HYPOTHESIS-BASED WEIGHT OF EVIDENCE: AN APPROACH TO HAZARD IDENTIFICATION AND TO UNCERTAINTY ANALYSIS FOR QUANTITATIVE RISK ASSESSMENT

Submitted to the

National Research Council

Committee on Improving Risk Analysis Approaches Used by the US EPA

On behalf of the

Aerospace Industries Association
1000 Wilson Boulevard, Suite 1700
Arlington, VA 22209

By

Lorenz R. Rhomberg, PhD

Gradient Corporation
20 University Road
Cambridge, MA 02138

July 2007

I. INTRODUCTION

Human health risk assessment consists of bringing to bear a diverse body of *in vitro* and animal testing results, epidemiologic studies, and background toxicological knowledge on the question of whether occupational or environmental exposures to a substance should be regarded as capable of impairing the health of an exposed population. The challenge is that the body of scientific information is rarely definitive — information is typically indirect (animal studies are applied to humans, high-dose studies are extrapolated to low doses), it is often incomplete (few strains/species are tested, epidemiologic studies may lack good exposure information), and it frequently contains apparent contradictions (endpoints may be discordant among studies, genotoxicity tests may include positive and negative outcomes).

As a result, conclusions about potential human risks from chemical exposure are not firm and certain deductions; rather they are statements about possibilities, the evidentiary basis for which may vary from very strong to tenuous depending on the case. The task, therefore, is not only to make reasoned inferences about the potential for human impact, but just as importantly, to gauge and effectively to communicate how *compelling* those uncertain inferences should be deemed to be, acknowledging and giving proper consideration to the existence of contrary data and alternative scientifically plausible interpretations. There are qualitative questions (whether the agent should be regarded as capable of causing particular effects in exposed humans—i.e., uncertainty in hazard identification) and quantitative questions (how accurate and precise the projected dose-dependent magnitude or probability of effect should be understood to be).

This paper memorializes points raised in Dr. Lorenz Rhomberg's oral presentation on behalf of the Aerospace Industries Association of America (AIA) to the National Research Council's *Committee on Improving Risk Analysis Approaches Used by the US EPA* on June 12, 2007. It begins with a critique of current approaches to hazard identification, especially as applied under the 2005 EPA *Guidelines for Carcinogen Risk Assessment* (EPA, 2005), and it proposes a modified approach which can be termed Hypothesis-Based Weight of Evidence.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

This approach stresses that human risk projections are in fact hypothesized scientific *generalizations* regarding particular observed toxicological phenomena (in human or animal studies); they are tentative assertions that some proposed underlying biological commonality justifies believing that the responses seen in a key study are also to be expected among humans in the population whose risk is being assessed. The weight accorded each hypothesized basis for inferring potential human risk ought to be evaluated endpoint-by-endpoint through testing its predictions against all the pertinent data.

Next, these comments apply this thinking to the problems of characterizing uncertainty in quantitative risk analysis (*i.e.*, dose-response analysis), arguing that efforts to find acceptable methods for uncertainty analysis of estimated potencies have tended to falter on just the sort of qualitative questions that the endpoint-specific hypothesis-based analysis can help to address, especially those having to do with mode of action and human relevance as they affect choices among alternative datasets and dose-response models for use in potency estimation. A tighter, more endpoint-specific linkage between hazard identification analysis and quantitative dose-response analysis is advocated as a way of improving the uncertainty characterization of each.

Finally, the comments address the perennial question: How much case-specific evidence is necessary to move beyond a default assumption? It is argued that default assumptions in risk assessment can also be addressed within the hypothesis-based weight-of-evidence framework; defaults can be regarded as generalized hypotheses (based on principles or on our collected larger experience with toxicity of other chemicals) about how an observed toxic effect may relate to human risk. As such, judging whether to replace defaults is seen as an exercise in evaluating whether a case-specific hypothesis that differs from the default constitutes a sufficiently more compelling explanation of the array of chemical-specific observations than does the default hypothesis.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

II. HYPOTHESES AND WEIGHT OF EVIDENCE

It has been pointed out that the phrase "weight of evidence," although much invoked in risk analysis, is used in very different senses by different people (Weed, 2005). In its metaphorical sense it may be used to indicate that findings depend on many sources of data, or that "all" data have been considered, or that findings have been made despite less-than-definitive information, or that findings have been made despite clear conflicts among the available data sources. The intent is usually to indicate that conclusions must be made based on objective scientific interpretations that integrate across sources of data and that evaluate how strongly one is justified in drawing conclusions (perhaps provisional conclusions) from less-than-definitive information. Rarely, however, is any specific methodology or program stated by which the evidence is to be weighed.

