Hypothesis-Based Weight of Evidence: An Approach to Assessing Causation and its Application to Regulatory Toxicology

Lorenz R. Rhomberg, Ph.D., FATS (Gradient)

Working Paper prepared for:

Methods for Research Synthesis:
A Cross-Disciplinary Workshop

Harvard Center for Risk Analysis

October 3, 2013

www.hcra.harvard.edu

***Corresponding author: lrhomberg@gradientcorp.com**

Disclaimer: The findings and conclusions of this paper are those of the author and do not imply endorsement by any component of Harvard University or other sponsors of this workshop. Comments should be directed to the author.

Acknowledgements: This article was prepared by the author with no external funding. Gradient provides consulting services to a variety of parties, including industry (*e.g.* utilities and engine manufacturers), governmental agencies, regulators, and law firms.

Hypothesis-Based Weight of Evidence: An Approach to Assessing Causation and its Application to Regulatory Toxicology

Lorenz R. Rhomberg

Gradient, Cambridge, MA

Abstract

Regulators are charged with examining existing scientific information and coming to judgments about the state of knowledge regarding toxicological properties of agents. The process needs to be seen as sound and objective. The challenge is that information is often far from definitive, containing gaps and outright contradictions. The particular results of studies must be generalized and extrapolated to apply to the target populations of the risk assessment. Existing weight-of-evidence approaches have been criticized as either too formulaic, ignoring the complexity and case-specificity of scientific interpretation, or too vague, simply calling for professional judgment that is hard to trace to its scientific basis. To meet these challenges, I discuss an approach – termed Hypothesis-Based Weight of Evidence (HBWoE) – that emphasizes articulation of the hypothesized generalizations, their basis and span of applicability. The approach stresses articulating what it is that makes data constitute evidence for a toxicologic concern in the target population. The common processes should be expected to act elsewhere as well – in different species or different tissues – and so outcomes that ought to be affected become part of the basis for evaluating success and defining the limits applicability. A compelling hypothesis is one that not only provides a common unified explanation for various results, but also has its apparent exceptions and failures to account for some data plausibly explained. *Ad hoc* additions to the explanations introduced to "save" hypotheses from apparent contradictions need to be recognized and the use of such *ad hoc* accommodations weakens the degree to which available data can be said to test the causal proposition in question. In the end we need an "account" of all the results at hand, specifying what is ascribed to hypothesized common causal processes and what to special exceptions, chance, or other factors. Evidence is

weighed by considering whether an account including a proposed causal hypothesis is more plausible than an alternative that explains all of the results at hand in different ways.

Keywords: hypothesis-based, weight of evidence, regulatory toxicology, causation, epidemiology, toxicity, systematic review

1 Introduction

The statutes that authorize government agencies to regulate the uses and permissible exposures to chemicals call on the regulators to make scientific assessments as to whether and under what conditions exposure to specific chemical agents may cause adverse consequences among specified target populations, either human populations or wildlife. That is, the regulators are charged with examining the existing scientific information and coming to judgments about the state of knowledge about toxicological properties of the agents, and these judgments are meant to reflect general well-informed scientific evaluation that should be widely acceptable as sound and objective.

The challenge is that the available scientific information on potential toxicity of chemicals is often far from definitive. Natural populations that may be exposed to the agent in question are subject to a wide variety of influences, and sorting out causal influences of single agents is difficult. Animal bioassay studies may be controlled, but applying their results requires extrapolation from observed effects at high doses to inferred possible effects in other species at much lower doses. *In vitro* studies and other investigations aimed at characterizing how agents interact with biological systems observe only potential components of a suite of processes that must interact to produce effects, and the consequences of any one component are uncertain. Importantly, the available data often contain information that is inconsistent or mutually contradictory. Some of this may be due to problems with the rigor or design of individual studies, but even well conducted studies often disagree, as when, for instance, the carcinogenic response to an agent in lifetime rodent tumor bioassays differs between rats and mice, or when effects appearing in animals seem to be unaffected in available human studies. In short, even though the statutes often presume that any scientifically competent analyst would be able to look at the available data and come to a clear and uncontroversial conclusion about the potential for the agent to cause toxicity, the facts of the matter are much more complex. A considerable degree of professional judgment is necessary to assess the degree to which available data support or do not support conclusions about toxicity on which regulatory decisions will be based.

