Weight of Evidence: A Review of Concept and Methods

Douglas L. Weed*

“Weight of evidence” (WOE) is a common term in the published scientific and policy-making literature, most often seen in the context of risk assessment (RA). Its definition, however, is unclear. A systematic review of the scientific literature was undertaken to characterize the concept. For the years 1994 through 2004, PubMed was searched for publications in which “weight of evidence” appeared in the abstract and/or title. Of the 276 papers that met these criteria, 92 were selected for review: 71 papers published in 2003 and 2004 (WOE appeared in abstract/title) and 21 from 1994 through 2002 (WOE appeared in title). WOE has three characteristic uses in this literature: (1) metaphorical, where WOE refers to a collection of studies or to an unspecified methodological approach; (2) methodological, where WOE points to established interpretative methodologies (e.g., systematic narrative review, meta-analysis, causal criteria, and/or quality criteria for toxicological studies) or where WOE means that “all” rather than some subset of the evidence is examined, or rarely, where WOE points to methods using quantitative weights for evidence; and (3) theoretical, where WOE serves as a label for a conceptual framework. Several problems are identified: the frequent lack of definition of the term “weight of evidence,” multiple uses of the term and a lack of consensus about its meaning, and the many different kinds of weights, both qualitative and quantitative, which can be used in RA. A practical recommendation emerges: the WOE concept and its associated methods should be fully described when used. A research agenda should examine the advantages of quantitative versus qualitative weighting schemes, how best to improve existing methods, and how best to combine those methods (e.g., epidemiology’s causal criteria with toxicology’s quality criteria).

KEY WORDS: Causal criteria; methods; quality criteria; risk assessment; systematic reviews; weight of evidence

1. INTRODUCTION

For at least 50 years, the phrase “weight of evidence” (WOE) has appeared in the scientific literature, most often in the context of risk assessment (RA). In the National Research Council’s 1983 “red book,” for example, the concept played an important role in describing key components of RA (especially hazard identification) and it continues to be used in many different kinds of publications, including federal government risk assessment guidelines and in countless published scientific papers from many different disciplines. “Weight of evidence” typically refers either to the interpretative methods of risk assessment or to claims about risk that emerge from their use. The central role that this concept plays in the practice of risk assessment makes it imperative that the many stakeholders be clear about its definition, its uses, and its implications. When we read that a “weight of evidence” approach was taken (a common and often undocumented statement in the literature), what exactly does that mean? What interpretative methods were employed? How were they applied to the available scientific evidence?

We are interested in answering questions like these in order to assist in the process of improving...
the methodological practice of risk assessment. At the center of the RA process is science and at the center of science are methods: the study methods used to generate scientific evidence and the methods used to summarize and interpret that evidence. Anyone familiar with this practice would likely agree that improvements in its interpretative methods are needed. These methods are used to summarize and synthesize evidence across several dimensions: large studies and small, strong studies and weak, old studies and new, human and animal studies, and studies involving human populations and studies of cellular systems. In addition to these obvious challenges, applying these methods involves values, both scientific and extrascientific, values that are not always made explicit. Uncertainty and underdetermination—the lack of definitive proof or disproof in science—are commonplace. Not uncommonly, claims about the WOE methods identified as such in the literature.

2. METHODS

“Weight of evidence” is a ubiquitous expression in biomedical science; a computerized PubMed library search using only that expression and without any limits generates a list of over 37,000 publications, dating from 1954 to the present. It follows that a systematic review of the WOE concept and methods requires some modifications to the methodologic guidelines for systematic narrative reviews designed to ensure that all publications on a topic are available for selection, summarization, and interpretation. We therefore constrained our search to identify relatively recent publications in which WOE was prominently featured.

We sought published articles in which the phrase “weight of evidence” appeared either in the title or the abstract using the National Library of Medicine’s search engine (PubMed). For the years 1994–2004 inclusive, 276 publications met those criteria. We selected for review all articles published in 2003 and 2004 (n = 71) as well as all publications from 1994 through 2002 in which “weight of evidence” appeared in the title (n = 21). These 92 publications were further categorized in terms of how “weight of evidence” was defined. Reference lists of these articles were reviewed to identify additional relevant documents (e.g., risk assessment guidelines from the Environmental Protection Agency and other government agencies as well as classic methodologic articles (e.g., Austin Bradford Hill’s classic 1965 article on causation)). These were used throughout the text for illustrative purposes.