The comments below stem from a central key realization: that data only become "evidence" to the extent that they are brought to bear on the evaluation of specific hypotheses. It is the strength of the *hypothesis* that is evaluated with respect to the data, not the other way around. What makes data into evidence — and what gives that evidence "weight" — is its ability to discriminate between true and false hypotheses. That ability is not absolute – we do not speak here of proof or disproof – but weighing evidence consists of judging whether the expected manifestations of a hypothesis actually appear in the relevant data and whether apparently contradictory observations ought to shake our faith in the hypothesis at hand.

Accordingly, in judging the extent to which an array of data on a chemical should be interpreted as indicative of potential human risk, it is essential to articulate a hypothesis about the proposed *basis* for such an inference that is specific enough to expose the logic of the inference about human risk to testing against the available data.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

III. HAZARD IDENTIFICATION

A. Critique of Some Current Practice

In essence, the critique of the current EPA process of Hazard Identification is that it is too unstructured. The former 1986 Cancer Guidelines (EPA, 1986) were rightly seen as overly prescriptive, tending to dictate the interpretation of study results through the rules of the process and giving insufficient room to examine human relevance of animal results or impacts of increased understanding of carcinogenic modes of action. The 2005 revision of the Guidelines substituted a "holistic" approach, dispensing with intermediate separate evaluative steps on human and animal studies, and basing the weight-of-evidence classification on professional judgment informed by an overall synthesis across all the pertinent data and all carcinogenic endpoints.

One could say that the 1986 Guidelines attempted to embody the judgment about bearing of studies on human risk in the rules of the Hazard Identification process itself. In practice, for many chemicals being assessed under this scheme, the main question was whether two valid positive animal bioassays exist; if so, the interpretation as indicative of a "probable human carcinogen (B2)" was nearly automatic, reflecting a presumption built into the process that clear existence of animal tumor responses (shown by repeated positive bioassays, even if for different tumor types) presage human carcinogenicity by virtue of the presumed commonality of carcinogenic processes across mammals. The 2005 Guidelines, recognizing that such prescriptive interpretation often led to problematic results, opened up the process and made it essentially a process of case-by-case argument.

In practice, documents tend to review evidence by categories of study (first metabolism and kinetics, then human studies, then animal bioassays, and then in vitro studies on potential modes of action) rather than being sorted by endpoints of potential concern (particular tumor types) or modes of action. Within each category, each study is listed according to its positive outcomes – i.e., the tumor sites or types that appear to be elevated

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

upon dosing – and documentation of lack of effect for particular tumors or any lack of corroboration of such findings from study to study is only implicit. By covering all tumor responses together (and by attending only to positive results), this organization tends to focus thinking on a generalized endpoint-independent "carcinogenicity" rather than on the rigor with which particular effects have been demonstrated.

There are a large number of "factors" named in the Guidelines that should tend to increase or decrease the overall weight, but there is no explicit means for combining such considerations or for deciding how to trade off between factors tending to increase or to decrease the weight of the evidence. The emphasis is on "positive" evidence – positive results for particular tumors in epidemiologic studies or in animal bioassays – with very little role for examination of "negative" evidence – the lack of the same effects in other human studies or animal tests. As noted, evidence is typically blended across specific tumor endpoints, even if there is no particular basis for considering the array of effects as manifestations of a single carcinogenic process. As a result, evidence for an extensively tested compound that has a number of unrepeated and disparate responses across studies is seen as strong (because the endpoints entertained are numerous), even though the same data could be used to argue that the lack of ability to repeat findings across studies shows them as weak bases for projections of human risk, because the sporadic results may constitute false positives.

