categories of evidence. Its weight is the overall strength and quality of the body of evidence with respect to its logical implication (represented by the boxes in the second row of Fig. 4). That is, the weighted body of evidence expresses how well the evidence supports or weakens the hypothesis.

3.3.2. Methods for weighing a body of evidence

The most difficult question with respect to weighing evidence is, how should a body of heterogeneous evidence (weighted or not) be combined? This section lists methods that have been used to combine observational studies such as field surveys, experimental studies such as toxicity tests, and general knowledge such as mechanistic information. The order of appearance of the methods is for the convenience of presenting concepts. Advantages and disadvantages of the methods are presented in Table 1.

Expert judgment is the most flexible and common method for combining evidence. If the assessor has sufficient experience, including feedback concerning the results of his prior judgments, the process might truly be called an expert judgment. If not, the judgments of outside experts may be elicited. Computer based decision support systems provide immediate results to the assessors of the effect of applying different weights or of including certain evidence (Marcomini, et al. 2009). This helps to identify the evidence and weights that do and do not tend to affect the assessment's findings.

Criteria-guided judgment**,** is a step up in rigor in the application of the assessor's judgment, guided by a set of criteria or issues to consider. The most commonly applied set of considerations is Hill's

Table 1
Strengths and weaknesses of methods for weighing evidence.

Method	Greatest strengths	Greatest weaknesses
Expert judgment	Quick, flexible	Not transparent, requires faith in expert
Criteria guided judgment	Flexible, transparent	Requires some expert judgment, takes time
Check list	Quick, transparent	Inflexible, dichotomous
Independent applicability	Protective, quick, transparent	Conservative, inflexible
Numerical indices	Consistent	Pseudo-quantitative, inflexible
Logic tables	Transparent, consistent	Inflexible (requires standard set of high quality data)
Sequential logic	Transparent, consistent, efficient	Inflexible, limited to simple alternatives
Case-specific logic	Flexible	Logic may not be accepted
Legal weighing of evidence	Flexible, decisive	Influenced by quality of each sides presentation
Combined statistical weights	Quantitative, transparent	Limited to quantitative evidence
Statistical weighing	Quantitative, transparent	Limited to quantitative evidence
Hypothetico-deductive method	Convincing, transparent	Seldom applicable

"criteria" for judging a proposed cause of observed effects. In particular, they are used as issues to consider when determining whether a chemical is a carcinogen (U.S. EPA, 2005). If the evidence is weighted, the judgment is guided by the weights as well as how many criteria or types of evidence support a hypothesis. Hence, the weight of the body of evidence is the aggregate weight of the relevant pieces and categories of evidence. However, characteristics of the body of evidence as a whole also influence its quality. A body of evidence has greater weight if it has the following qualities (Cormier et al., 2010).

Credibility — the body or evidence is based on relevant and high quality information weighted by relevance, and quality of study design and execution.

Coherence — the body of evidence is internally consistent, consistent with scientific knowledge and theory, and together logically explains the facts in the case as judged by internal consistency and consistency with theory.

Strength — the body of evidence includes pieces of evidence that are logically compelling (e.g., the effect occurred before the cause) or that present quantitatively strong relationships (e.g., high correlation coefficients or relative likelihoods) (see section above). Strength is weighted based on logical implication and by the independence, quantitative strength, and specificity of the evidence.

Diversity — many sources of evidence and characteristics of causation are represented in the body of evidence including a variety of types of evidence and evidence from different datasets.

Check lists provide a simple and rigorous but inflexible method. That is, a list of dichotomous properties must be checked off before a hypothesis is accepted. The best known example is Koch's postulates, three or four pieces of evidence (depending on the version) that must be provided for a pathogen to be proven to cause a disease. Koch's postulates have also been adapted to environmental contamination (Sec. 4.3.3.3 in Suter et al., 2007).

Independent applicability is the simplest formal method; any sound evidence is sufficient to demonstrate impairment. This practice is used in enforcement of the U.S. Clean Water Act (U.S. EPA, 1991) and other situations in which protection is the paramount goal. If measurements of any chemical exceed water quality criteria, the system is impaired; if the water is toxic, it is impaired; and if the biotic community fails to achieve biological criteria, the system is impaired. This policy recognizes that we do not have all three types of evidence in every case, and, even when we do, each type of evidence has weaknesses that may cause an impairment to be missed (Yoder and Rankin, 1998). For example, most biological criteria are based on invertebrates, but they are less sensitive than fish to some contaminants such as selenium. Similarly, standard toxicity tests of ambient water are short-term sub-chronics that cannot detect the bioaccumulation or reproductive toxicity of selenium in fish. Of the routine methods, only chemical criteria will detect impairment caused by selenium. Hence, although independent applicability has been criticized as overly protective (Lee and Jones-Lee, 1995), it is in keeping with legal mandates to protect biological integrity.

Numerical indices based on the ratio to reference are used in some biological condition indices (Hawkins et al. 2000, Wright et al. 2000) and were the original method for combining Chapman's triad of evidence (Chapman, 1990; Long and Chapman, 1985). For example, data sets may be normalized to a 1–100 scale, averaged within types and then averaged across sites to create an ecotoxicological index (U.S. EPA, 1994). Chapman (2003) and others have come to realize that reducing the triad or other bodies of evidence to index numbers was not useful. In particular, it destroys the distinct information provided by each type of evidence. For example, it might give the same score to a case in which all lines of evidence were weakly positive as to a case in which none of the measured chemicals were elevated but the sediment was highly toxic and the biotic community was highly degraded.

Table 2
Inference based on the sediment quality triad (modified from Chapman 1990).