3. RESULTS

“Weight of evidence” has at least three characteristic uses: metaphorical, methodological (with several subcategories), and theoretical, roughly in order of their relative prevalence. See Table I.

3.1. “Weight of Evidence” as a Metaphor

The most common use of the phrase “weight of evidence” is to refer to a body of scientific evidence that has been examined for some purported risk, without reference to any interpretative methodology. “Weight of evidence” in this context can therefore be considered symbolic or metaphorical; the phrase could be replaced by the words “summary interpretation of the evidence” or “synthesis of the evidence.” This category also includes those publications in which the results of a single study were reported as...
Table I. Uses of “Weight of Evidence” (WOE) in Current Practice (1994–2004)

<table>
<thead>
<tr>
<th>WOE Category</th>
<th>Description</th>
</tr>
</thead>
</table>
| Metaphorical (no method described) | • WOE collection of studies  
• Single study contributing to a WOE  
• WOE approach |
| Methodological        | • WOE method versus a “strength of evidence” approach  
• WOE method using “all” rather than a selected subset (e.g., standard test assay) of the evidence  
• WOE method pointing to other “established” or familiar interpretative methodologies  
• Systematic narrative review  
• Quality criteria for toxicologic studies  
• Epidemiology’s causal criteria  
• Meta-analysis  
• Mixed epidemiology-toxicology methods  
• WOE method employing a quantitative weighting scheme |
| Theoretical           | • WOE theory of pattern recognition in cognitive science  
• WOE and the court’s evidentiary gate-keeping role |

**WOE = weight of evidence.**

**Note:** Categorization arose from 92 published scientific papers in which “weight of evidence” appeared in the abstract ($n = 71$) in 2003 and 2004 or appeared in the title ($n = 21$) from 1994 through 2002.

We have not described but could be more or less inferred from a careful between-the-lines reading of our paper.” In sum, this metaphor is a kind of scientific shorthand, collapsible in this particular case to the simple acronym, WOE, often seen in print but rarely spoken.

This metaphorical category of “weight of evidence” also highlights an important problem in the current practice of risk assessment: lack of transparency; that is, a tendency to underreport, even omit, the details of the interpretative methodology used. We will return to this topic in the discussion.

3.2. “Weight of Evidence” as Methodology: General and Contrastive Approaches

The second category in Table I is methodological. In this literature, the phrase “weight of evidence” is sometimes used to refer to a methodological approach with a fairly simple premise: that all available evidence should be examined and interpreted. For example,

> “[t]he weight of evidence evaluation is a determination of what is a reasonable conclusion in view of all available information without numerical safety factors or uncertainty factors. . . . while exercising one’s best judgment.”

Interestingly, EPA’s account of guidelines for carcinogen risk assessment (98) conforms to this meaning. That document reserves the use of the term “weight
of evidence” for what is called a summary narrative, that is, a “single step after assessing all of the individual lines of evidence” whose purpose is to “summarize the results of the hazard assessment and provides a conclusion with regard to human carcinogenic potential.”

“Weight of evidence,” in this sense—using all the evidence—is sometimes compared to another, apparently less desirable, alternative that uses a subset of the evidence, sometimes referred to as a “strength of evidence” approach. For example:

Historically regulatory classification of a xenobiotic as a carcinogen has relied upon strength of evidence; that is, the degree of positive evidence from even a single study showing a statistically significant result. By contrast, weight of evidence considerations integrate together all toxicologic and mode of action information—positive, negative, and evidence on relevance to humans—that relate to the determination.(94)

In this example, the “strong” evidence is that which is both statistically significant and positive. In the next example, “strong” refers to unbiased epidemiological evidence:

In assessing the human data within the overall weight of evidence, determination about the strength of the epidemiologic evidence should clearly identify the degree to which the observed association may be explained by other factors, including bias or confounding.(98)

Finally, we also include here those examples in which a “weight of evidence” approach (meaning all the evidence) is contrasted with an approach that uses standardized tests. For example:

Emphasis on a weight-of-evidence approach to immunotoxicity evaluation as opposed to implementing a standard set of tests on every investigational drug.(80)

It is important to keep in mind that in many of the articles in this category, very little information was provided to define what is meant by “all” the evidence (i.e., whether issues of quality, peer review, or other standards were used to exclude studies from the risk assessment). In addition, no specific interpretative method may be described. The emphasis is primarily on ensuring that “all” rather than some evidence is interpreted in the RA process.