The approach taken to this challenge is to employ a "weight-of-evidence" evaluation -- the application of professional judgment to consider the strengths and weaknesses of individual

studies, to compare and contrast their findings, and to try and reconcile or explain inconsistencies so as to arrive at a characterization of what potential toxicological properties are sufficiently supportable to justify the regulatory decisions that will be made. The challenge is for such a process to be sufficiently flexible to apply to a wide variety of arrays of data and patterns of agreement and disagreement, and at the same time sufficiently prescribed so that the results will not be seen as arbitrary, having consistent application of principles and standards of proof from case to case, applied in a way that is seen as transparent and objective.

In a review of the use of the weight-of-evidence concept in regulatory toxicology, Weed (2005) noted that the term is often used metaphorically with no proffered method, but that in some uses it does refer to a method, though it is important that each analysis specify what that method may be. The adequacy of existing weight-of-evidence approaches identifying chemical hazards in regulation has recently been called into question by a committee of the National Academy of Sciences (NRC, 2011) in its review of an assessment of formaldehyde carcinogenicity by the US Environmental Protection Agency (EPA), stating that "EPA might direct effort at better understanding how weight-of-evidence determinations are made with the goal of improving the process." In the present paper, I (1) analyze and comment on some of the challenges in arriving at a robust yet flexible framework, (2) note the epistemological questions that are raised in trying to infer toxicological causality from the kinds of information available, (3) note the differences between the context of such inference as a pure scientific question and the particular application to supporting regulatory action, (4) briefly review some existing approaches to defined weight-of-evidence evaluation procedures, with comments about their rationales, strengths, and weaknesses, and finally (5) propose and describe a method we have developed and applied -- Hypothesis-Based Weight of Evidence. I end by briefly describing some of the published applications of this method to some current questions in regulatory toxicology, demonstrating the method's ability to be useful in producing a transparent evaluation with an explicitly explained basis as one addresses different kinds of quandaries that present themselves in regulatory toxicological evaluation.

2 Challenges for Systematic Weight of Evidence Evaluation

In a way, weight of evidence is a very general challenge for all of science. What is particular about the application to regulatory toxicology is that the regulatory process cannot sustain the suspended judgment and ongoing questioning that characterizes the methods of pure science and that serves to direct future research directions; decisions are necessary, and evaluations of the extent of knowledge as it stands at a moment is needed. Moreover, the evaluations of the robustness of such evaluations need to be made by particular analysts authorized to make the judgments on behalf of the regulatory process. This raises questions about who is doing the judging, and it puts a premium on objective and operational methodology, transparency, and defined processes to ensure the legitimacy of the delegated authority to make judgment. There is a distinction between analysts making their own best professional judgments, based on their knowledge, expertise, and experience and the analysts serving as representatives of the larger body of scientific opinion, attempting to discern judgments that are not theirs alone but reflect the wider body of opinion and fairly representing the spectrum of views in any ongoing debates. Because opinions on the most supportable judgment vary among scientists, there is a tension between representing individual versus collective judgments in doing the weighing of evidence.

This tension is addressed through the structure of the weight-of-evidence process. We have recently (Rhomberg *et al.*, 2013) surveyed a number of existing weight-of-evidence frameworks and attempted to draw some insights into how and why they vary as they address the challenges of providing a way to ensure accountability and transparency of the judgment process, aim at making it transparent and consistent across cases, and deal with the challenges that judgments may receive from those who harbor different views. One can discern three basic tendencies among these methods.

One is to stress systematic review and presentation of relevant data according to rigorous and predefined criteria, noting study methodology strengths and shortcomings, ability to control extraneous influences, and other factors that bear on the reliability of results. The aim is to identify the most reliable results, done in the hope that the assembled data evaluated in this way will collectively point to the most reliable conclusion. The strength of this method is that it is very objective, operational, and transparent, but the challenge is that it does not easily deal with

seemingly reliable but nonetheless discordant results and it does not provide a clear means to integrate across markedly different lines of evidence.

A second tendency is to provide a set of rules or procedures for evaluating evidence, along with defined presumptive interpretations for given sets of outcomes, such as rules for resolving conflicts or preferring results from certain kinds of studies compared to others. In such approaches, the intent is to codify the collective wisdom of appropriate interpretation into the rules, so that different analysts would come to similar conclusions, interpretations would be consistent across applications, and each analysis could be held accountable for whether it followed the prescribed procedure and interpretations. The advantage is consistency, clarity, and transparency, and also that the collective expertise and judgment of the wider field of study can be brought to bear whenever the method is applied by technically competent analysts. The challenge is that the results are only as good as the success at codifying the collective wisdom into the rules, and it can devolve into application of conventional wisdom and rely overly much on precedent while failing to modify interpretations as the relevant science advances.