Situation	Chemicals present	Toxicity	Community alteration	Possible conclusions
1	+	+	+	Strong evidence for pollution-induced degradation
2	−	−	−	Strong evidence that there is no pollution-induced degradation
3	+	−	−	Contaminants are not bioavailable, or are present at non-toxic levels
4	−	+	−	Unmeasured chemicals or conditions exist with the potential to cause degradation
5	−	−	+	Alteration is not due to toxic chemicals
6	+	+	−	Toxic chemicals are stressing the system but are not sufficient to significantly modify the community
7	−	+	+	Unmeasured toxic chemicals are causing degradation
8	+	−	+	Chemicals are not bioavailable or alteration is not due to toxic chemicals

Logic tables combine types of evidence providing a standard conclusion for each possible outcome of a set of standard types of evidence (Table 2). Most notable among these is the logic triad developed for contaminated sites by Chapman et al. (2002). The standard body of evidence is:

1. Chemical analyses of the contaminated media which are compared to either reference concentrations or to single chemical toxicity data.
2. Toxicity tests of the contaminated media, which are compared to tests of reference or standard media.
3. Biological surveys of the biotic community inhabiting the contaminated media, which are compared to reference communities.

The logic table is applied to the three types of evidence to determine whether chemical contamination is responsible for biological impairment. That is, it is an ecoepidemiological assessment, combining condition and causal assessments. For example, if the concentrations of chemicals are not elevated, but the sediment is toxic and the biological metrics are below reference levels, the conclusion is that the impairment is due to unmeasured chemicals. This logic is impeccable if the tests and measurements are complete and the quality of the data and its analyses is high. In particular, the logic requires that the right tests and biological metrics be employed. For example, it would fail to identify the impairments due to selenium as discussed in the previous paragraph. Because the score for Se would be +, −, −, the logic table would conclude that the elevated Se is "not bioavailable or is present at nontoxic levels." The table allows the possibility of unmeasured chemicals, but not unmeasured modes of toxicity or unmeasured biological effects. Hence, the standard logic is helpful in most cases but will trip up assessors in some cases. Chapman has recognized these limitations of the standard triad logic and has encouraged flexibility in its application.

Sequential logic is useful if the pieces or categories of evidence can be reduced to a two-part logic (yes, no) or three-part logic (yes, no, uncertain) and may be depicted by a logic diagram (Fig. 5). Examples include the SALE system (Hull and Swanson, 2006), a system for assessing selenium risks (McDonald and Chapman 2007), and a causal assessment in West Virginia (Gerritsen et al. 2010). Another method begins with a simple sequential logic but ends with a guided judgment (Burkhardt-Holm and Scheurer, 2007; Forbes and Calow, 2002). The chief advantage of sequential logic is that data may be

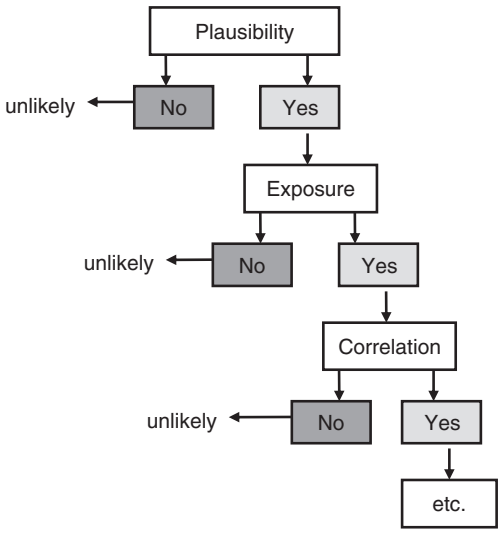

Fig. 5. An example of weighing evidence by applying a system of sequential logic; the first three steps of the system of Burkhardt-Holm and Scheurer (2007).

generated sequentially, and the process may reach a conclusion without generating or analyzing all types of evidence.

Case-specific logic involves devising a logic for interpreting the multiple types of evidence to fit the case at hand. A standard logic such as depicted in Table 2 may not be used because it does not suit the needs of assessment or because of problems with the standard data (McPherson et al., 2008), unequal quality or strength of the types of evidence, or the need to include additional types of evidence (Suter et al., 1999, 2000).

Legal weighing of evidence is used in courts of law where evidence is weighed by a neutral party. If science courts were established, this approach could be used for high profile or contentious issues without resorting to the actual legal system.

Vote counting is simply counting the pieces of evidence for each alternative and choosing the winner. It can be used if a few clearly defined alternatives must be weighed and the pieces of evidence are similar. For example, one might count the number of tests that find an increase in tumors and the ones that do not.

Statistical weighing, also termed meta analysis, includes various statistical methods for combining equivalent quantitative results of multiple similar studies (Gates, 2002; Weed, 2000). These include weighted means of estimates, model averaging, and others. For example, Cormier et al. (2008) averaged the results of multiple models used to estimate a threshold for impairment of stream communities from deposited sediment.

Combined statistical weights of evidence is possible because logs of ratios are additive, the WoE for different types of evidence, as defined by Good (see Section 2.3), may be added to obtain an overall WoE (Smith et al., 2002).

Hypothetico-deductive method treats each body of evidence as a hypothesis, deductively generates predictions and chooses the one that performs the best by comparison to reality. A familiar example is the prediction of the path of hurricanes. Alternative models generate predictions, the model that best predicts the storm's behavior by a statistical or criterion is used to make official predictions for the next time interval, and the process is repeated. More complex cases involving multiple pieces or categories of evidence per hypothesis can be devised as long as they provide deductions of clearly different but verifiable phenomena.

3.4. Comparison of hypotheses and expression of results

Although assessments often address a single hypothesis, the most reliable approach to assessment is the comparison of alternative hypotheses (the bottom row in Fig. 4). The results of weighing multiple pieces of evidence with respect to multiple hypotheses may take at least two distinct forms.