3.3. Familiar “Weight of Evidence” Methods

“Weight of evidence” can also be used to refer to well-known methods for summarizing and interpreting scientific evidence on health (and environmental) risks as well as methods for assessing clinical treatments and preventive services. For example:

Best evidence synthesis combines the strength of meta-analysis and traditional (narrative) reviews and provides reviewers with an approach to put forth conclusions about where the weight of the evidence lies.(102)

Analyzing the contribution of evidence from a body of human data requires examining available studies and weighing them in the context of well-accepted criteria for causation.(97)

Thus, a “weight of evidence” method may refer to systematic narrative reviews, to criteria-based methods of causal inference, to the statistical technique of meta-analysis, or to some combination of these well-known (and oft-debated) techniques, some more qualitative than others. Clinical reviews of the “weight of evidence” may point to the hierarchy of study designs commonly used to guide recommendations for medical treatments or preventive services.(85) Randomized clinical trials appear at the top of these lists (due to their ability to validly test a specific hypothesis) whereas case reports and expert opinion (in the absence of evidence) appear at the bottom of these lists.

3.4. Systematic Narrative Reviews

Systematic narrative reviews have received much attention in the past 20 years in the scientific and medical literature in response to careful analysis—“reviews of reviews”—revealing a general lack of clarity, transparency, and rigor in this important if underappreciated form of scientific publication.(103) Increasingly, scientific journals require authors of narrative reviews to follow guidelines; the Journal of the National Cancer Institute, for example, provides the following methodologic guidelines for review articles;(3)

1. Statement of purpose
2. Literature search methods
3. Inclusion and exclusion criteria for the literature reviewed
4. Criteria used for study validity and quality
5. Methods for summarizing and interpreting evidence
6. Criteria for conclusions and recommendations made

The purpose of a systematic review, beyond its obvious role in describing the “state of the science” (i.e., a summary of the studies to date), may be
to make research recommendations, to make claims about causality (or risk), or to make preventive (public health or clinical practice) recommendations. The stated purpose will then help to determine what methods are appropriate. For example, public health recommendations (but not claims about causation) could require cost-benefit analysis; claims about causation (but not research recommendations) could require the use of criteria-based methods of inference and/or meta-analysis when appropriate. The most salient point for this review of the concept and methods of WOE is that guidelines for systematic narrative reviews require the author(s) to state how they went about their business. As a result, readers can better assess whether the stated methods are appropriate given the purpose of the review, the extent to which the methods were used correctly, and, perhaps most importantly, whether the conclusions and recommendations were warranted.

A key step in any systematic narrative review is to determine which studies will be included in the application of the interpretative methods used and which will be excluded. These considerations typically appear in the methods section of the review, detailing the library search techniques used to identify the initial list of studies, often supplemented by careful examination of the reference lists of studies identified in the search. (See Section 2 of this article for an example.) Exclusions can be based on concerns about quality, relevance, or reliability. For example, clinical reviews may exclude individual case reports; reviews of public health topics may exclude small underpowered (in the statistical sense) studies often inadequately described in letters to the editor. Many systematic reviews exclude prior reviews published on the same topic although a reasonable case can be made to include them as a way to document and evaluate the basis for earlier claims about risk and how these claims may have changed in the face of new research.

3.5. Quality Criteria for Toxicologic Studies

In the WOE literature, Klimisch et al. describe an approach that addresses the quality of toxicological studies, an approach that could be used more generally to assign any scientific study into one of four reliability categories:

1. Reliable without restriction (i.e., conforming to good laboratory practices (GLP) or some other set of quality criteria)

2. Reliable with restriction (i.e., well documented and scientifically acceptable, but falling short of GLP)

3. Not reliable (not well documented or used unacceptable methods)

4. Not assignable (e.g., abstracts)

In this scheme, evidence considered reliable (with or without restriction) is subsequently used in the risk assessment; evidence judged to be “not reliable” or “not assignable” is not automatically included, but may be used on a case-by-case basis depending upon expert judgment. Note that this approach is still consistent with a WOE method using “all” the evidence, with some evidence weighted more reliable than others.