The third tendency is to rely on the personal judgments of panels of experts, guiding them procedurally but ultimately relying on individual professional judgments from experts recognized as having appropriate knowledge and insights into the scientific questions at hand. The advantage is flexibility of method for different kinds of data, and production of thought-through arguments that are tailored to the questions as they present themselves and answer the particular challenges of the case at hand, but the challenge is that the authority to make the judgments is delegated to the chosen judges and may not be easily explained to the wider public nor guaranteed to be consistent with other cases that might be deemed similar. To a large degree, most existing weight-of-evidence frameworks can be seen as tending toward one of these three poles, while trying to find ways to gain some of the advantages of the remaining approaches while minimizing their shortcomings.

3 Hypothesis-Based Weight of Evidence

We have developed our approach – the Hypothesis-Based Weight of Evidence (HBWoE) approach – in an attempt to gain the advantages of a rigorous systematic review of data while also producing issue-specific, thought-through arguments for the application of those data. The rationale of HBWoE is based on two realizations. The first is that when we apply results of a study to a question of a chemical's potential toxicity in the target (human) population, we are making a hypothetical *generalization* – we are asserting that there is something about the causal forces operating in the study that its subjects share in some way with the target population. It is this asserted commonality that makes the data from the study constitute evidence about the different question of toxicity-causing processes in the population whose risks we are assessing. As a generalization, the common processes might be supposed to act elsewhere as well – in different species or different tissues – and so outcomes that ought to be affected become part of the basis for evaluating the success of the hypothesized generalization. Defining the limits to and span of applicability of the asserted general causative process -- where it should happen, where it should not, and what accounts for the differences – is part of the evaluation of the bearing of evidence. When there are discordant data, they pose a challenge to the asserted generalization unless reasons for the discordance can be proposed, in which case those reasons become part of the body of assertions about the bearing of data that need to be evaluated. In short, the "hypotheses" of hypothesis-based weight of evidence are proposed reasons for why individual pieces of data should be considered as evidence for the overarching hazard question at hand. The second realization is that, in the end, we need some tentative account of all of the results in the array of studies at hand, including the ones that seem discordant. A compelling hypothesis is one that not only serves to provide a common unified explanation for various results, but one that also has its apparent exceptions and failures to account for some data at hand reasonably and plausibly explained. Moreover, a hypothesis is compelling when it constitutes this kind of explanation of the array of results at hand better than does any competing hypothesis.

That is, our approach is to evaluate the credibility of the hypothesized basis for an inference of potential human risk, in view of (1) the plausibility of the hypothesis (in view of what we know about biology and the invoked causal processes and their variation among species); (2) the degree to which we need to fill data gaps or complete inferences with assumptions that have not

been empirically verified; (3) the plausibility of those specific assumptions and inferences, that is, an assessment of the a priori likelihood that they would prove true if tested; (4) the degree to which specific hypothesized causative phenomena have indeed been observed rather than merely assumed or proposed; (5) the degree to which the causative elements have been tested and affirmed in other species or in other tissues; (6) the degree to which the observed lack of the same causative elements in other settings is associated with a lack of observed effect (*i.e.*, the complementary affirmation of its role); and (7) the degree to which the hypothesized elements have not merely been crafted to accommodate or explain apparent incongruities in the data that would seem to be inconsistent with a more straightforward version of the hypothesis. By making explicit the logic and content of the hypothesized basis for using an animal response as an indicator of potential human hazard, and by evaluating that logic using the numbered points above against all the available data, one can come to a transparently explained evaluation of how credible the proposed existence of the human hazard potential ought to be judged.

The procedure for HBWoE can be set out in the following steps: (1) systematically review individual studies potentially relevant to causal question at hand (*e.g.*, epidemiology, mode of action, pharmacokinetic, toxicology), with focus on evaluation of the quality of all individual studies (both negative and positive, of varying qualities); (2) within a realm of investigation (*e.g.*, epidemiology, animal toxicology, or mode of action studies), systematically examine the data for particular endpoints across studies, evaluating consistency, specificity, and reproducibility of outcomes; (3) identify and articulate lines of argument (or "hypotheses"), newly proposed or those already put forth (if available), that bear on the available data and discuss how available studies are used for each hypothesis to infer the existence, nature, or magnitude of human risk; (4) evaluate the logic of the proposed hypotheses with respect to each line of evidence to determine how well the hypotheses are supported by the available data; (5) evaluate the logic of the proposed hypotheses with respect to all lines of evidence together so that all of the data are considered and integrated and allowed to inform interpretation of one another; (6) describe and compare (if more than one hypothesis has been put forth) the various alternate accounts of the observations at hand, describing how well each overarching hypothesis is supported by all of the available data, discussing the uncertainties and inconsistencies in the data set and *ad hoc* assumptions required to support each hypothesis; (7) formulate discussion and conclusion

regarding the WoE, and proposed next steps. These steps are intended to provide general guidance on how to weigh all of the evidence in a systematic way, but are also intended to be flexible, because every causal question has a different data set that will require a somewhat different specific approach for presentation and systematic review of the data at hand, but should generally follow these seven steps.