The weight-of-evidence categories ("Human Carcinogen," "Likely, … *etc.*) are defined in terms of the strengths of conclusions and do not directly address anything about the logic behind reaching those conclusions or about specific contributions of individual studies. As a consequence, it is difficult to have a scientific discussion about the soundness of weight-of-evidence conclusions produced by the agency; if one disagrees with the bottom line, there is no evident way to trace that back to the interpretation of particular study results or to trade-offs between positive and negative factors, and it is difficult to have a practical scientific dialogue about the analysis or bearing of particular data. As a result, scientific disagreements about weight-of-evidence conclusions often devolve into squabbles about perceived biases among the "evidence-weighers" rather than focusing on the weight of the evidence itself.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

In view of these criticisms, what is needed? The 2005 Guidelines were right in abandoning the prescriptive rules of interpretation from the 1986 edition; a case-specific weight-of-evidence argument is necessary, but there must be some structure to the process and some organized process for the evaluation of evidence. The aim should be to articulate the rationale for projecting potential risk to target human populations in a way that shows how conclusions derive from data and their interpretation (and, as necessary, from science-policy defaults). This makes the process transparent and focuses attention on the scientific arguments and data, not on the evaluators.

As has been argued above, evaluation of evidence can only be done in the context of particular hypotheses that articulate the basis being invoked if one is to apply outcomes from animal studies and/or high-dose occupational human studies to the target human population. In order to address particular biological bases for extrapolation to humans, it would seem necessary to evaluate hypotheses about particular tumor endpoints, or at least particular modes of action. It should be possible – indeed it will nearly always be the case – that some endpoints constitute more compelling bases for concerns about potential human risk than do others. Keeping such distinctions clear is essential to informative hazard identification, and (looking ahead) they are important in informing the quantitative dose-response analysis, which in the end needs to be based on particular data and endpoints chosen from among those reviewed in the Hazard Identification step. By setting out the evidence for and against such hypotheses, and by noting what further assumptions must be made in order for them to be compatible with the data and to apply to the target population, the arguments are rendered more transparent, and the confidence to be placed in projections of various kinds of responses in humans – from nearly sure to merely possible – can be ascertained.

B. Hypothesis-Based Weight of Evidence Applied to Hazard Identification

The process herein proposed to address some of the shortcomings of current Hazard Identification can be termed Hypothesis-Based Weight of Evidence. It aims to create a systematic process for data evaluation that records and tabulates important intermediate

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

conclusions. All of the pertinent studies – and pertinent results from those studies – should be examined. Of course, this includes the presence of effects thought potentially indicative of human risk, but it also includes the *absence* of such effects in other studies, in the opposite sex, and in different strains or species. If one is to assess whether a response seen in a particular study can be generalized to apply to humans, it is necessary to note these discordances as potentially informative about how widely the generalization should be supposed to hold. Noting negative outcomes also applies to human studies, i.e., the failure of animal-study tumors to be detected as affected in epidemiologic studies needs to be noted. It may be that the lack of parallel findings in humans can be attributed to lack of statistical power, but this is an analysis of the data that needs to be made, not grounds for excluding such results from the analysis.

The results should be examined endpoint by endpoint across the span of animal, human, and mode-of-action studies (e.g., noting precursor events and incipient pathophysiological changes), and instances of concordance and non-concordance should be noted. Only by such organization can the generality of particular effects be assessed.

The outcome is a separate weight-of-evidence evaluation *for each endpoint*. Some endpoints may have compelling evidence that they can be generalized to humans (e.g., repeated and consistent findings across studies with only readily understood exceptions) while others may be much more marginal in their evident applicability (e.g., a singleton response that is uncorroborated in other animal studies with no indication of effect in epidemiologic studies). If one is positing a mode of action (e.g., genotoxicity of a systemic metabolite) that is expected to engender tumors at multiple sites, then the data pertinent to that mode of action can be examined as a suite, but in such a case all the negative outcomes for any tumor that is part of the argument (and indeed for any potential tumor site that ought to be exposed to the genotoxic metabolite) becomes part of the relevant suite of data to be considered.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

