The Weight of Scientific Evidence in Policy and Law

Sheldon Krimsky, PhD

The term "weight of evidence" (WOE) appears in regulatory rules and decisions. However, there has been little discussion about the meaning, variations of use, and epistemic significance of WOE for setting health and safety standards.

This article gives an overview of the role of WOE in regulatory science, discusses alternative views about the methodology underlying the concept, and places WOE in the context of the Supreme Court's decision in *Daubert v Merrell Dow Pharmaceuticals, Inc* (1993). I argue that whereas the WOE approach to evaluating scientific evidence is gaining favor among regulators, its applications in judicial processes may be in conflict with some interpretations of how the Daubert criteria for judging reliable evidence should be applied. (*Am J Public Health.* 2005;95:S129–S136. doi: 10.2105/AJPH.2004.044727)

In the narratives describing the historical development of natural science, nothing captures the drama of discovery as effectively as the "crucial experiment" (an *experimentum crucis*). For it is such an experiment, according to most historical accounts, that finally resolves competing explanations and/or theories, bringing to a close contested schools of thought. For example, history of science texts tell us that it was a crucial experiment that put to rest the theory of spontaneous generation in favor of the germ theory of disease and that launched a critical blow to the Phlogiston theory of combustion. Also widely acclaimed as a crucial experiment in the early part of the 20th century were the measurements made by British physicists, among them Sir Arthur Eddington, of the sun's rays during a solar eclipse. From their measurements they concluded that light bends in a gravitational field, which provided evidence in support of Einstein's over Newton's theory of light.[1] Such experiments have gained iconic status in the history of science.

But there is a significant and lively debate among philosophers and historians on whether it is meaningful to talk about "crucial experiments" in science. Sir Karl Popper believed that crucial experiments could play a role in falsifying scientific theories ("It should be noted that I mean by a crucial experiment one that is designed to refute a theory (if possible) and more especially one which is designed to bring about a decision between two competing theories by refuting (at least) one of them—without of course, proving the other.")[2] In contrast, Pierre Duhem and Thomas Kuhn were leading voices against the view that scientists are influenced by "crucial experiments" in deciding between competing paradigms.

Notwithstanding this debate, I believe there are influential or determinative experiments that crystallize a new scientific consensus, particularly in fields like physics, chemistry, and engineering.

However, it is very rare to find determinative experiments in environmental health sciences. A single, well-constructed experiment almost never resolves a critical issue on the cause of a disease, especially but not exclusively, diseases resulting from exposure to toxic substances. Rothman provides an example where the etiology of "toxic shock syndrome" was resolved through a crucial experiment.[3] As long as we do not permit controlled experiments where we would intentionally harm a human subject, when there are no possible benefits to them, for the mere sake of scientific inquiry, no single experiment can provide decisive data on the effects of a foreign substance on a human group. In so far as we depend on a number of experiments, some with greater statistical or explanatory power than others and information from diverse forms of evidence, we need to have some way of aggregating or weighing the results across different modalities of evidence.

The term "weight of evidence" (WOE) is used to characterize a process or method in which all scientific evidence that is relevant to the status of a causal hypothesis is taken into account. In criminal law, juries are given the responsibility to decide the WOE in regards to guilt or innocence. Judges weigh the evidence of legal precedent in justifying their rulings. Clinicians use a form of WOE in making diagnoses, and judges may defer to it when they offer opinions on the reliability of evidence. In the policy sectors of government, regulatory agencies or risk analysis panels use WOE to assess the total value of the scientific evidence that a substance may be dangerous to human health. Sometimes the term is used as if there were some algorithm or rational decision process by which the "weighing of evidence" is accomplished. Other times, the term WOE refers to nothing more than a subjective assessment on the part of a reviewer who takes relevant data from a given body of published research into consideration in order to ascertain whether a hypothesis is more likely to be true than false.

A distinction has been made between WOE and "strength of evidence" (SOE).[4] The latter is associated with the gravitas and relevance of information related to a specific indicator, such as the number of tumors produced in animals. In contrast, WOE includes all varieties of evidence, positive and negative, mechanistic and nonmechanistic, in vivo and in vitro, as well as human and animal studies. In risk assessment, the trend has been to widen the lens of relevant empirical and theoretical evidence, thus moving from approaches that utilize "strength of evidence" to those that utilize WOE. In this article I shall speak exclusively of WOE and assume that it encompasses the use of strength of evidence.

The WOE approach has been introduced into ecological risk assessment since the early 1990s in response to the need for better risk analyses of Superfund sites and impacted natural ecosystems.[5,6] One consensus report on WOE defined it as "the process by which multiple measurement endpoints are related to an assessment endpoint to evaluate whether a significant risk of harm is posed to

the environment."[7] In his widely cited book *Ecological Risk Assessment*, Suter notes the significance of WOE in evaluating different classes of evidence generated by alternative ecological models. He wrote, "the separate lines of evidence must be evaluated, organized in some coherent fashion, and explained to the risk manager so that a weight of evidence evaluation can be made."[8]

A number of benefits have been attributed to a WOE framework in regulatory decisions. Walker[9] cites three objectives of a WOE analysis: (1) it provides a "clear and transparent framework" for evaluating the evidence in a risk determination; (2) it offers regulatory agencies a consistent and standardized approach to evaluating toxic substances; and (3) it helps to identify the discretionary assumptions in risk determinations from experts. However, in selecting a WOE approach a certain number of nontestable *a priori* assumptions must be adopted, which may narrow the scope of scientific opinion and consensus on how different modalities of evidence should be aggregated, thereby failing to meet Walker's objective.