4 Comparison of Hypothesis-Based Weight of Evidence to Other Systems

Given the theme of the symposium to which the present paper belongs -- cross-disciplinary comparisons of approaches to integrating complex datasets to make inferences -- it may be useful to examine how the case of assessing toxicological causation may be similar to or different from other related applications. In particular, evidence-based medicine, which has well developed formal methodologies, is sometimes held as an example that toxicology can follow, since there is a good deal in common between evidence-based medicine's task of assembling and integrating evidence on the efficacy of therapeutic approaches in preventing or treating physiological dysfunction and toxicology's task of assembling and integrating evidence on the potential to cause such dysfunction. An important difference, however, is that in evidence-based medicine, the studies being assembled and integrated are all designed to observe the question of ultimate interest directly -- that is, they are trials of the therapy being evaluated, designed to measure its efficacy directly on actual patients in the actual context of interest. The main question to be evaluated is whether the various studies consistently show a common effect, and the statistical power to evaluate and describe effects is increased through multiple observations of what is thought to be the same underlying phenomenon.

In contrast, toxicological evaluations entail the assembly of a diverse array of studies, most or even all of which are not direct observations of the central motivating question about the ability of an agent to cause adverse effects at prevailing or anticipated exposure levels as they are experienced among members of a target population (which is often, but not always, the general human population). We observe effects in animals not to assess potential risks to mice and rats, but to use these mammals as surrogates for what might be expected to happen in humans. We test high exposure levels to ensure statistical power to detect effects that might be possible but

rare at lower exposures, as well as to know what happens when one pushes the living system beyond its limits of tolerance and accommodation, but we wish to characterize the potential for such effects at much lower exposures that cannot practically be tested. We do studies with rigid schedules of constant exposure to single agents when we seek insights into the possible effects of intermittent and time-varying exposures that may interact with other agents in the environment that also have inconstant patterns of presence or concentration over time. We investigate biochemical and cellular changes, often in highly artificial constructs designed to enable isolation and measurement of particular biological processes, that are not themselves toxicity but which we think may be components of more complex pathways of toxicological causation or at least may be biomarkers for other unobserved underlying mechanistic effects that could be relevant to such pathways.

In short, the issue is less one of assessing consistency among studies (in the sense of replication) than of finding the way in which each component bears on the ultimate question about toxicological causation. This is not to say that the development of standards for the process of identifying relevant studies, abstracting information in an objective, thorough, and consistent way, and evaluating study strengths and weaknesses -- which are all well developed in evidence-based medicine -- are not also important for toxicological weight of evidence. But such systematic review of the data will not by itself make the appropriate conclusions evident, and the challenges of the evaluation and integration component need to be addressed.

5 Integration of Evidence

The integration question is one of assembling a body of diverse information, the various pieces of which may have different ways of being informative about the overarching question. Some *in vivo* testing studies are intended to be surrogate cases of the whole process of potential toxicity generation of the agent, based on our understanding of the commonality in anatomy, physiology, and biochemistry among mammals, but results must be interpreted in light of our larger experience and biological knowledge about the degree of commonality and the potential impacts of species differences that are also known to exist. Especially when animal study outcomes disagree with one another, the question arises whether humans ought to be supposed to

be like responding species or resistant ones, and investigating the potential reasons for such differences, and how they might bear on the relevance to humans, becomes part of the evidence evaluation. In mechanistic studies, indications of metabolic, biochemical, and cellular processes that could be part of a larger mode of toxic action must be interpreted in light of how these parts of a more complex causal network may or may not be indicative of the possibility of generating apical toxicity. It is not just a matter of counting up various lines of evidence, each with its own basis for relevance and questions about sufficiency, as a set of independent indicators; one must also consider whether the various lines are mutually compatible in the sense of describing what are plausibly understood to be different aspects of the same underlying set of causative processes, with assumptions or hypotheses introduced during interpretation of the bearing of some data not contradicting those introduced to make use of other lines of evidence. Deciding whether the various relevant studies and their interpretation together describe a story that is sufficiently internally consistent, plausible, lacking in gaps or untested assumptions, and with apparent inconsistencies reconciled in reasonable and nonarbitrary ways that the body of data as a whole can be considered to suffice for making inferences about the target question. What is at issue is not only the result of each study, but also the reasoning by which it is taken to serve as evidence for the main question at hand. This second aspect brings into the consideration our background knowledge of the applicable biology, along with our experience of how reliable inferences of the type in question have proved in the past, the possible pitfalls to interpretation and how likely they are to come into play -- all aspects that themselves depend on a yet wider body of evidence and its interpretation.