C. Framing the Hypothesis

As has been argued above, in order to gauge how compelling is the case for a projection of human risk, one needs to articulate the hypothesis to be evaluated. The appropriate hypothesis is the answer to the question, What is the *proposed basis* for inferring that a *particular* phenomenon seen in studies of a chemical's effects will also happen in the environmentally (or occupationally) exposed humans? This recognizes that the proposed extrapolation (say, from a rodent study to environmentally exposed humans) is actually based on a *generalization* of the toxicity phenomenon to include both the test species and the target species. As a generalization, it should apply to all cases within its realm of applicability. The issue is to define what manifestations of the hypothesis are expected and not expected under the conditions of the various studies on hand – the predictions of the hypothesis for these conditions can then be tested against all the actual relevant experience with different sexes, strains, species, and exposure levels. The elements of the hypothesis are to propose the commonality of (a) the material basis, (b) the causal processes, and (c) the progression of events that serve as the basis for supposing that responses in the test system indicate potential toxicity in the target human population.

That is, the hypothesized basis consists of whatever argument one is relying on to apply test results to the risk assessment. It may include details of mode-of-action but it need not do so. There is no "minimum" specification of underlying biology required; for instance, it may be that the argument for applicability simply consists of the observation that most mammals share a good deal of anatomy, physiology, and biochemistry, and the control of cell division and differentiation (which are key to malignant transformation) are largely similar across species. In essence, this simple hypothesis is the basis for extrapolation to humans embodied in the rules of the 1986 Guidelines. Such a hypothesis predicts similar responses in all mammalian species, and so the observation of a tumor type in mice but not rats makes an apparent counterexample that must be evaluated. If the generalization of carcinogenic effect from mice applies to mammals generally, why do rats not respond, and what is the basis for supposing that humans will be like mice rather than like rats in this regard?

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

Definitive evidence on such questions is not expected; indeed, the point is to assess the likelihood that the proposed account is true, in view of the actual results of the relevant studies (positive and negative) and of any subsidiary assumptions (e.g., perhaps rats are unusual for mammals in being resistant to the particular actions that cause cancer in mice and that we are presuming would also do so in humans).

D. Testing the Hypothesis Against All the Available Data

The weight of the evidence for a hypothesis is strengthened when (a) the hypothesis tends to explain results in several studies as a consequence of the proposed basis for generalization; (b) relatively few experimental facts appear inconsistent with it; (c) it makes predictions (especially non-obvious "risky" predictions) that are subsequently confirmed experimentally; (d) it is relatively free of after-the-fact subsidiary assumptions that are included to explain particular outcomes that otherwise might not clearly be expected under the hypothesis; and (e) that whatever subsidiary assumptions one does need to make are reasonably likely to be true ("not a stretch") in view of our larger experience and toxicological understanding. The weight of the evidence for a hypothesis is weakened to the extent that there is a failure to repeat responses across those sexes, strains, species, and doses that the hypothesis, if true, suggests should be responding and to the extent that unpredicted but clearly relevant phenomena do occur in those studies. A hypothesis can often be reconciled with apparent refutations by modifying it or by adding subsidiary assumptions, but such "accommodation" entails a weight "penalty" because it makes the truth of the hypothesis contingent on further unproven assumptions and because the additions are ad hoc, introduced to explain apparent inconsistencies. Again, this is not to say that the further explanations could not be true, but it is clear that invoking them after the fact weakens the degree to which one can say the evidence compels belief in the central hypothesis.

Each hypothesis about the relevance of experimental results to human risk has a counterpart that is the inverse – the hypothesis that the observed studies do not indicate risk to humans. It is important to evaluate the inverse hypothesis against the data as well, with the

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

onus to show that there is a lack of human risk despite the positive studies that have been done. The findings of animal tumors now become the inconsistencies in this inverse hypothesis that must be accounted for. This is not a zero-sum game – a poorly supported argument for human risk does not mean that the inverse argument against such risk is strong. A final but very important criterion for judging the weight of the evidence in favor of a hypothesis is whether alternative hypotheses – ones that predict no human risk – might also constitute good explanations of the array of study results at hand. A hypothesis is most compelling when it is not just a good explanation, but the only good explanation at hand.

E. Structure and Procedure for the Hypothesis-Based Weight of Evidence Evaluation

In the end, evaluation of the weight that evidence lends to a hypothesis is a matter of scientific argument and professional judgment; it is not easily codified into rules, and such rules would likely be counterproductive. Nonetheless, it is important to follow a process that ensures a systematic examination of all the pertinent data and that fosters articulation of the rationale for using particular observations to affect the degree of perceived support for the named hypotheses. Details of such a systematic process need to be worked out, but some suggestions are presented below.