3.6. Epidemiology’s Causal Criteria

When epidemiological data are available in a particular risk assessment, criteria-based methods of causal inference are often used. Discussed in the epidemiological literature since the early 1950s, these so-called criteria continue to evoke spirited discussions. The most widely recognized are “Hill’s” criteria, appearing in a 1965 article on the causes of occupational diseases written by the British statistician, Austin Bradford Hill. One year earlier, a closely related list of causal criteria appeared in the 1964 U.S. Surgeon General’s Report on Smoking and Health. Hill’s now classic article provides a list of nine so-called criteria or what he called “considerations” for causation, given a body of statistically significant epidemiological evidence and some laboratory-based (biological) evidence. Put another way, Hill assumed the existence of a statistical association between exposure and disease before “applying” the following list of considerations to the body of available scientific evidence:

1. Consistency (of association)
2. Strength (of association)
3. Dose response
4. Temporality
5. Experimentation
6. Specificity
7. Biologic Plausibility
8. Coherence
9. Analogy

It is beyond the scope of this review to comprehensively discuss this important approach to causal inference in epidemiology and public health. From more theoretical inquiries as well as from studies on the use of these criteria in practice, it can be said that this is a mixed qualitative and quantitative approach
to examining a body of evidence.\(^{(106-108)}\) Temporality, specificity, coherence, and analogy, for example, are all qualitative concepts not easily expressed (nor satisfied by the evidence) in quantitative terms. Consistency and strength of association, however, have well-known quantitative interpretations; an assessment of strength, for example, involves estimating the summary magnitude of the relative risk estimate (typically greater than 1.0) across all studies, taking into account the impact of bias and confounding on that quantitative estimate.

Those who practice causal inference typically exercise considerable latitude in selecting the criteria to be employed in any specific application; it is common for users to select a subset of Hill’s criteria without justification or explanation. In addition, users of this approach often fail to describe the “rules of evidence” assigned to each criterion, i.e., what characteristics of the evidence would lead the user to say that the criterion has been satisfied. For example, the “rule” for strength of association involves what magnitude of a summary effect measure (odds ratio or relative risk) should be considered “weak.” No consensus has emerged on where the threshold value for “weakness” lies; some have argued for 2.0, others point out that today’s larger epidemiological studies can reliably detect relative risk values less than 2.0.

It is important to point out that causal criteria can be considered a “weight of evidence” methodology using implicit weights: for example, criteria ignored in an analysis are weighted “zero.” On the other hand, some criteria are almost always used; in causality assessments in cancer epidemiology the criteria of consistency, strength, dose response, and biologic plausibility are almost always used together.\(^{(106)}\) Textbook descriptions, however, often emphasize additional criteria, most notably, temporality and specificity. Put another way, there is some evidence of mismatch between how this method is used in practice and how it is “supposed” to be used in theory.

### 3.7. Meta-Analysis

Meta-analysis, a more quantitative than qualitative approach to summarizing evidence from several human population studies, can also be considered a “weight of evidence” methodology. The contribution of the result of each individual study is weighted by the inverse variance of the effect estimate. Meta-analysis has an extensive theoretical and practical literature. Of interest here is the relationship of meta-analysis to the causal criteria, given that meta-analysis has been largely used for summarizing either epidemiological or clinical trial evidence.\(^{(109)}\) Meta-analysis alone is not sufficient for making claims about causation (or hazard); it can, however, provide a reproducible weighted average of the estimate of effect across several studies, and thus a measure of the consistency of that evidence (when heterogeneity can be ruled out). Meta-analysis also provides more precise estimates of the overall magnitude of the effect and the dose-response relationship, but the causal relevance of these estimates remains a matter of judgment.

### 3.8. Mixed Epidemiology-Toxicology Methods

The use of the causal criteria (and meta-analysis when appropriate) brings up one of the most intriguing and relevant problems for risk assessment: how to combine epidemiological evidence with animal model studies and other forms of laboratory-based biological evidence. “Biological plausibility” is the criterion in Hill’s original list that attempts to do this; yet in practice, it has three very different (and increasingly rigorous) interpretations:\(^{(110)}\)

1. A biologically plausible association is one for which a mechanism can be hypothesized, but for which no biologic evidence exists.
2. Simply suggesting a mechanism for a factor-disease association is insufficient to satisfy the criterion of “biologic plausibility.” Some lab-based evidence supporting the mechanism is also necessary.
3. An association is considered biologically plausible if there is sufficient evidence to show how the factor influences a known disease mechanism.