3.4.1. A numerical result

When the weighing process addresses alternative data sets or models that estimate a quantitative variable, the result is an estimate based on all of the evidence. It may be a weighted mean numerical value, the range of alternative estimates, the best (weightiest) estimate with an expression of variance, or some other combined result.

3.4.2. The hypothesis best supported by the evidence

This approach simply presents the weightiest of the alternative hypotheses. It may be a most likely cause, a category of carcinogenicity, a preferred remedial action, or some other discrete assessment outcome. The result may include a score indicating the WoE for the hypothesis. The other alternatives are sometimes ranked or categorized in some way, such as lower risk or no risk.

Sometimes, weighing the evidence is not sufficient. Often, the evidence is insufficient and additional studies must be performed and the assessment must be repeated. Other cases are complex or ambiguous and may require a more complex inferential process than weighing the evidence. We refer to this option as building a case.

4. Building a case

If comparison of weighted bodies of evidence does not lead to acceptance of a hypothesis, then one can attempt to build a case by reexamining hypotheses and reconsidering the evidence. The process of building a logical case can take two forms.

1. For single hypotheses, if weighing the evidence for and against a hypothesis led to an incomplete or inconclusive result, the addition of new assumptions or previously unconsidered prior knowledge or reinterpretation of the evidence can build a consistent case for the original hypothesis or a modified hypothesis.
2. For multiple hypotheses, if weighing the evidence for and against each alternative hypothesis did not reveal a best hypothesis, the *a priori* hypotheses may be integrated to create new and better hypotheses (i.e., hypotheses that are more consistent with the evidence).

Clearly, building a case provides important opportunities to provide a useful result when simply weighing evidence fails. However, it is also clear that it provides opportunities for the creation of "just so stories." *Post hoc* hypotheses are discouraged in scientific inference because fitting the hypothesis to the evidence can lead to fudging or self deception. Also, the theoretical or practical knowledge required for building a case is often unavailable to the assessor. The best protection against inadequate insight and knowledge are not unlike those against poor analysis: clearly articulated arguments, caution with respect to over interpretation, and, involvement of multiple assessors with a different perspectives. And as emphasized in this paper, a clear and consistent method should be used to build the case.

When building a case for an individual hypothesis, the most generally useful approach is to 1) list the inconsistencies, 2) ask for each what could explain it, and then 3) find and evaluate evidence relevant to the explanations. A common example in causal assessments is positive evidence from laboratory tests but negative evidence from exposure–response in the field. For example, copper concentrations in a stream may exceed levels that cause effects in the laboratory, but that evidence is contradicted by the lack of a relationship between copper concentrations and effects in the field. This may be explained by differences in bioavailability of copper between the laboratory and the field. The inconsistent evidence may also be explained by poor field data. These plausible explanations are potentially sufficient to discount or retain the hypothesis that copper caused the impairment. However,

without independent evidence, neither explanation can be supported. For the first explanation, such evidence might include a finding that dissolved organic carbon concentrations are high in the stream, which would reduce bioavailability. The second explanation might be supported by the discovery that different sites were sampled by different crews, in different seasons, using different protocols. If the number of possible explanations and the amount of evidence are substantive, a formal weighing of the evidence concerning the revised hypotheses is appropriate.

When building a case from multiple hypotheses, the most generally useful approach is to create a synthetic conceptual model that combines the credible relationships in the individual *a priori* conceptual models into a new causal structure. This is a process of pulling in the strong relationships (i.e., box–arrow–box combinations from the conceptual models), deleting those that had been disproved, and looking for ways that the remaining relationships might interact to create a new conceptual model that is more consistent with the evidence. The links come from prior scientific knowledge and should be based on established mechanisms. They can be created from:

1. Combined causes. Co-occurring causal agents may have additive or interactive combined effects.
2. Proximate causes converted into intermediate causes. For example, a chemical might not cause the loss of a population through direct toxic effects, but it may reduce food or habitat.
3. Hypothetical linking processes. For example, an unknown immunotoxic effect of a chemical may enhance the effect of a pathogen that is not known to be lethal and would link a chemical to a pathogen through a new component — susceptible organisms.
4. Hypothetical agents or events. For example, an unknown septic field might be the source of organic matter that lowered dissolved oxygen.
5. Extreme events that were thought to be too improbable to include initially.

As in building a case for a single hypothesis, these synthetic hypotheses should be supported by evidence, preferably independent of the evidence used to generate them.

Although conceptual models are the best general tool for building a case, it is desirable to convert them into prose. If the logic of the case is sound, it can be converted into a narrative that will be compelling to decision makers and stakeholders who are put off by "spaghetti diagrams." In addition, a narrative can clarify the mechanisms that cause the links in the model. Finally, translating between a graphic and narrative form can act as a quality check for the assessors' logic.

5. Our approach to weighing evidence

We have developed a generic approach to inference by WoE that is based on experience in both risk assessment and causal assessment (Fig. 6). Note that this represents the synthesis stage of an assessment (e.g., the risk characterization of a risk assessment) (Fig. 2). It must be preceded by an analysis stage in which data are converted into evidence concerning the assessment problem defined in the planning stage. Although we do not argue that it is the best approach in all instances (see the alternative approaches discussed above), we do believe that it is generally useful and that its adoption would improve most assessments.

- Assemble all relevant pieces of evidence for each hypothesis and categorize them into consistent sets.
- Weigh evidence.
  Weight each piece of evidence. If evidence will be quantitatively combined, apply quantitative weights. Otherwise, apply qualitative weights based on logical implication, strength, and quality.
  Weigh the evidence (i.e., combine weighted pieces or categories of evidence) quantitatively if appropriate, otherwise use criteria guided judgment to combine. Weigh based on credibility, coherence, strength, and diversity of the body of evidence.
  Compare alternative hypotheses based on their weighted bodies of evidence. (If only one hypothesis is assessed, compare its weighted body of evidence to a standard of acceptability.)
- Build a case for a new hypothesis if the evidence is weak or inconsistent for all prior hypotheses or if it provides a more convincing result.
- Iterate the assessment based on new evidence if results are ambiguous or if a new hypothesis requires support.