I begin with the observation that there is virtually no discussion in the scientific literature of the epistemic meaning of WOE. We cannot tell whether it is used as a methodology, a heuristic, a ranking system, or simply a subjective process of setting a causal threshold for cumulative indirect evidence. In the spirit of these questions, this article will do the following: (1) discuss the problem of aggregating different forms of evidence; (2) review uses of WOE in health science publications; (3) examine some applications of WOE by federal agencies; and (4) discuss how WOE enters judicial proceedings, particularly in the context of the admissibility of expert witnesses.

In this discussion, I shall argue that the concept of WOE, as it is currently applied in the health sciences, largely involves a qualitative approach to rating and assessing the aggregation of different forms of scientific evidence in relationship to a causal hypothesis. Currently, qualitative or quantitative frameworks that guide the use of a WOE method are more or less *a priori* heuristics that adopt certain norms about the status and relevance of alternative types of information, but their application largely depends on the tacit expertise of scientific evaluators. Moreover, no canonical frameworks for weighing scientific evidence have emerged. When experts use the term WOE in publications or in the courtroom, they are almost always referring to the outcome of a process in which scientists, working as individuals or in groups, examine a body of relevant scientific studies on the relationship between a compound and a disease outcome. These scientists, operating within an accepted framework, apply their tacit knowledge of a field to reach a "yea," "nay," or "probabilistic conclusion" about the relationship between the compound and a disease outcome. Most applications of WOE in support of public policy that are cited in the literature seem to (by inference or lack of specification) use a process methodology that is low on transparency and high on subjectivity.

MODALITIES OF EVIDENTIARY SUPPORT

If the modality of evidence considered for evaluating the human health effects of a chemical compound was of one type, let us say epidemiological studies, then the WOE might refer to how many studies support the hypothesis about health risks, what the individual power of a study is, or what the combined power of all the studies are in a meta-analysis. But each modality of evidentiary support is limited. For example, some scientists argue that epidemiological studies cannot demonstrate causation or mechanism, but only association.[10] Controlled animal studies do not yield direct information about people. Comparison of chemical structure between suspected and known toxins (known as structure activity analysis) does not provide information on how the chemicals function in a live organism. The term WOE has come to mean not only a determination of the statistical and explanatory power of any individual study (or the combined power of all the studies) but the extent to which different types of studies converge on the hypothesis. The WOE approach has become likened to "triangulation," namely, approaching the target question from many directions. Where no single epistemic modality (by which I mean a specific method or technique for acquiring information) can yield the definitive answer to an environmental health question, we refer to multiple epistemic modalities. The problem is: how does the evidence from these modalities add up? Does the accumulated data from several epistemic modalities mitigate against the insufficiency or shortcomings of evidence from a single epistemic modality?

A similar problem is presented in decision analysis. Multiattribute Utility Theory applies to cases where there are different dimensions of value associated with outcomes that, on the face of it, are not reducible to a common metric.[11] Thus, a decision to build a dam will have both positive and negative impacts of a social and ecological variety. These attributes are incommensurable, such as the additional hydropower gained by the dam and the loss of fish spawning in the river. In Multiattribute Utility Theory, a decision analyst develops a ranking and a utility function for the attributes and then undertakes an empirical investigation to determine the actual value of those attributes (how many fish will be lost and how much energy will be produced). Thus, the final outcome of applying Multiattribute Utility Theory is the aggregation of incommensurable variables through the adoption of a numerical schema.

For evaluating the human health effects of a chemical agent, there are different modalities of evidence, including human epidemiology, wildlife studies, experimental laboratory animal studies with rodents, primate studies, *in vitro* cell studies, and chemical structure activity analysis. Each type of study may provide some evidence, but each has its limitations. Human epidemiology may be valued highly for its relevance but less so for its scientific power, especially if the findings are unrelated to a postulated or known biological mechanism. Experimental animal studies may be dependable for the mechanistic knowledge they offer but questionable for their relevance to human cases.

If a chemical were known to be one of the causal agents responsible for a human disease, then we would expect a series of evidentiary pathways to converge on that conclusion. The chemical might manifest genotoxic or gross chromosomal effects in human cells studied *in vitro*. Or the chemical might be associated with wildlife abnormalities. But not all of the

evidence may be consistent with the result. It is possible that the chemical may be harmless to certain species and yet cause disease in others. Nevertheless, we gain confidence when one epistemic modality (rodent studies) is consistent with the results of other epistemic modalities (epidemiological studies) that make up the architectonic of evidence.

When we do not *know* whether a chemical causes a human disease but have the type of circumstantial evidence we would expect to acquire if the substance were known to cause the disease, then, building on a coherence theory of truth, the weight of the circumstantial or related evidence elevates our confidence in the hypothesis connecting the substance to the disease.