The chief interpretation issue is to identify what is hypothesized to be in common between the test system and the target question, since it is this commonality that serves to make the results of the test system indicative of some aspects of the ability of the target population to be affected. The "hypotheses" in Hypothesis-Based Weight of Evidence consist of just such proposals for why, owing to underlying commonality, each study result should be considered as evidence regarding the potential for toxicity causation in the target population. The method realizes that such applications of study results as indicators are not just extrapolations, they are generalizations, representing the tentative assertion (which then needs to be evaluated) that both the study subjects' responses and the target population subjects' inferred potential for responses

are instances of a more general biological phenomenon. The question then arises, Where else should this generalization apply, and does it indeed to so? In this way, the larger base of data can be used to test assertions of broader relevance of single experimental results, at least to a degree.

The challenge is to try to ascertain not only what the potential generalization of causative elements may be across studies, but also to ascertain the limits to the applicability of the generalization. This must be done in the face of a limited number of observations of potential cases, and it is usually not easy to test generalizations by adding experiments designed to test the hypotheses. For instance, if there is nonconcordance of tumor responses in rats and mice, we recognize that this can happen and indeed has happened for other toxicants. It is not easy, however, to conduct bioassays on an array of further species to test how widely the tumor responsiveness is or is not shared, nor is it easy to repeat bioassays within a species to determine if the discordance might be attributable to a one-study false positive or false negative. There are potential reasons for such discordance (sometimes understood, sometimes not) that bear on how widely (and to what other species, including humans) the effect seen in responding species should be thought to apply. But such a case shows that to use a rodent response as evidence for a possible human response cannot simply rest on basic assumptions about commonality among mammals but also needs to consider (at least hypothetical) reasons for why humans should be supposed to be like the responding rodent species and not the nonresponding one.

As noted, rigorous, unbiased, and nonselective assembly of data is important but won't by itself make the decisions about integration. When weight of evidence is conducted in support of regulatory risk assessment, the challenge of integration is both scientific (making sound judgments, well justified and communicated) and procedural (since the evaluation and judgment is delegated to a process that must be held to standards of valid conduct and case-to-case consistency). Much of the diversity of practice in regulatory weight-of-evidence frameworks (surveyed and commented on by Rhomberg *et al.* 2013) represents attempts to grapple with these challenges. The approaches tend to be those that try to embody sound judgment into rules of evidence that can be followed operationally and those that invoke more generalized case-by-case application of professional judgment, but requiring a process for identifying the judges (and perhaps naming the data and evaluations they should consider but not providing a specific set of

rules or processes for doing so). Both have shortcomings. Approaches based on pre-set interpretation rules work only as well as the rules succeed in capturing the important principles of interpretation, and rules can easily become ossified by practice into conventionalized interpretations that, when applied to particular instances, are seen to miss key aspects of interpretation. On the other extreme, invoking unstructured "professional judgment" depends on the insightfulness and accepted authority of the chosen judges. Though case-by-case special aspects of interpretation can be brought in to the consideration, the judgments can seem arbitrary since by their nature they transcend the specifics of the component studies. In settings where interpretations can be contentious (as in many regulatory applications), this tends to raise questions about the choice of judges and their perceived objectivity.

6 Tenets of Hypothesis-Based Weight of Evidence

Hypothesis-Based Weight of Evidence attempts to find a balance between these extremes in the form of relying on professional scientific judgment, but also asking that the basis for and reasoning behind such judgment be laid out explicitly. The heart of the method is to specify what it means to lay out the basis and reasoning. It is not simply naming the data examined, nor simply noting the results that "support" a certain conclusion or interpretation, nor even these along with supplying reasons why apparently contradictory information does not necessarily invalidate the conclusion. The key is to appreciate that the question is not "how much" evidence there is that suggests (or is in some sense consistent with) any one particular causal interpretation, but rather whether hypothesizing an underlying causal process more successfully provides an explanation for the whole array of relevant results we have on hand than does an alternative set of explanations which does not include the causal process being tested, and instead explains the existing results in other ways (say, as the result of confounding or extraneous causative factors not adequately controlled or even simply chance fluctuations in responses). That is, we have a set of observations and facts before us, and they all have some reason for being as they are, although we may not know all the reasons. Our question is whether we can significantly better explain the set of facts if the hypothesized causal process by the agent under consideration were true than we can if it were not true. The analogy is less to a hypothesis

test in statistics than it is to a likelihood ratio test. In a likelihood ratio test, one has competing models, and for each model one calculates the probability that, if that model were true, it would produce the set of data we have observed. We prefer one model over the other when the observed data are substantially more likely to have been produced if one model were true than if its competitor were true.