6. Demonstration of our approach

This general approach to weighing evidence can be illustrated by a causal assessment for a decline in the abundance of a San Joaquin kit fox population on the Elk Hills Naval Petroleum Reserve (NPR-1) in California (U.S. EPA, 2009). The impairment is a decline in kit fox abundance from at least 160 individuals to 40. It co-occurred in space and time with increased oil drilling and production. The decline was greatest on the developed portions of the field, so comparisons could be made between developed and undeveloped portions of NPR-1, between NPR-1 and another oil field (NPR-2), and between the oil fields and other habitats. The potential causes were prey availability, habitat quality, accidents, disease, toxicity, and predation. Two of the causes, prey abundance and habitat quality, were further divided by source: due to climate or due to physical disturbance. The assessment was performed using CADDIS (http://www.epa.gov/caddis/). Because some types of evidence are addressed by more than one piece of evidence, the tiered approach is used (Fig. 4). This section will not present the full causal assessment, but rather will present portions of the assessment that illustrate components of the approach.

6.1. Assemble pieces of evidence

The decline in kit fox abundance was revealed by monitoring performed to meet the U.S. Department of Energy's obligations to manage an endangered species. Subsequently, additional studies were preformed to address some of the possible causes. Finally, relevant information from the literature was used to support the generality of potentially causal relationships on the site. Each piece of evidence for each candidate cause was associated with one of the 17 types of evidence in CADDIS.

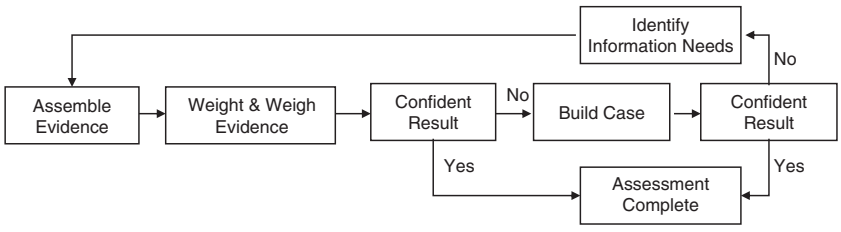

Fig. 6. Diagram of a general approach for weighing evidence in the synthesis stage of an environmental assessment.

6.2. Weight pieces of evidence

The pieces of evidence were qualitatively weighted using the CADDIS scoring system. The numbers of + or − signs are assigned to a piece of evidence as follows.

+++ or − − −	Convincingly supports or weakens
++ or −−	Strongly supports or weakens
+ or −	Somewhat supports or weakens
0	No effect

The CADDIS website provides guidance on weighting. For example, co-occurrence of a candidate cause and an effect only somewhat supports (+), but lack of co-occurrence convincingly weakens (− − −) the case for causation. This guidance is based on the fact that co-occurrences may be coincidental, but lack of co-occurrence precludes causation.

In some cases multiple pieces of evidence for a candidate cause were related to a type of evidence. For example, evidence of a causal pathway (from source to exposure) was available from soil analyses, a waste survey, water analyses, and records of petroleum spills and sumps (Table 3).

Table 3
Pieces of evidence and their weights for one type of evidence, causal pathway, for toxic chemicals as a candidate cause.

Soil — routes of exposure to soil exist, but contaminant concentrations were not elevated in random soil samples from developed NPR-1.	−
Wastes — spills and deposits of wastes were present and available for direct or indirect exposure.	+
Water — waste waters were highly contaminated, but there was no evidence of drinking by kit foxes.	0
Petroleum — spills and sumps were available to foxes and at least one died in a spill.	+

6.3. Combine pieces of evidence within types

Most types of evidence have only one piece of evidence, if any. For others, the weights are combined logically into an overall score for the type. For example, although the evidence of a causal pathway for toxic chemicals was mixed (Table 3), the combined score for that type of evidence is + (Table 4). That score is appropriate because only one pathway is needed for exposure to occur. However, it is only somewhat supportive (i.e., only one +), because the existence of a pathway is not strong evidence of actual exposure, much less effects.

Pieces of evidence were combined quantitatively for one type of evidence, mechanistic sufficiency. All of the pieces of demographic evidence including survivorship, fecundity, and emigration data from various studies (radio collared foxes, den surveys, road kills, etc.) were combined using a projection matrix population model. That model determined that the decline was due to high mortality in the first year of life, and predation by coyotes accounted for 80% of the mortality. Hence, predation was given a score of +++ for mechanistic sufficiency (Table 4).

6.4. Weigh the body of evidence for each candidate cause

In CADDIS, the body of evidence for each candidate cause is weighed in terms of two integrative types of evidence: consistency and reasonable explanation of the evidence (Table 4). For example, predation is consistently supported by the evidence and one type of evidence, mechanistic sufficiency, was convincing. Therefore, it received +++ for consistency and required no explanation. At the other extreme, the evidence for disease was negative, except that diseases have caused declines in other species of foxes at other locations, and no explanation could account for the negative evidence or build a case for disease as a plausible cause.

Table 4
Comparison of the strength of evidence for the candidate causes. Types of evidence with no evidence for any candidate cause were excluded. Codes are identified in a footnote[a].