But how can we aggregate the evidence from a variety of modalities in a WOE approach, when no single study is definitive, and we cannot justifiably reach a conclusion from the limited evidence that a specific compound is likely the cause of human illness? Aggregating evidence across different epistemic modalities is like adding incommensurables. It can only be done if *a priori* constructs provide a basis for developing a common metric. More evidence, albeit inconclusive, may mean you are closer to demonstrating causality, but you cannot know by how much. And sometimes, different modalities of evidence do not converge on a single hypothesis and may even be inconsistent.

USES OF THE TERM WOE IN HEALTH SCIENCES

Usually WOE methods are applied when no single study and no individual modality of evidence (e.g., animal studies, human studies, *in vitro,* etc.) is conclusive in demonstrating a cause-effect relationship. Other times it may be used even when there is a solid epidemiological study showing a large increased risk from the exposure to some substance in order to build a stronger argument for regulation. Alternatively, WOE has been introduced to assess the "strengths and weaknesses of various measurements, and of the nature of uncertainty associated with each of them."[12] However, while the term is applied quite liberally in the regulatory literature, the methodology behind it is rarely explicated. We might be told that the decision to regulate was decided on the WOE rather than a crucial study demonstrating causality. Without an explication of how evidence is "weighed" or "weighted," the claim WOE seems to be coming out of a "black box" of scientific judgment.[13] One article that uses the term WOE in its title does not refer to the term elsewhere in the text.[14] Other articles assign scaling factors or qualitative terms to the evidentiary attributes.

A report issued by the US Agency for Toxic Substances and Disease Registry (ATSDR) of the Department of Health and Human Services stated that a necessary and reasonable alternative to causal determinations when establishing policy "may be a critical assessment of the overall "weight of evidence" of available science to serve as a surrogate of 'causality.'" The implication is that when causality is out of reach, we must use a surrogate called WOE. The ATSDR states: "'The weight-of-evidence' approach is an assessment method that includes reviewing site-specific doses, epidemiologic studies, and chemical-specific toxicity data to evaluate exposures and potential health effects in a community."[15]

In law, when direct material evidence of a crime or direct eyewitness testimony is not available, the term "circumstantial evidence" is used. This type of evidence comes in "bundles" and eventually must be "weighed" by the jury in its role of determining guilt or innocence. Each piece of the "bundle" of circumstantial evidence is insufficient to make a case. It is the entire "bundle" that convinces the jury. The concept of "circumstantial evidence" has a counterpart in environmental health.

The ATSDR uses the metaphor of the microscope as the rationale for applying the WOE approach to examining the human effects of polychlorinated biphenyls, by aggregating the results of disparate studies.

"Each of the studies, whether an epidemiologic study, a laboratory study, or the findings of wildlife biologists, could be compared to the lens of a microscope. Like the lens of a microscope, they can vary in terms of their resolving power and quality. They are also focused on different populations at different points in time . . . . Despite the limits and weaknesses of individual pieces of research, the collective weight of evidence indicates that certain polychlorinated biphenyl/dioxin-like compounds found in fish in the Great Lakes-St. Lawrence basin and elsewhere can cause neurobehavioral deficits."[16]

The concept of WOE is used widely but rarely explicated in the scientific and policy literature. Menzie et al.[17] state that, "although the term 'weight-of-evidence' is used frequently in ecological risk assessment, there is no consensus on its definition or how it should be applied." Often when WOE is cited, it is assumed that readers know what it means. Sometimes it is used to signify that evidence must reach a certain critical threshold before it can support regulation. Other times it refers to a process that examines both positive and negative studies and determines by the number and strength of the studies whether a causal relationship can be inferred. As regulatory bodies and scientific review panels depend increasingly on WOE methods, questions surrounding their use will inevitably enter litigation either in torts or contested regulations, where the elusive methodology behind WOE is ripe for Daubert challenges. Therefore, it is important to understand how WOE is being interpreted and what, if any, criteria are implicit or explicit in its application.

After an extensive review of the appearance of WOE in public health studies and regulatory documents, I have uncovered what I believe are four general uses of the term.

Intensive Literature Review

This interpretation of WOE takes the form of an intensive literature review, including some qualitative discussion of the studies, without assigning any weights to the studies. In the words of one medical group, "the more inclusive method of literature review involves assessing the 'weight of evidence' . . . the importance of the findings from each piece of research should be judged: this is termed 'Signal.' This is then balanced by the strength of the evidence or design weaknesses (termed 'Noise')."[18] Those who use the term WOE in this context assume that the reviewers have applied their expertise in interpreting both the quantity (number of positive studies) and the quality (statistical power) of the evidence without any explicit reference to a methodology. Readers may justifiably assume that the

reviewers are basing their interpretation of the aggregate value of the selected studies on their years of experience and tacit knowledge, rather than a fully developed analytical framework.[19]

Seat-of-the Pants Qualitative Assessment

According to this view, WOE is a vague term that scientists use when they apply implicit, qualitative, and/or subjective criteria to evaluate a body of evidence. Experts cite the general grounds for their opinion, but no specific parameters or methodologies are given for how the evidence is weighed. Thus, one might see general statements such as: A decision was made based on WOE standards, such as number of studies, strength of association, breadth and consistency of evidence, correlational power, and biological plausibility. A number of papers use the term WOE in the title without explaining a methodology or process that is used to do the weighting.