In Hypothesis-Based Weight of Evidence applied to assessing toxicological causation, this idea is implemented by first making sure the array of facts and study findings is presented and available for consideration, and then creating alternative "accounts" that comprise a set of tentative and hypothesized reasons for how those results came to be as we see them. One such account includes the hypothesis of a causal role for the agent being evaluated. The nature of this causal process -- including where and how it is being proposed to apply and where it is not -- needs to be put forward with enough specificity to address its hypothesized role in generating outcomes that were seen. A competing account attempts to explain all the results at hand with a set of plausible explanations that omits the causal role of the evaluated agent. That is, the observations are arrayed out in a table, and the hypothesized reasons that these outcomes occurred under each account are specified, with each competing account compelled to specify how it proposes to explain each notable observation. It is important to note that this does not simply list the items of evidence supporting each account; each account also has to deal with its failures or contradictions, and it must therefore include subsidiary arguments for why those inconsistencies could have come out as they did while still maintaining the hypothetical truth of the main causal tested propositions. That is, the accounts comprise a whole set of sub-hypothesis, assembled as necessary to explain the patterns as well as the discrepancies. These subsidiary explanations can be specific (there was a disease in the animal colony that was the actual cause of the observed lesion) or general and *ad hoc* (for some unknown reason, rats are susceptible to an effect that also affects humans, but mice are not susceptible, justified because instances of this pattern have been seen in other cases). Obviously, there is a continuum that spans between firmly based and specifically mechanistic explanations for particular outcomes on the one hand and more and more vague or general ones on the other, until the reasons given become more akin to speculations and excuses than real explanations. But the point is to make these hypothesized explanations as compelling as one can in view of what is known about the

agent and its actions and also in view of wider knowledge of biology and experience of the phenomena that have been seen with other toxicants. Arguments that are only speculations or excuses need to be shown, because it is that very status that makes the account in which they appear be seen as less supportable than an alternative account that explains the same facts in a more scientifically compelling way.

In the end, under this approach, the way one "weighs" evidence is to consider and compare how the competing accounts fare in providing explanations for all of the relevant data, not just the parts that are "supporting" of each account's main driving hypothesis, but all of the facts on hand. The evidence for a causal effect of the agent is stronger or weaker to the extent that an account relying on such a hypothesized causal process appears to be superior to its rivals. By stressing the articulation of additions, supplementary hypotheses, accommodations for otherwise inexplicable facts, and so on as part of the accounts, the method emphasizes that if one accepts the main causal propositions of one account, one is also accepting all these ancillary parts of the account as well. Even if one rejects the notion that agent X can cause effect Y (because the appearance of Y after exposure to X is inconsistent or possibly caused by other factors) one still has the occurrences of effect Y that were observed to explain, and if one does so without the effect of X one needs to say how else the observations of Y are to be accounted for (as chance fluctuations, or as species-specific effects for some hypothesized or merely speculated reason, or as due to experimental error or to the action of an extraneous cause not sufficiently controlled, which must then be accepted as not just potentially affecting Y but actually responsible for it). This general approach is very much in accord with the vision of Sir Austin Bradford Hill in his seminal 1965 paper on assessing causality in epidemiology. In that paper, he identified ". . . the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?" (Hill, 1965).

7 Application

We have applied HBWoE to the following cases:

- assessment of the human carcinogenicity of inhaled naphthalene, where the central question is the relevance of modes of action in rodents to humans and resolving apparently discordant roles for cytotoxicity in the noses of rats (which have tumors) and mice (which do not) (Rhomberg *et al.*, 2010);
- assessment of the human leukemogenicity of inhaled formaldehyde, where the need is to resolve the patterns of risk in human studies (some but not all of which suggest an association) with negative results in animals and also the biological implausibility of formaldehyde's ability to attack the target cells for leukemia in view of its lack of systemic distribution (Rhomberg *et al.*, 2011);
- assessment of the human developmental toxicity potential of chlorpyrifos, which entails questions about systemic distribution as well as evaluation of confounding factors in human studies (Prueitt *et al.*, 2011);
- assessment of the human carcinogenicity potential of methanol, which requires resolving discordant bioassay results where there are significant questions about study quality (Bailey *et al.*, 2012).