Types of evidence	Prey[c]		Habitat		Predation	Toxics	Accidents	Disease
	Disturbance	Climate	Disturbance	Climate				
Evidence that uses data from the case								
Spatial/temporal co-occurrence	++[b]		+	−	+	+	+	−
	+	−	+	−	+	+	+	−
Temporal sequence	o[b]		o[b]		o	o	ne	ne
Evidence of exposure or biological mechanism (pathway independent)	++[b]		ne	ne	+ +	+ +	+ +	− −
Evidence of exposure or biological mechanism (by pathway)	−	+	ne	ne	+ +	+ +	+ +	- -
Causal pathway[b]	+	−	+ +	−	+	+	+	o
Stressor–response relationships from the field (pathway independent)	+ + +[b]		− −	o	ne	ne	ne	ne
Stressor–response relationships from the field (by pathway)	−	+	− −	o	ne	ne	ne	ne
Manipulation of exposure	+[b]		ne	ne	+	ne	ne	ne
Symptoms, starvation	−[b]		ne	ne	ne	ne	ne	ne
Symptoms, reproductive (pathway independent)	+[b]		ne	ne	ne	ne	ne	ne
Symptoms, reproductive (by pathway)	+	−	ne	ne	ne	ne	ne	ne
Mechanistic sufficiency	−[b]		o[b]		+++	−	+	− −
Evidence that uses data from elsewhere								
Mechanistically plausible cause	+	+	+	+	+	+	+	+
Stressor–response relationships from other field studies	+[b]		ne	ne	o	o	+	+
Stressor–response relationships from laboratory studies	ne	ne	ne	ne	ne	−	ne	ne
Evaluating multiple lines of evidence								
Consistency of evidence	−	−	−	− −	+++	−	++	-
Explanation of the evidence	C[d]	C[d]	C[d]	C[d]	na	−	na	-

[a] + supports the candidate cause, − weakens the cause, o neither weakens nor supports, ne = no evidence, na = not applicable, C = a contributing factor in the final case.
[b] A single score for Prey indicates that the same data were used to evaluate both pathways.
[c] An additional causal pathway for prey abundance, competition for prey by coyotes, was ambiguous.
[d] The candidate cause is a contributor to the final case.

6.5. Compare across candidate causes

Comparison of evidence across candidate causes is relatively simple in this case (Table 4). Predation is a consistent and convincing cause and appears to have been sufficient alone. Accidents are also consistent but less strongly supported. The body of evidence for each of the other potential causes is predominately negative or has significant inconsistencies.

6.6. Building a case by integrating hypotheses

The occurrence of two apparent causes and positive evidence for other potential but inconsistent causes suggested that building a case might be informative.

1. Predation by coyotes was clearly the dominant cause, but accidents, which accounted for 15% of mortality, were a contributing cause with the same mode of action, elevated mortality. Hence those two causes were joined as joint proximate causes.
2. Disease, climate, and toxicity were not contributors and were discarded.
3. The remaining two causes, reduced prey and reduced habitat quality, are not causes in themselves, but they could be part of the causal network to the proximate causes by increasing exposure to predators and vehicles. That is, reduced habitat quality resulted in reduced prey which increased time spent hunting which, in turn, increased exposure to predation and accidents. That pathway was added as consistent with the evidence and general knowledge.
4. Disturbance (oil field development) may have improved coyote habitat (food subsidies and protection from shooting and trapping) but when a predator control program was imposed, it reduced their abundance. Those pathways were added as consistent with the evidence and general knowledge.

The case built from the integrated body of evidence for all potential causes is presented as a conceptual model in Fig. 7. It accounts for the evidence better than any of the individual potential causes, and provides a more complete account of the system that could be used to design management interventions or to generate new evidence.

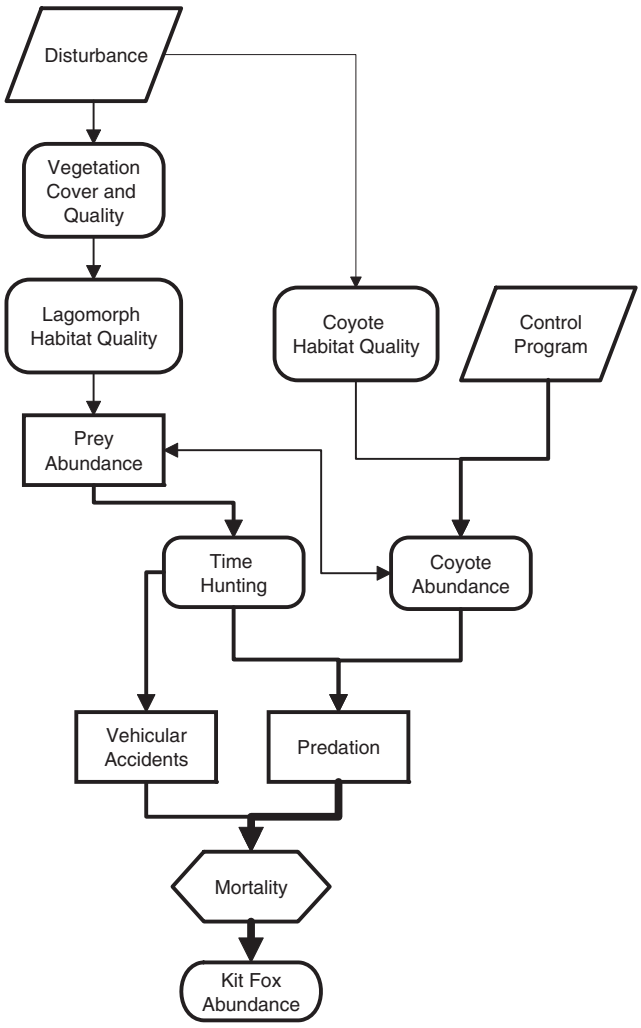

Fig. 7. The final conceptual model for the cause of the kit fox decline (U.S. EPA, 2009). The thickness of the arrow lines indicates the relative weights among the causal connection.