Sometimes the application of WOE involves a taxonomic presentation of studies. An example can be found in a 2001 study of "disinfection by-products."[20] These are the potential human hazards of chlorination. The authors created a table of evidence, which listed the summary data of studies for each adverse reproductive effect focusing on sample size, exposure assessment, relative risk, and odds ratios. They describe as the goal of the paper "to view the totality of the evidence in order to judge the overall weight of evidence concerning 'disinfection by-products' and reproductive and developmental effects."[21] After commenting on the categories listed in their taxonomy (odds ratios, uncertainties, and statistical significance), the authors conclude that the weight of evidence shows that low birth weight is not associated with "disinfection by-products" exposure. But the outcome they reach is not logically or rigorously derived by a methodology. The justification for the use of WOE could be enhanced if criteria for weighing the evidence were established at the outset.

Aggregating Diverse Evidentiary Modalities

In this particular use of WOE, an effort is made to aggregate the evidence through some combination of qualitative and/or quantitative techniques. For example, ATSDR incorporates an assessment method that includes reviewing site-specific doses, epidemiological studies, and toxicity data. A dose level injurious to humans is found from different types of research protocols.

The World Health Organization's Global Assessment of Endocrine Disrupting Chemicals uses "overall strength of evidence" as a qualitative evaluation of the outcome of concern and an exposure to a substance—assessing the strength of association as weak, moderate, or strong based on the qualitative values of each of five evaluation factors.[22]

Calabrese et al.[23] have proposed a quantitative ranking scheme to evaluate the endocrine effects of chemicals on human health. In their scheme endocrine disruption is considered a multistage process, where they assume the probability of achieving the end result, namely a clinical endocrine pathology, rises as one progresses through the process. The authors identify five levels of evidence that correspond with the stages of the multistage process, level 1 being the weakest and level 5 the strongest. Then they introduce a point system based on a geometric progression ($a + ar + ar^2 + ar^3 + ar^4$), which is normalized to 10 points when stage 5 is reached. Stages 1-4 are weighted as 0.6, 1.3, 2.5, and 5.0, respectively. The causal chain is neither linear nor deterministic. Stage 3 will not always reach stage 5, but only does so at a certain probability. Therefore, by attaching a weight to each stage, one is essentially assigning probability estimates to the evidence. Thus, these weights represent the probabilities that the specific stage will proceed to the next stage.

In theory it is possible to come up with weighting factors that are empirically verifiable. Let us suppose we are trying to determine whether a chemical is a human endocrine disruptor (that it will adversely affect the human endocrine system) and that there are five stages in the causal chain. If we had evidence that the chemical induced stage 5 effects, then we can declare the substance a human endocrine disruptor. Let us assume we have evidence the chemical induced a stage 3 effect. If we had a toxicological database with thousands of entries that allowed us to calculate the percentage of those chemicals that induced a stage 3 effect and the frequency among those that also induced a stage 5 effect, we would have an empirically based system to develop weighting factors.

However, there is no generally accepted rationale for such *a priori* weightings within a discipline. And if there were an accepted framework of weightings, the selection would be premised on achieving consistency among expert evaluators rather than on some consensus about causality.

WOE in Hypothesis Testing

Sometimes the term WOE refers to a methodology used for selecting between two competing hypotheses. In this context, authors refer to WOE in the quantitative evaluation of a hypothesis relative to the null hypothesis, based on *a priori* evidence.[24] It is common to find Bayesian methods of analysis used, where the probability of a hypothesis is based on current evidence and prior probabilities. This use of WOE is discussed in a published report that examines whether a DNA profile of a suspect is unique in the population.[25] A suspect's DNA is compared to the DNA found at the crime scene. The comparison is presented in the form of a probability estimate that the suspect's DNA and the DNA found at the crime scene are a perfect match. The weight of evidence is synonymous with the probability estimate.[26]

THE FEDERAL AGENCY USE OF WOE

US Federal agencies, as well as international agencies like the International Joint Commission,[27] have begun to incorporate WOE in both their internal risk assessment analysis and in their advisory processes where they engage with external experts. The approaches taken are usually qualitative and avoid compressing all of the data to some WOE numerical value. The ATSDR uses a WOE approach to evaluate the synergistic effects of chemical mixtures.[28] The ATSDR describes the objectives of and factors to consider in a WOE analysis in its *Public Health Assessment Guidance Manual*, without providing any details on how evidence is actually "weighed" or scaled.[29]

"A weight-of-evidence analysis involves the balanced review and integration of relevant

exposure, toxicologic, epidemiologic, medical, and health outcome data to help determine whether exposure to contaminant levels under site-specific conditions might result in harmful effects. . . . The goal of the weight-of-evidence analysis is to decide whether or not harmful effects might be possible in the exposed population by weighing the scientific evidence and by keeping site-specific doses in perspective."[30]

The Occupational Safety and Health Administration (OSHA) has incorporated WOE in its regulations. In OSHA's air contaminants standard the agency stated:

> In response to those commenters who argued that none of the studies described by OSHA presented sufficient dose-response data to be used as a basis for establishing a limit, the Agency emphasizes that it is not relying on any single study to determine that wood dust presents a significant risk of material health impairment. Instead, OSHA is making this determination on the basis of the findings in the dozens of studies reporting on the respiratory, irritant, allergic, and carcinogenic properties of wood dust. The Agency finds the results of these studies biologically plausible and their findings reproducible and consistent. It is true that some of these studies, like all human studies, have limitations of sample size, involve confounding exposures, have exposure measurement problems, and often do not produce the kind of dose-response data that can be obtained when experimental animals are subjected to controlled laboratory conditions. What the large group of studies being relied upon by OSHA to establish the significance of the risk associated with exposure to wood dust do show is that the overall weight of evidence that such exposures are harmful and cause loss of functional capacity and material impairment of health is convincing beyond a reasonable doubt.[31]