I will expand on the application to formaldehyde as a postulated cause of leukemias as an example that serves well to illustrate some of the principles. The nature of HBWoE arguments is that they are complex and lengthy, necessitated by the role of laying out all the observations to be accounted for and examining alternative ways to account for them. The reader is referred to the original publication (Rhomberg *et al.*, 2011) for the specifics. What is presented below is not the full analysis but only a recounting of some notable points to serve as an illustration of the notion and utility of the approach of alternative accounts.

Briefly, formaldehyde is a very reactive compound, albeit one formed naturally in all tissues as a product of normal metabolism. Consequently, the body has a number of defenses against damage by formaldehyde's reactivity with local macromolecules. Rats inhaling formaldehyde at

substantial concentrations develop nasal tumors, which appear to be due to marked cytotoxicity in the nasal tissues that first encounter the inhaled material and have their defenses overwhelmed. Formaldehyde does not cause leukemias or other hematopoietic cancers in rodents after inhalation. In some, but by no means all, human studies, there are signs of increased risk of a variety of hematopoietic cancers in occupationally exposed people, though these effects tend to appear in occupations with lower rather than higher typical exposure levels. The specific kinds of hematopoietic cancers (which typically have distinct and differing causes among types) differ somewhat among studies, and types that are affected in one study appear unaffected in some others. Importantly, in the study with the most apparent increased risk, the relationship is not with average or long-term cumulative exposure, but rather with having experienced particularly elevated peak exposures (though there is no association with the estimated numbers of such peaks experienced over a working history). There are several other questions that bear on the interpretation of human studies, such as possible confounders, indications of especially low risks in unexposed (rather than especially high ones in exposed workers) compared to population rates, and others.

If one examines the epidemiological data alone, one could argue that, even though there is some inconsistency, there are nonetheless several observations of associations of apparently elevated hematopoietic cancer risk with occupational exposure. It is true that the data had to be analyzed against several alternative measures of exposure before the association with peak levels was found, and this practice of multiple parallel analyses of the same data risks seizing on a chance association that is retained simply because it is the one among several that gives a positive result, but examples of dependence of toxicity on peak exposures are known for other chemicals, and often have a mechanistic explanation in overwhelming of the body's defenses, so it could be that the association of cancers with peaks is actually a discovery of a biologically important phenomenon. Formaldehyde adducts are detectable in the proteins and DNA of humans who have been exposed, and adducted DNA is a means for potentially causing somatic mutations that could transform cells to malignancy. All in all, one could consider the human studies by themselves to provide limited but notable evidence for an effect on leukemias and perhaps related hematopoietic cancers. The fact that the rodent studies showed no such cancers could be read simply as the failure of animal data to increase the concern for human risks, since species

differences in responsiveness are known for other agents, and humans are after all the species of concern, so apparent effects in humans might reasonably take precedence.

This line of interpretation is complicated, however, by consideration of the fate of inhaled formaldehyde. Owing to its reactivity, virtually all the exogenous formaldehyde that is inhaled reacts in the immediate nasal and upper respiratory tract tissues it first encounters. It has been shown that the levels of tissue formaldehyde, and the adducts formed, are not increased in tissues remote from the immediate respiratory tract lining by inhalation to even substantial amounts of formaldehyde. Labeled formaldehyde studies show that the adducts detectable in non-nasal tissues have their origin in the naturally metabolically generated formaldehyde and adducts from labeled inhaled formaldehyde are not detectable. Moreover, the cells at risk of malignant transformation for the cancers involved reside in bone marrow, so it is not clear how formaldehyde in its reactive form can get to the targets it would need to affect. To reconcile these observations, it has been proposed that the stem cells that are targets are able to migrate from the marrow into the blood, where they could be exposed as the blood passes through the lungs, and then the cells could return to the marrow. There is evidence of such migration possibilities in rats, making the possibility somewhat plausible, but then again rats do not get leukemias from formaldehyde inhalation, so one would have to argue that the migration phenomenon seen in rats also occurs in humans but for some reason leads to risk in humans but not in rats. Moreover, the migration would be a continuous process, and the kind of impact on the cells as they pass through the lung -- the formation of adducts from reactive formaldehyde -- does not have an obvious way to depend on peak exposure but logically ought to be more relatable to ongoing average exposure, the patterns that did not show association with leukemias in the human studies. In addition, the labeled formaldehyde studies in rats did not show adducts from inhaled formaldehyde in bone marrow tissue, where the lung-exposed cells would have to migrate back.