7. Conclusions

Ecological assessors should emulate the medical assessors who have standard methods and guidance for weighing evidence from clinical trials [e.g., the Jadad scale (Jadad et al., 1996) and CONSORT Statement (Moher et al., 2001)], from reviews of clinical trials [e.g., the Cochrane handbook (Higgins and Green, 2008)], and from bodies of evidence including clinical trials, epidemiology, and diagnostic tests [e.g., GRADE (GRADE Working Group, 2004) and SORT (Ebell et al., 2004)]. Systems for weighing medical evidence have been rated and features that lead to different conclusions have been identified (Lohr, 2004). In contrast, the weighing of evidence in environmental assessments is almost always performed informally, ad hoc and with little regard for how the method for weighing evidence influences the outcome. The exceptions such as the sediment quality triad tend to be designed for a particular set of pieces of evidence and a particular assessment problem.

The general approach presented here is flexible, transparent, and defensible. During its development, it has been applied to risk assessments for contaminated sites (Jones et al., 1999; Suter et al., 1999) and to causal assessments in aquatic and terrestrial systems (Bellucci et al., 2010; Gerritsen et al., 2010; Hicks et al., 2010; Haake et al., 2010; U.S. EPA, 2009; U.S. EPA, 2010; Wiseman et al., 2010). It is intended to balance the need for rigor and discipline with the need for sufficient flexibility to accept all relevant evidence and generate creative solutions to difficult problems. When weighing the evidence, flexibility comes from a qualitative analysis while discipline comes from using a pre-defined scoring system and criteria. When building a case, flexibility comes from the creative use of prior scientific knowledge and discipline comes from requiring a mechanistic basis for the novel aspects of the case and from requiring independent supporting evidence. In any case, the inferential procedure should be defined as clearly as a statistical test or mathematical model.

Acknowledgements

We thank Michael McManus and Amy Meyers and three anonymous reviewers for their helpful comments. The paper is based on work supported by the U.S. EPA (both authors) and the U.S. Department of Energy (Glenn Suter). Tom O'Farrell and his staff provided data and advice for the kit fox case and Larry Barnthouse provided the population modeling. The manuscript has been reviewed and cleared by the U.S. EPA, but does not necessarily reflect Agency policies.

Appendix A. Why combine multiple pieces of evidence?

At least a dozen arguments may be presented for making inferences by combining multiple pieces of evidence.

1. If all evidence is not considered, false conclusions are more likely. In particular, the wrong single line of evidence may be chosen because:
   a) assessors are likely to choose their own evidence, their favorite type of evidence, or the most easily obtained evidence, and
   b) assessors are likely to avoid evidence that is unfamiliar, difficult to understand, controversial, or contrary to preconceptions or self interest.
2. One line or type of evidence may compensate for the limitations of another. This is methodological replication. If you cannot study replicate systems, you can study the same system using multiple methods (Cavalli-Sforza, 2000). The body of results from the multiple methods may reveal the best-supported outcome.
3. Each type of evidence tells you something different about the situation being assessed. For example, one piece of evidence may tell you whether the cause preceded the effect while another tells you whether the cause was of sufficient magnitude.
4. The mistakes made in generating one piece of evidence may not be repeated in others and random errors tend to balance out. Hence, the average of multiple estimates is likely to be more reliable than any one estimate.
5. It is seldom the case that one piece or category of evidence is superior to all others in all respects.
6. Even if one piece of evidence is superior to all others, supporting evidence from other sources can increase confidence in results from the superior study.
7. "Because many judgments must be based on limited information, it is critical that all reliable information be considered" (The Presidential/Congressional Commission on Risk Assessment and Risk Management, 1997). That is, to not weigh evidence is to waste a limited resource — scientific evidence.
8. Including and weighting multiple risk estimates allows decision makers to make better informed decisions (Gray, 1994).
9. Weighing evidence makes risk assessment consistent with legal practice (Walker, 1996; Krimsky, 2005) in which evidence is weighed to determine which side is supported by the preponderance of evidence.
10. Including all relevant evidence reassures stakeholders that evidence is not being ignored or covered up (National Research Council, 1994).
11. Weighing evidence is consistent with natural cognitive processes (Gold and Shadlen, 2002).
12. Weighing evidence is recommended by agencies and authoritative panels (U.S. EPA, 1998; National Research Council, 2009; The Presidential/Congressional Commission on Risk Assessment and Risk Management, 1997).

Appendix B. Reasons to not combine multiple pieces of evidence

Although we do not find them convincing, we must acknowledge that there are arguments against including all relevant evidence.

1. It is unnecessary because in most cases one piece of evidence is right or one type is best. Therefore, no weight should be given to the others.
2. It is qualitative and subjective.
3. It is too difficult or time consuming.
4. It is too complicated and confusing.
5. It dilutes good evidence with poor evidence.
6. Traditional knowledge should not be weighed with scientific results (Chapman 2007).

Some authors, who advocate particular quantitative methods for inference, criticize WoE, as less rigorous than their methods (Newman et al., 2007; de Zwart et al., 2009). Although quantitative methods using one type of evidence may be more rigorous, they are not more likely to be correct, because they often discard information that does not fit their methods. Rather, we find practical reasons for rejecting WoE to be more compelling. If one very strong and high quality piece of evidence is available that adequately answers the assessment question, use it. There is no need to use a complex method if it will not improve the assessment or its acceptability to the decision makers or stakeholders.

References

Agency for Healthcare and Quality. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment 47. Washington, DC: U.S. Public Health Service; 2002.

Akaike H. Information theory as an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second International Symposium on Information Theory. Budapest: Akademia Kiado; 1973. p. 267–81.

Anderson DR. Model based inference in the life sciences: a primer on evidence. New York, NY: Springer; 2008.