The EPA has used WOE in the assessment of Superfund sites, endocrine disruptors, and carcinogens. In its 1986 carcinogen assessment guidelines, the EPA introduced the term WOE to describe how it combined tumor findings in animals and humans as the principal elements of its WOE analysis to ascertain the carcinogenicity rating of a compound. In subsequent years, the EPA expanded its framework for a WOE evaluation of carcinogenicity by including a wider range of evidentiary sources beyond rodent and human epidemiological studies. In its recent policy document, "Proposed Guidelines for Carcinogen Risk Assessment"[32] the EPA stated that the agency would include structure-activity relationships (computer models of chemical substances) of other carcinogenic agents, modes of action of carcinogenic agents at cellular and subcellular levels, and knowledge of toxicokinetic and metabolic processes, in addition to the more conventional sources of evidence.

In 1986, the EPA issued a summary ranking of five grades for possible carcinogenic agents (A through E, A signifies that a chemical is a human carcinogen, B a probable human carcinogen, etc., until we get to E, not a carcinogen). In 1996, the EPA replaced the letters with three designations: known/likely a human carcinogen, cannot be determined, and not likely a human carcinogen. The change in the carcinogen guidelines accompanied a more expansive view of the acceptable sources of evidence, which the agency defines as a WOE approach. The EPA referred to a WOE evaluation as a "collective evaluation of all pertinent information so that the full impact of biological plausibility and coherence are adequately considered."[33]

The EPA notes that for a WOE approach, no single "weighing factor" determines the overall weight; moreover, "the factors are not scored numerically by adding pluses and minuses."[34] The factors are judged in combination, and there is no algorithm to aggregate the modalities and quality of evidence. The EPA does provide a guidance document that indicates when the weight goes up or down. Evidence is weighted more highly when time between exposure and outcome is short; there are consistent results in independent studies; a strong association exists between a compound and an effect; there are reliable exposure data; there is a dose-response relationship; there are no biases and confounding factors; there is a high level of statistical significance; and positive results are found in multiple species, sites, and sexes. The agency wrote: "Generally, the weight of human evidence increases with the number of adequate studies that show comparable results on populations exposed to the same agent under different conditions."[35] These qualitative weighting factors are consistent with the Bradford-Hill criteria for inferring causation.[36]

As previously noted, the EPA defined three descriptors for carcinogenicity (I, known/likely; II, cannot be determined; and III, not likely) and asserted that: "Applying a descriptor is a matter of judgment and cannot be reduced to a formula."[37]

What happens when you bring scientists together and ask them to apply a WOE qualitative heuristic and reach a conclusion on whether a substance is, is likely, or is unlikely to be harmful? Several studies have evaluated expert panels' use of WOE to determine whether there is consistency and convergence on the application of the criteria.[38] Some panel studies have introduced weighting factors for specific evidentiary modalities (e.g., in one case, studies that show direct mechanistic evidence for an effect receive a ranking of "1.0," whereas mechanistic data on related compounds receive a ranking of "0.71.") and measured the degree of consensus among experts.[39] The results in the study were mixed. The six teams of experts could not always agree on the direction of the interaction effect of two chemicals after reviewing and ranking the same data and applying the same *a priori* ranking scheme.

One of the key factors behind the reliability of science is the accuracy and replicability of measurement. The term WOE may suggest that a measurement is involved, but that is a false implication of the term. Weighing the evidence, in the way it is carried out by regulatory bodies, is based on human judgment. Such judgments are rarely, if ever, tested for interrater reliability. Those who are considered experts in "weighing" evidence are considered so because they have a good grasp of the type and variety of evidence that, according to standards in their discipline, are sufficient to justify a claim of cause and effect.

WOE IN LEGAL TESTIMONY

In law and public policy, three standards of evidence are generally recognized: preponderance, clear and convincing, and beyond a reasonable doubt. By preponderance of evidence, it is usually meant that a hypothesis under consideration need only be proven more trustworthy (more probable) than its negation. Most civil proceedings use a preponderance of evidence as a standard of proof.

A higher standard is found in the phrase "clear and convincing evidence." The

supporting evidence under this standard has to have more than a marginal edge over the alternative hypothesis. It has been described as evidentiary support "sufficiently strong to command the unhesitating assent of every reasonable mind."[40]

Finally, evidentiary criterion that meets the standard "beyond a reasonable doubt" is the highest burden and the one used in criminal trials to minimize false positives (convicting an innocent person).