Toxicologists and others in the risk assessment community have proposed more detailed evidentiary considerations for combining human and animal evidence. One such method can be found in Proctor et al.\(^{(67)}\) These authors, in their examination of the carcinogenic nature of ingested hexavalent chromium, use a weight of evidence combinatorial approach derived from the Environmental Protection Agency’s Guidelines for Carcinogen Risk Assessment.\(^{(96)}\) The authors list eight considerations for human evidence (the first involving “multiple independent studies with consistent results;” the other seven described as “causal criteria”), then five considerations for animal evidence, and then an additional six considerations
under the heading “other key evidence.” See Table II, reproduced from Proctor et al.\(^{(67)}\) In summary, “weight of evidence” in this particular example (and in the EPA’s guidelines) refers to a criteria-based method of causal inference similar to but not identical with that of Austin Bradford Hill\(^{(99)}\) coupled with a number of additional considerations to be used to judge nonhuman evidence.

### 3.9. Weight of Evidence as Methodology: Quantitative Weighting Schemes

“WOE” may also refer to methods that quantitatively weight scientific evidence; three recent examples are briefly described.\(^{(10,57,111)}\) Calabrese et al.\(^{(10)}\) proposed a “toxicologically based weight-of-evidence methodology” for ranking chemicals on their endocrine disruption potential. Each candidate chemical is scored on each of the following:

1. **Multistage process of endocrine disruption.** Specifically, how many stages of the multistage process does the chemical disrupt? The greatest weight is assigned to the final state: clinical manifestations.
2. **Phylogenetic considerations.** Specifically, how close is the test species to the target species?

<table>
<thead>
<tr>
<th>Human evidence</th>
<th>Animal evidence</th>
<th>Other key evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Multiple, independent studies with consistent results</td>
<td>• Multiple independent studies with consistent results</td>
<td>• Robust data set available</td>
</tr>
<tr>
<td>• Causal criteria satisfied</td>
<td>• Same site across species and structural analogs</td>
<td>• Physical/chemical information</td>
</tr>
<tr>
<td>1. Temporal relation consistent with cause and effect</td>
<td>• Multiple observations by sex, species, and sites</td>
<td>• Structure-activity relation</td>
</tr>
<tr>
<td>2. Strong associations</td>
<td>• Severity and progression of lesions, including early-life tumors and malignancy, dose response, uncommon tumor type</td>
<td></td>
</tr>
<tr>
<td>3. Reliable exposure association</td>
<td>• Similar route of exposure to humans and relevant exposure levels</td>
<td>• Comparable metabolism and toxicity between species</td>
</tr>
<tr>
<td>4. Dose-response relationship evident</td>
<td>• Biomarker data</td>
<td>• Mode of action supports causal interpretation of human and animal evidence</td>
</tr>
<tr>
<td>5. Free from bias and confounding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Biologically plausible</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. High level of statistical significance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Table II.** Weight of Evidence Considerations for Determining Confidence of Causation\(^{a}\) |

\(^{a}\)See Reference 67.

3. **Model system.** Specifically, greater weight is assigned to *in vivo* rather than *in vitro* studies.
4. **Estrogenic potency.** Specifically, the most points are assigned to the highest potency measured (relative to the standard, estradiol).

Scores are added, divided by the maximum number, and multiplied by 100.

A more complicated example of using explicit weights can be found in Menzie *et al.*\(^{(57)}\) This approach is the product of a workshop on evaluating ecologic risks. In their words:

The weight of evidence approach is the process by which measurement endpoints are related to an assessment endpoint to evaluate whether a significant risk of harm is posed to the environment.

An example of an assessment endpoint is the community structure of a songbird population; a measurement endpoint might be the concentration of a potentially harmful chemical in sediment. In this approach, desired characteristics of measurement endpoints—called “attributes”—are listed. Each is assigned a scaling factor (0 to 1). Selected examples of attributes in this example include: strength of association, site specificity, quality of study, temporal representativeness, and “use of a standard method.” Then, each measurement endpoint (i.e., each relevant...
study) is scored with respect to each attribute (1 to 5). An overall weight is calculated for each measurement endpoint across all attributes; the weighted endpoints can then be compared to one another after a further ranking by their capacity to cause harm and by the magnitude of the response.