Bellucci C, Hoffman G, Cormier S. An iterative approach for identifying the causes of reduced benthic macroinvertebrate diversity in the Willimantic River, Connecticut. EPA/600/R-08/144. Cincinnati, OH: U.S. Environmental Protection Agency; 2010.

Burkhardt-Holm P, Scheurer K. Application of the weight-of-evidence approach to assess the decline of brown trout (*Salmo trutta*) in Swiss rivers. Aquat Sci 2007;69: 51–70.

Burton GA, Batley GE, Chapman PM, Forbes VE, Smith EP, Reynoldson T, et al. A weight-of-evidence framework for assessing sediment (or other) contamination: improving certainty in the decision-making process. Hum Ecol Risk Assess 2002;8:1675–96.

Cavalli-Sforza L. Genes, peoples, and languages. New York: North Point Press; 2000.

Chapman PM. The sediment quality triad approach to determining pollution-induced degradation. Sci Total Environ 1990;97/98:815–25.

Chapman PM. The sediment quality triad: then, now and tomorrow. Internat J Environ Pollut 2003;13:351–6.

Chapman PM. Traditional ecological knowledge (TEK) and scientific weight of evidence determination. Mar Pollut Bull 2007;54:1839–40.

Chapman PM, McDonald BG, Lawrence GS. Weight-of-evidence issues and frameworks for sediment quality (and other) assessments. Hum Ecol Risk Assess 2002;8: 1489–515.

Cormier SM, Paul JF, Spehar RL, Berry WJ, Shaw-Allen P, Suter II GW. Using field data and weight of evidence to develop water quality criteria. Integr Environ Assess Manag 2008;4:490–504.

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

Cormier SM, Suter II GW. A framework for thoroughly integrating environmental assessments. Environ Manag 2008;42:543–56.

Dagnino A, Sforzini S, Dondero F, Fenoglio S, Bona E, Jensen J, et al. A "weight-of-evidence" approach for the integration of environmental "triad" data to assess ecological risk and biological vulnerability. Integr Environ Assess Manag 2008;4: 314–26.

Dale VH, Biddinger GR, Newman MC, Oris JT, Suter II GW, Thompson T, et al. Enhancing the ecological risk assessment process. Integr Environ Assess Manag 2008;4: 306–13.

de Zwart D, Posthuma L, Gevrey M, von der Ohe P, de Dekere E. Diagnosis of ecosystem impairment in a multiple-stress context — how to formulate effective river basin management plans. Integr Environ Assess Manag 2009;5:38–49.

Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman J, Ewigman B, et al. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. Am Fam Physician 2004;69:548–56.

Forbes VE, Calow P. Applying weight-of-evidence to retrospective ecological risk assessment when quantitative data are limited. Hum Ecol Risk Assess 2002;8: 1625–39.

Fox GA. Practical causal inference for ecoepidemiologists. J Toxicol Environ Health 1991;33:359–73.

Garnett ST, Brook BW. Modelling to forestall extinction of Australian tropical birds. J Ornithol 2007;148(Suppl. 2):S311–20.

Gates S. Review of methodology of quantitative reviews using meta-analysis in ecology. J Anim Ecol 2002;71:547–57.

Gerritsen J, Zheng L, Burton J, Boschen C, Wilkes S, Ludwig J, Cormier S. Inferring causes of biological impairment in the Clear Fork Watershed, West Virginia. EPA/600/R-08/146. Cincinnati, OH: U.S. Environmental Protection Agency; 2010.

Gold JI, Shadlen MN. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. Neuron 2002;36:299–308.

Good IJ. A causal calculus. Brit J Philo Sci 1961;11:305–18.

Good IJ. Good thinking: the foundations of probability and its applications. Minneapolis: U Minnesota Press; 1983.

GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1–8.

Gray GM. Complete risk characterization. Risk Perspective 1994;2:1–2.

Haake DM, Wilton T, Krier K, Stewart AJ, Cormier SM. Causal assessment of biological impairment in the Little Floyd River, Iowa, USA. Hum Ecol Risk Assess 2010;16: 116–48.

Hawkins CP, Norris RH, Hogue JN, Feminella JW. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecol Appl 2000;10:1456–77.

Hicks M, Whittington K, Thomas J, Kurtz J, Stewart A, Suter II GW, et al. Causal assessment of biological impairment in the Bogue Homo River, Mississippi using