In *Daubert v Merrell Dow Pharmaceuticals, Inc,* the US Supreme Court issued a ruling clarifying standards for federal judges on the admissibility of expert testimony in the courtroom. According to the *Daubert* standard, admissible expert testimony must meet a standard of relevancy and reliability. Moreover, some judges apply the standard to each study on which the expert relies, as well as the expert's overall conclusions. This interpretation of *Daubert* would have each study stand on its own. McGarity calls this interpretation of *Daubert* the "corpuscular approach to expert testimony."[41] He writes:

"If the plaintiff fails to establish the relevance and scientific reliability of a sufficient number of individual studies, the trial judge will exclude the expert's testimony and (in the absence of other relevant and reliable expert testimony on causation) grant the defendant's motion for summary judgment before the jury ever enters the picture."[42]

If McGarity is correct on how *Daubert* has been applied, then we will begin to witness a divergence between judicial and regulatory approaches to evidence. In regulation, the strands of evidence are not assumed to stand by themselves. Rather, they are seen as pieces of a puzzle. McGarity notes: "corpuscular approach effectively prevents the expert in toxic tort cases from applying the 'weight-of-evidence' approach that regulatory agencies universally employ in addressing the risks that toxic substances pose to human beings."[43] He likens the WOE approach in risk assessment to the jury's role in civil trials in weighing the quality and credibility of various testimonies.

Because there is no algorithm or canonical methodology for determining WOE in circumstances where no single study is definitive and there is no determinative experiment that can foster a consensus on causality, experts will exercise their judgment on the strength of evidentiary support when a subset of the pieces of the puzzle are assembled. The term puzzle solving is an apt metaphor for the practice of science. Thomas Kuhn used it in his classic study *The Structure of Scientific Revolutions* to describe the role of scientists engaged in normal research problems. "Bringing a normal research problem to a conclusion is achieving the anticipated in a new way, and it requires the solution to all sorts of complex instrumental, conceptual, and mathematical puzzles. The man who succeeds proves himself the expert puzzle-solver."[44] The metaphor has also been cited by Susan Haack in connection with the *Daubert* decision: " . . . scientists are like a bunch of people working, sometimes in cooperation with each other, sometimes in competition, on this or that part of a vast crossword. . . ."[45]

Two experts may easily disagree on the WOE. Who should decide whether the WOE has been met for a given hypothesis when there are contested views? After the corpuscular interpretation of *Daubert*, a judge applies the reliability standard to the admissibility of every piece of evidence in expert testimony without seeing it as part of the entire evidentiary record. By disqualifying the evidence as unreliable on its own weight, jurors may never hear the total weight of scientific evidence. McGarity concludes: "It is not at all clear that lay judges have the wherewithal to distinguish unreliable expert testimony from reliable testimony based on scientific studies that have been 'deconstructed' by paid industry consultants."[46]

When an agency reports, "according to a WOE determination chemical X causes (does not cause) a human disease," a number of possible presuppositions are implicit in the decision process including:

- a socially constructed heuristic for classifying studies or evaluating data,[47]
- an *a priori* numerical weighting scheme, and
- a constructed consensus from a panel of scientists through an interactive consultative process, such as the Delphi Process.

Studies that have measured the variance in expert judgments on the use of WOE in evaluating a hypothesis demonstrate that the application of WOE is not strictly a science but depends on the experience, as well as other tacit factors associated with the expert, such as their familiarity with or financial connection to the substance being evaluated. Experts who apply a WOE analysis to evaluate the human health hazards of a substance draw from their personal knowledge of similar compounds; situate the properties of the compound in a ranking system; and, based on the diversity and quality of the evidence, reach an informed, albeit subjective, judgment on whether the likelihood that the substance is the cause of a human disease is strong, moderate, or weak (e.g., the substance is a human carcinogen, a reproductive toxicant, or an endocrine disruptor).[48] Without an accepted canonical methodology or standard of weighing and combining information streams, and because subjective factors inevitably shape the outcome of the process, judges may not be in any better position than jurors to decide which WOE analysis used by expert witnesses is more credible or reliable.

CONCLUSIONS

As a metaphor, the term WOE turns a cognitive and subjective process, as in the case of juries "weighing the evidence," into something that connotes a purely rational and objective process. If we add the term "scientific" to the phrase, as in "weight of scientific evidence," it suggests even more precision by drawing its symbolic meaning from the terms "weighing" (from the weights and measures) and "science" (the most dependable self-correcting system for fixing belief). In this metaphor there is a triple dose of constructed rationality. Our first realization is that the "weighing instrument" for "weighing evidence" is human cognition, which has never been calibrated to the task. In fact, "weighing evidence" has little if anything in common with weights and measures. Secondly, evidence for a hypothesis generally appears in gradations, with the exception of the evidence from a crucial experiment. Generally, there is more or less evidence or conflicting evidence, or more or less uncertainty in the evidence. The approach that uses WOE applies a

method that treats evidence as a continuous variable and turns it into a dichotomous (below or above the threshold) or triadic variable: "yes," "no," or "probably." (I am indebted to Susan Haack for suggesting this point.) Third, the process of assigning values (qualitative or quantitative) to different evidentiary modalities or to studies of different quality within the same modality is generally constructed *a priori* (independent of empirically based evidence) for each specific case. Where frameworks or models have been developed for this purpose, they have not been standardized.[49]

Writing about the environmental etiology of childhood diseases, Debaun and Gurney highlight the essential role of a conceptual framework for weighing the evidence. "Informed recommendations require systematic assessments of the weight of evidence from available studies and placement of the studies into a conceptual framework that allows for available data to be reviewed in the context of epidemiology principles of causal inference."[50] Presuppositions within these frameworks about the value of different forms of evidence may bias the outcome of a WOE analysis. For example, some WOE approaches give higher weight to mechanistic information over epidemiological data. Where mechanistic knowledge may be unavailable for a particular substance, the value of excellent human epidemiological data may be reduced in the weighing schema because of *a priori* assumptions about evidence.