The extent to which chemicals have interactive effects when mixed (e.g., in toxic dump sites) provides another example of an explicit weighting scheme for scientific evidence. For each pair of chemicals suspected of being hazards, six weights are assigned to the body of evidence available, one for each of the following categories:

1. Direction of interaction (positive, negative, or no interaction)
2. Classification of mechanistic understanding
   a. Direct mechanistic data
   b. Mechanistic data on related compounds
   c. Inadequate or ambiguous mechanistic data
3. Classification of toxicological significance
   a. Direct demonstration
   b. Inferred or demonstrated in related compounds
   c. Unclear
4. Modifier: exposure duration and sequence (anticipated or different)
5. Modifier: in vivo versus in vitro data
6. Modifier: route of exposure (anticipated or different)

The direction of interaction is assigned either a 1.0 (positive), –1.0 (negative), or zero (no interaction). Each additional category (via its related subcategories) is assigned a weight between 1.0 and 0.32; these values were arbitrarily assigned so that the maximum possible weight, obtained by multiplying together the six individual weights, is 1.0 and the minimum weight for any body of evidence is 0.05. These weights are then incorporated into a calculation of the hazard index (HI).

3.11. “Weight of Evidence” in Theory

Cognitive science and the law provide theoretical interpretations of the concept “weight of evidence.” See the third general category in Table I. A “weight of evidence” theory has been suggested as a way to understand how visual patterns (a relatively simple example would be the sequence of letters: aabbcc) are perceived as regular phenomena; cognitive scientists refer to this regularity feature of patterns as “figural goodness.” The application of this theory to risk assessment is not readily apparent, although its use of the “weight of evidence” bears some resemblance to Bayesian notions.

In the law, it has been suggested that the gatekeeping role of the American courts (regarding scientific evidence) may provide a conceptual framework for a “weight of evidence” approach to risk assessment. Four concepts support this framework:

1. Relevance (the extent to which any single piece of evidence could have the tendency to make a fact more or less probable)
2. Reliability (the extent to which the evidence is of a sort reasonably relied upon to form an opinion or inference)
3. Sufficiency (the threshold “weight” of the totality of the evidence needed to infer a claim)
4. Standard of Proof (levels of proof needed for the sufficiency of different types of legal opinions or inferences, e.g., in civil versus criminal cases)

4. DISCUSSION

Although primarily a scientific activity, risk assessment has important implications for commerce, public and environmental health, science policy, governmental regulations, and the law. “Weight of evidence,” as it appears in its various guises in the published scientific literature, is clearly connected to RA in many ways: to its interpretative methods, to the evidence used in those assessments, and to its theoretical foundations. Identifying (and solving) the problems that emerge in the use of WOE could have important, even profound, consequences for all sectors of society that RA impacts.

4.1. The Problem of Multiple Definitions and Uses

On the face of it, the most obvious problem is the multiplicity of WOE definitions and applications. This review has identified at least eight distinct yet interrelated uses of WOE (i.e., the eight subcategories in
Table I) ranging across metaphorical, methodological, and theoretical categories. Given that only a third (92/276) of the articles published in the past decade that featured WOE (as described in Section 2) were reviewed, it is possible that other meanings and uses exist. A more comprehensive review could test this hypothesis. Put another way, this study sample (i.e., the 92 articles selected) may not be representative of the entire WOE literature. It is fair to say, however, that this review has demonstrated the variability of the phrase “weight of evidence” and its many uses in current practice. WOE has no single meaning. Such variability suggests that efforts to “harmonize” the risk assessment process around the concept of WOE will be challenging.

An intermediate step along the way to harmonization would be to encourage authors to define what they mean by “WOE,” thus reducing the lack of transparency that plagues this literature. A practical solution to the problem would be to require authors to define WOE and to describe the details of the WOE methods used in their research. Journal guidelines and the peer-review process could help in making these changes. As noted above, some high impact journals currently require methods sections for reviews. One way to think about such a shift in the preparation and review of scientific publications is to consider risk assessments a form of systematic narrative review. As noted above, a systematic review includes a description of the literature search, exclusions and inclusions, interpretative methods, quality criteria, and the like.