the U.S. EPA's stressor identification methodology. EPA/600/R-08/143. Cincinnati, OH: U.S. Environmental Protection Agency; 2010.
Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions. Version 5.0.1. Cambridge, UK: The Cochrane Collaboration; 2008http://www.mrc-bsu.cam.ac.uk/cochrane/handbook/. Available at.
Hilborn R, Mangel M. The ecological detective: confronting models with data. Princeton, NJ: Princeton U. Press; 1997.
Hull RN, Swanson S. Sequential analysis of lines of evidence — an advanced weight-of-evidence approach for ecological risk assessment. Integr Environ Assess Manag 2006;2:302–11.
Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1-12.
Jones DS, Barnthouse LW, Suter II GW, Efroymson RE, Field JM, Beauchamp JJ. Ecological risk assessment of a large river-reservoir: 3. benthic invertebrates. Environ Toxicol Chem 1999;18:599–609.
Krimsky S. The weight of scientific evidence in policy and law. Amer J Public Health 2005;95:S129–35.
Lee GF, Jones-Lee A. Independent applicability of chemical and biologicalcriteria/standards and effluent toxicity tests. Nat Environ J 1995;5:60–3.
Linkov I, Long EB, Cormier SM, Satterstrom FK, Bridges T. Weight-of-evidence evaluation in environmental assessment: review of qualitative and quantitative approaches. Sci Total Environ 2009;497:5199–205.
Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care 2004;16:9-18.
Long ER, Chapman PM. A sediment quality triad: measures of sediment contamination, toxicity and infaunal community composition in Puget Sound. Mar Pollut Bull 1985;16:405–15.
Marcomini A, Suter II GW, Critto A. Decision support systems for risk-based management for contaminated sites. New York: Springer; 2009.
McDonald BG, Chapman PM. Selenium effects: a weight-of-evidence approach. Integr Environ Assess Manag 2007;3:129–36.
McDonald BG, DeBruyn AMH, Wernick BG, Patterson L, Pellerin N, Chapman PM. Design and application of a transparent and scalable weight-of-evidence framework: an example from Wabamun Lake, Alberta, Canada. Integr Environ Assess Manag 2007;3:476–83.
McPherson C, Chapman PM, DeBruyn AMH, Cooper L. The importance of benthos in weight of evidence sediment assessments — a case study. Sci Total Env 2008;394:252–64.
Menzie C, Henning MH, Cura J, Finkelstein K, Gentile J, Maughan J, et al. A weight-of-evidence approach for evaluating ecological risks: report of the Massachusetts weight-of-evidence work group. Hum Ecol Risk Assess 1996;2:277–304.
Moher D, Schulz KF, Altman D. The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001;285:1987–91.
Mumtaz MM, Furkin PR. A weight-of-evidence scheme for assessing interactions in chemical mixtures. Toxicol Ind Health 1992;8:377–406.
National Research Council. Science and judgment in risk assessment. Washington, DC: National Academy Press; 1994.
National Research Council. Science and decisions: advancing risk assessment. Washington, DC: National Academies Press; 2009.
Newman MC, Zhao Y, Carriger JF. Coastal and estuarine ecological risk assessment: the need for a more formal approach to stressor identification. Hydrobiolgia 2007;577:31–40.
Pope C, Mays N, Popay J. Synthesizing qualitative and quantitative health evidence: a guide to methods. Maidenhead, UK: Open University Press; 2007.
Smith EP, Lipkovich I, Ye K. Weight-of-evidence (WOE): quantitative estimation of probability of impairment for individual and multiple lines of evidence. Hum Ecol Risk Assess 2002;8:1585–96.
Stahl Jr RG. Issues addressed and unaddressed in EPA's ecological risk guidelines. Risk Policy Report April 1998;17:35–7.
Susser M. Rules of inference in epidemiology. Reg Toxicol Pharma 1986;6:116–86.
Suter II GW. Risk characterization for ecological risk assessment of contaminated sites. ES/ER/TM-200. Oak Ridge, TN: Oak Ridge National Laboratory; 1996.
Suter II GW, Barnthouse LW, Efroymson RE, Jager H. Ecological risk assessment of a large river-reservoir: 2. fish community. Environ Toxicol Chem 1999;18:589–98.
Suter II GW, Efroymson RA, Sample BE, Jones DS. Ecological risk assessment for contaminated sites. Boca Raton, FL: Lewis Publishers; 2000.
Suter II GW, Norton SB, Cormier SM. A methodology for inferring the causes of observed impairments in aquatic ecosystems. Environ Toxicol Chem 2002;21:1101–11.
Suter II GW, Cormier SM, Norton SB. Ecological epidemiology and causal analysis. In: Suter GW, editor. Ecological Risk Assessment. 2nd ed. Boca Raton, FL: CRC Press; 2007. p. p. 39–68.
Suter II GW, Cormier SM. When is a formal assessment process worth while? Hum Ecol Risk Assess 2010;16:1–3.
The Presidential/Congressional Commission on Risk Assessment and Risk Management. Risk assessment and risk management in regulatory decision-making. Washington, D.C: U.S. Government Printing Office; 1997.
U.S. EPA. Technical support document for water quality-based toxics control. EPA/505/2-90-001. Washington, D.C: U.S. Environmental Protection Agency; 1991.
U.S. EPA. ARCS assessment guidance document. EPA 905-B94-002. Chicago, IL: Great Lakes National Program Office; 1994.
U.S. EPA. Guidelines for ecological risk assessment. EPA/630/R-95/002F. Washington, DC: U.S. Environmental Protection Agency; 1998.
U.S. EPA. Stressor identification guidance document. EPA/822/B-00/025. Washington, DC: U.S. Environmental Protection Agency; 2000.
EPA US. Guidelines for carcinogen risk assessment. EPA/630/P-03/001F. Washington, DC: U.S. Environmental Protection Agency; 2005.
EPA US. Analysis of the causes of a decline in the San Joaquin kit fox population on the Elk Hills, Naval Petroleum Reserve #1, California. EPA/600/R-08/130. Cincinnati, OH: U.S. Environmental Protection Agency; 2009.
EPA US. A field-based aquatic life benchmark for conductivity in central Appalachian streams. (external review draft).EPA/600/R-10/023A. Cincinnati, OH: U.S. Environmental Protection Agency; 2010.
Walker VR. Risk characterization and the weight of evidence: adapting gatekeeping concepts from the courts. Risk Anal 1996;16:793–9.
Weed DL. Weight of evidence: a review of concept and method. Risk Anal 2005;25:1545–57.
Weed DL. Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. Int J Epidemiol 2000;29:387–90.
Weinhold B. Building on success: assessment categories for experimental noncancer endpoints. Environ Health Persp 2009;117:A113–5.
Wiseman CD, LeMoine M, Cormier S. Assessment of probable causes of reduced aquatic life in the Touchet River, Washington, USA. Hum Ecol Risk Assess 2010;16:87-115.
Wright JF, Sutcliffe DW, Furse MT. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Ambleside: Freshwater Biological Association; 2000.
Yoder CO, Rankin ET. The role of biological indicators in a state water quality management process. Environ Monit Assess 1998;51:61–88.