The use of all the relevant evidence for assessing the health effects of a substance is certainly an advance over restricting assessment to a few choice evidentiary modalities, where information derived from these modalities is scarce or the results highly uncertain. A legal process that rejects the use of WOE or restricts its utilization seems to be at odds with current practices in regulatory science, where knowledge about a potentially hazardous product is pursued through a triangulation of evidentiary streams. Moreover, the same legal processes that acknowledge the value of WOE must also acknowledge that its use is not a rigorous science and, therefore, must be open to public view and interpretation. When WOE is used consistently and uniformly by a regulatory body, it enables that body to develop a strong comparative approach for assessing the potential health and environmental effects of products. On the other hand, the transparency of WOE will enable jurors and stakeholders to fully grasp the norms and *a priori* assumptions that enter into the analysis. The Daubert decision and subsequent related procedures should neither serve as an excuse for "disbarring" WOE analysis in risk assessment nor prevent jurors from learning about the value and limitations that it may bring to litigation. ■

About the Author

The author is with the Department of Urban and Environmental Policy and Planning at Tufts University.

Request for reprints should be sent to Sheldon Krimsky, PhD, Department of Urban and Environmental Policy and Planning, Tufts University, Medford, MA 02155 (e-mail: sheldon.krimsky@tufts.edu).

This article was accepted July 27, 2004.

Acknowledgments

This work was supported in part by the Project on Scientific Knowledge and Public Policy.

Special thanks to the SKAPP Planning Committee, especially David Ozonoff, and participants at the Coronado Conference in 2003 for their constructive comments on an earlier version of the paper.

References

1. A.S. Eddington, *Space, Time and Gravitation.* (Cambridge, UK: Cambridge University Press, 1920).

2. K.R. Popper, *The Logic of Scientific Discovery.* (New York: Harper, 1959), 277.

3. K. Rothman. *Causation and Causal Inference in Epidemiology.* Draft paper delivered to the Coronado Conference on Scientific Evidence and Public Policy, March 3-4, 2003.

4. C.C. Willhite. "Weight-of-Evidence versus Strength-of-Evidence in Toxicologic Hazard Identification: Di(2-Ethylhexyl)Phthalate (DEHP)," *Toxicology* 160 (2001): 219-226.

5. J.M. Culp, R.B. Lowell, and K.J. Cash. "Integrating Mesocosm Experiments with Field and Laboratory Studies to Generate Weight-of-Evidence Risk Assessments for Large Rivers." *Environmental Toxicology and Chemistry* 19 (2000): 1167-1173.

6. L.W. Hall and J.M. Giddings. "The Need for Multiple Lines of Evidence for Predicting Site-Specific Ecological Effects." *Human and Ecological Risk Assessment* 6 (2000): 679-710.

7. C. Menzie, M.H. Henning, J. Cura, et al. "A Weight-of-Evidence Approach for Evaluating Ecological Risks: Report of the Massachusetts Weight-of-Evidence Work Group." *Human Ecological Risk Assessment* 2 (1996): 277-304.

8. G.W. Suter II, ed. *Ecological Risk Assessment.* (Chelsea, MI: Lewis Pub. Co, 1993), 86.

9. V.R. Walker, "Risk Characterization and the Weight of Evidence: Adapting Gatekeeping Concepts from the Courts." *Risk Analysis* 14 (1996): 793-799.

10. G.E. Dallal, Chief, Biostatistics Unit, *The Little Handbook of Statistical Practice* (The Jean Mayer USDA Human Nutrition Research Center on Aging, Tufts University), available at http://www.tufts.edu/~gdallal/LHSP.HTM.

11. R.A. Chechile "Probability, utility, and decision trees in environmental decision analysis," in *Environmental Decision Making: A Multidisciplinary Perspective.* (New York: Van Nostrand, 1991), 64-91.

12. Menzie, 1.

13. M.A. Ibrahim, G.G. Bond, T.A. Burke et al. "Weight of the Evidence on the Human Carcinogenicity of 2,4-D." *Environmental Health Perspectives* 96 (1991): 213-222.

14. R.L. Cooper and R.J. Kavlock. "Endocrine Disruptors and Reproductive Development: A Weight of Evidence Overview. *Journal of Endocrinology* 152 (1997): 159-166.

15. Agency for Toxic Substances and Disease Registry (ATSDR), "The Assessment Process: An Interactive Learning Program," available at http://www.atsdr.cdc.gov/training/public-health-assessment-overview/html/module2/sv18.html. Accessed March 24, 2005.

16. Agency for Toxic Substances and Disease Registry (ATSDR), "The Assessment Process: An Interactive Learning Program," available at http://www.atsdr.cdc.gov/training/public-health-assessment-overview/html/module2/sv18.html. Accessed March 24, 2005.

17. Menzie.

18. A. Edwards, G. Elwyn, K. Hood, and S. Rollnick "Judging the Weight of Evidence in Systematic Reviews: Introducing Rigour into the Qualitative Overview Stage by Assessing Signal and Noise." *Journal of Evaluation in Clinical Practice* 6 (2000): 177-184.