4.2. The Problem of Different Kinds of Weights

It is important to point out that detailed descriptions of WOE methodology (as noted above) may not reduce the variability observed in practice; indeed, that variability may actually increase as more participants in the RA process voice their personal views on the meaning of “weight of evidence” and the specific methodologic choices within the array of methods applicable to risk assessment. As just one example, consider the various ways this review has revealed that evidence can be weighed.

1. Weighing individual studies on grounds of quality or reliability
2. Weighing individual studies on their capacity to test a causal hypothesis (e.g., by study design type)
3. Weighing summary characteristics of evidence (e.g., using some causal criteria, ignoring others; meta-analysis inverse variance weights of summary effect measures)
4. Weighing human evidence relative to animal evidence

With so many different interpretations and applications of the concept of “weighing,” it should be clear that explicit descriptions of a “weight of evidence” approach used in any single risk assessment will likely add to rather than reduce the observed variability.

In addition, there is the question of how best to go about weighing. These weights can be either qualitative or quantitative and it is not immediately obvious which approach is better. Certainly, in a statistical technique such as meta-analysis, the inverse variance (quantitative) weighting of effect measures makes good sense. On the other hand, arbitrarily assigning numerical weights to evidentiary criteria does not have a strong theoretical foundation and may not improve decision making.

4.3. Judgment, Weight of Evidence, and Risk Assessment

Another concern is the role of judgment in WOE approaches to risk assessment. Many who write about WOE methods (in theory or in the practice of risk assessment) emphasize judgment. Yet why is it so important? One line of argument goes like this: it seems reasonable to assume that if we can agree on a particular WOE method, RA decision making may improve. But even in the face of such agreement, this method—part qualitative, part quantitative, containing several weighting procedures, and different kinds and qualities of evidence—will never determine the outcome. That is too much to ask of any method, given that the outcome is, at its core, a decision regarding whether the purported risk is in fact a risk, a hazard, that is, something that causes harm to health or to the environment. The method, then, does not (cannot) determine the outcome; the method requires judgment. Metaphorically, judgment is a kind of intellectual glue, cementing together the evidence and the methods.

Given the essential role for judgment in the RA process, it will be important to understand how it is obtained, fostered, measured, and evaluated. How values impact judgment will require careful analysis.

5. A FINAL COMMENT ON THE FUTURE OF “WEIGHT OF EVIDENCE” IN RISK ASSESSMENT

The risk assessment community is faced with three choices regarding the role of “weight of evidence” in its future.
Option 1: Encourage (even demand) that the WOE concept and its methods be fully described when used. The goal of this approach is to work toward a consensus on the meaning and methods of weight of evidence, such that a recognizable standard can be created for and accepted by the risk assessment community. Reaching this goal will require more than full disclosure of meaning and methodology. A research agenda will need to be developed that examines issues such as: the advantages and disadvantages of quantitative and qualitative weighting schemes, how to improve existing interpretative “WOE” methods, and how best to combine those methods. Some of the most obvious problems to be solved were described above.

Option 2: Interpret the diversity of views and lack of clarity on WOE as evidence that the concept is a passing metaphorical fancy, not really an appropriate overarching focus for risk assessment and the sectors of society it serves. Develop instead a research agenda centered upon the familiar interpretative methods that, along with the evidence and expert judgment, form the foundation of RA, such as causal criteria, meta-analysis, and various mixed epidemiology/toxicologic approaches, including the EPA’s approach to risk assessment methodology. Reserve the use of the term “weight of evidence” to the specific (and still important) activities of actually weighing evidence using quantitative and/or qualitative schemes, including weighing bodies of evidence of different types (e.g., human versus animal).

Option 3: Accept the diversity of views on and uses of the WOE concept and methods. Encourage the community to describe its meaning and the methods employed, allowing for (but not advocating) a consensus to develop but expecting at best that a diminution in diversity will ensue.

This article has embraced the first of these options. There is a case to be made, however, for Option 2, preserving a highly specific and literal interpretation of WOE, applied only when evidence is actually weighed. There is still much to be done in this more limited (but more precise) interpretation of WOE.

ACKNOWLEDGMENTS

The advice and assistance of the Risk Assessment Methodology Committee of Health and Environmental Sciences, Inc (HESI) is gratefully appreciated.

REFERENCES


105. Surgeon General’s Advisory Committee on Smoking and Health. (1964). *Smoking and Health: Rockville, MD: U.S. Public Health Service (DHHS publication no. (PHS) 1103).*