In short, if one holds that the association of elevated leukemia risk in some human studies is real, one has to accept that the association with peaks is a discovery about mode of action, and the lack of association of effect with ongoing average levels is not refuting. One also has to assert that the lack of effect in animals stems from some yet unidentified species difference that constitutes an exception to the usual argument that rodents can serve as surrogates for humans in

cancer causation studies. Moreover, the apparent implausibility of formaldehyde getting to its needed targets needs to be seen as somehow in error, with some version of the cell migration theory invoked (along with reasons for why it fails to apply to rats) as well as an explanation for how this mechanism nonetheless is consistent with the observation of dependence of effect on peak exposure. On the other hand, if one denies that the positive results in human studies reflect actual causation by formaldehyde, one nonetheless has to explain how it is that some studies show apparent (and under this view, false) associations and that there is some degree of agreement among at least some studies that such associations exist. That is, the phenomena and results are all still there, and rejecting one explanation of them entails accepting some alternative. This example illustrates the importance of tracing the consequences of hypothesized causal processes through all of the data, not just the ones the processes were originally invoked to address. The apparent association of risk with peak exposure is plausible in itself, since such patterns are known, and the migration of stem cells could be taken as a plausible postulate to get the target cells to the chemical being supposed to affect them, but the two assumptions, each reasonable on its own, do not fit well together, and this observation is part of weighing the evidence.

8 Conclusions

HBWoE comes down to evaluation of alternative "accounts." An account is a proposed set of explanations for the set of observed phenomena across the body of relevant observations. The explanations need not be proven—what is important is that one set out what is being proposed as causal and generalizable phenomena, what is the proposed basis for deviations that lead to observations that do not fit (*i.e.*, that would otherwise be counterexamples or refutations), what assumptions are made that are *ad hoc* (to explain particulars, but for which the evidence consists of their plausibility and the observations they are adduced to explain), what further assumptions have to be made (and how reasonable they are), and what is relegated to error, happenstance, or other causes not relevant to the question at hand. There are competing accounts, and one should evaluate the main ones as to how the evidence supports them; what is necessary to assume; and overall, how the weight of evidence for each suggests how compelling the account may be.

The hope is that by explicitly articulating the arguments for how the data should be interpreted as evidence on the overarching question at hand, and by explicitly showing how the conclusions are asserted to follow from the data, one has in essence "opened up" the process of expert judgment, allowing the benefits of case-specific reasoning and well thought-out analysis that can be flexible from case to case, and making the basis for judgments the subject of reasoned scientific debate, which can then focus on the soundness of the arguments rather than on disputes about the choice of the judges or the basis for their opinions.

References

Bailey, LA; Prueitt, RL; Rhomberg, LR. 2012. "Hypothesis-based weight-of-evidence evaluation of methanol as a human carcinogen." *Regul. Toxicol. Pharmacol*. 62:278-291.

NRC (National Research Council). 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Committee to Review EPA's Draft IRIS Assessment of Formaldehyde. National Academies Press, Washington, DC.

Prueitt, RL; Goodman, JE; Bailey, LA; Rhomberg, LR. 2011. "Hypothesis-based weight of evidence evaluation of the neurodevelopmental effects of chlorpyrifos." *Crit. Rev. Toxicol*. 42(10):822-903.

Rhomberg, LR; Bailey, LA; Goodman, JE. 2010. "Hypothesis-based weight of evidence: A tool for evaluating and communicating uncertainties and inconsistencies in the large body of evidence in proposing a carcinogenic mode of action—naphthalene as an example." *Crit. Rev. Toxicol*. 40(8):671-696.

Rhomberg, LR; Bailey, LA; Goodman, JE; Hamade, AK; Mayfield, DB. 2011. "Is exposure to formaldehyde in air causally associated with leukemia? – A hypothesis-based weight-of-evidence analysis." *Crit. Rev. Toxicol*. 41(7):555-621.

Rhomberg, LR; Goodman, JE; Bailey, LA; Prueitt, RL; Neck, NB; Bevan, C; Honeycutt, M; Kaminski, NE; Paoli, G; Pottenger, LH; Scherer, RW; Wise, KC; Becker, RA. 2013. "A Survey of Frameworks for Best Practices in Weight-of-Evidence Analyses." *Crit. Rev. Toxicol*. 43(9):753-784.

Weed, DL. 2005. "Weight of evidence: a review of concepts and methods." *Risk Analysis* 25:1545-1557.

Hill, AB. 1965. "The environment and disease: Association or causation?" *Proc. R. Soc. Med*. 58:295-300.