19. R.L. Cooper and R.J. Kavlock "Endocrine Disruptors and Reproductive Development: A Weight-of-Evidence Review." *Journal of Endocrinology* 152 (1997):159-166.

20. C.G. Graves, G.M. Matanoski, and R.G. Tardiff "Weight of Evidence for an Association between Adverse Reproductive and Developmental Effects and Exposure to Disinfection By-Products: A Critical Review." *RegulatoryToxicology and Pharmacology* 34 (2001): 103-124.

21. Ibid, 110.

22. World Health Organization, IPCS Global Assessment of the State of the Science of Endocrine Disruptors: Chapter 7. "Causal Criteria for Assessing Endocrine Disruptors—a Proposed Framework," 123-128 http://www.who.int/ipcs/publications/new_issues/endocrine_disruptors/en/. Accessed March 24, 2005.

23. E.J. Calabrese, L.A. Baldwin, P.T. Kostecki, et al. "A Toxicologically Based Weight-of-Evidence Methodology for the Relative Ranking of Chemicals of Endocrine Disruption Potential." *Regulatory Toxicology and Pharmacology* 26 (1997): 36-40.

24. E.P. Smith, I. Lipkovich, and K. Ye. *Weight of Evidence (WOE): Quantitative Estimate of Probability of Impact.* Working Paper. February 10, 2002.

25. D.J. Balding "When Can a DNA Profile Be Regarded as Unique?" *Science and Justice* 39 (1999): 257-260.

26. I.W. Evett, L.A. Forman, G. Jackson, et al. "DNA Profiling: a Discussion of Issues Relating to the Reporting of Very Small Match Probabilities." *Criminal Law Review* (May 2000): 341-355.

27. International Joint Commission (IJC). *Sixth Biennial Report on Great Lakes Water Quality.* Washington, DC: International Joint Commission, 1992.

28. H.R. Pohl, N. Roney, M. Fay, et al. "Site-Specific Consultation for a Chemical Mixture." *Toxicology and Industrial Health* 15 (1999): 470-479.

29. Agency for Toxic Substances and Disease Registry (ATSDR). *Public Health Assessment Guidance Manual*, Ch. 8, Health effects evaluation: weight-of-evidence analysis, available at http://www.atsdr.cdc.gov/HAC/PHAmanual/ch8p1.html. Accessed June 5, 2004.

30. Agency for Toxic Substances and Disease Registry (ATSDR), Chapter 8, http://www.atsdr.cdc.gov/HAC/PHAmanual/ch8p1.html, pp. 2-3. Accessed March 24, 2005.

31. 29CFR Par. 1910. Air Contaminants, Sec. VI. *Health Effects Discussion and Determination of Final Pel*, January 1989.

32. Environmental Protection Agency. Proposed Guidelines for Carcinogen Risk Assessment. *Federal Register* 61(79):17960-18011 (April 23, 1996). Hereafter, EPA 1996.

33. Ibid., 17981.

34. Ibid.

35. Ibid.

36. A. Bradford-Hill, "The Environment and Disease: Association or Causation?" *Proc. Royal Soc. Med.* 58 (1965): 295-300.

37. EPA 1996, 17985.

38. M.E. Anderson, M.E. Meek, G.A. Boorman, et al. "Lessons Learned in Applying the U.S. EPA Proposed Cancer Guidelines to Specific Compounds." *Toxicological Sciences* 53 (2000): 159-172.

39. M.M. Muntaz, P. Furkin, G.I. Diamond, et al. "Exercises in the Use of Weight-of-Evidence Approach for Chemical-Mixture Interactions." *Journal of Clean Technology, Environmental Toxicology, and Occupational Medicine* 5 (1996): 339-345.

40. V.R. Walker "Risk Characterization and the Weight of Evidence: Adapting Gatekeeping Concepts from the Courts." *Risk Analysis* 16 (1996): 793-799.

41. T.O. McGarity "Proposal for Linking Culpability and Causation to Ensure Corporate Accountability for Toxic Risks." *William and Mary Environmental Law and Policy Review*, Fall 2001.

42. McGarity, 7.

43. Ibid., 8.

44. T.S. Kuhn. *The Structure of Scientific Revolutions.* (Chicago: University of Chicago Press, 1962), 36.

45. S. Haack "An Epistemologist in the Bramble-Bush: at the Supreme Court with Mr. Joiner." *Journal of Health, Politics, Policy & Law,* 26 (2001): 217-248.

46. McGarity, 12.

47. A Baysian statistical approach to WOE is vien in: E.P. Smith, I. Lipkovich, and K. Ye. *Weight of evidence (WOE): Quanitative estimation of probability impact.* Unpublished manuscript. (Blacksburg, VA: Department of Statistics, Virginia Tech., 2002).

48. The ATSDR endorses "the use of a narrative statement incorporating "weight-of-evidence" conclusions in lieu of alphanumeric designations alone in conveying qualitative conclusions regarding carcinogenicity," available at: http://www/atsdr.cdc.gov/cancer.html, p. 9. Accessed March 24, 2005.

49. Op. cit. Ibrahim et al. 1991, p. 219.

50. M.R. DeBaun and J.G. Gurney "Environmental Exposure and Cancer in Children: a Conceptual Framework for the Pediatrician." *Pediatric Clinics of North America* 48 (2001): 1125.