First, a systematic process should be carried out to identify and lay out all the pertinent data, including instances of negative outcomes in animal and human studies for endpoints that are to be considered as candidates for human risk projection. Second, a set of hypotheses should be formulated and articulated, each one proposing a provisional basis by which results from the studies at hand could be used to infer the existence of potential human risk. There should be at least one such hypothesis for each candidate endpoint or mode of action. As stated, these should be accompanied by inverse hypotheses, under which the study outcomes are taken not to indicate human risk (overriding by some means the presumption of mammalian similarity outlined earlier). There may be more than one hypothesis for a given endpoint or mode if distinct bases worthy of consideration can be identified.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

Next, for *each* hypothesis, it would be useful to array the main observations in a table such as shown by example in Table 1. It is proposed that each key observation be put under one of the six categories listed. "Predicted" outcomes are those that arise from the hypothesis before the relevant data have been obtained, and when they are obtained, the predicted result occurs. This will rarely be available, since most data collection will occur after the hypothesis has been formulated. The second category, "Supporting," includes observations that unequivocally are expected from and consistent with the truth of the hypothesis. (Such observations have somewhat lesser weight than true predictions because it is possible to formulate a hypothesis so as to accommodate known data, so the consistency provides less of a check on truth.) The third category includes observations that are supporting so long as significant additional assumptions or *ad hoc* adjustments to the hypothesis are made, changes that may be reasonable but make the observation's support of the hypothesis conditional on further facts that are not known at present. The fourth category is for observations that are neither supporting nor are they apparent refutations. It is worthwhile listing them only to the degree that they illuminate reasoning or provide limits on tenable explanations. The fifth category is for observations that are refuting, unless one accepts significant *ad hoc* explanations. Here belong apparent counterexamples to the generality of the hypothesized phenomena (i.e., negative studies) that can only be reconciled with the applicability of the hypothesis to humans by invoking explanations (e.g., rats are inherently and uniquely insensitive to the mechanism in question) that have little support other than the need for them to be true in order for the main hypothesis also to be true. Finally, there is a category for unequivocally refuting observations. If there are any of these, it should be hard to maintain the potential truth of the hypothesis.

For each hypothesis, and thus for each such table, there should be accompanying text that sets out as clearly as possible the rationale for the placement and interpretation of each observation in the table. When subsidiary assumptions are made, the text should address how plausible they are in view of data at hand, known possibilities from other chemicals, and so on. To the extent that untested subsidiary assumptions are risky (i.e., possible but largely

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

unexpected based on general experience and background knowledge) they are clearly invoked to rescue a hypothesis that would otherwise be refuted. Again, the subsidiary assumptions might be true, but the likelihood of the truth of the main hypothesis now becomes affected by the likelihood of truth of the necessary subsidiary assumptions.

The tables and the accompanying text lay out the considerations for testing the likelihood of hypothesized bases for human risk projection against the body of relevant observations. They serve to articulate the ways in which each observation is used and the considerations applied in interpreting its bearing. The trade-offs between factors tending to support and those tending to weaken the weight of the evidence can be noted.

The next step would be to come to an individual endpoint weight-of-evidence conclusion for each effect that is a candidate for a risk applicable to humans. The results can be tabulated as shown by example in Table 2. The categories of endpoint-specific conclusions would need to be defined, and establishing clear and consistently applicable definitions would be a challenge, but the aim should be to create a qualitative scale regarding how compelling a case is made for the existence of human risk based on each hypothesis evaluated. As categories we propose the following: Compelling, Likely, Plausible (with further assumptions), and Unlikely. The first category holds hypothesized human risk sources that, although perhaps not absolutely proven, would be surprising to find false in view of the strength of the arguments on hand. The second likely category is for hypothesized risks that are more likely than not to be true. The data should be largely consistent, with any apparent counterexamples readily explained by plausible additional assumptions. The third category is for endpoints that require accepting significant further assumptions that, although possibly true, have a substantial likelihood of not being the case, in view of our larger base of knowledge. Accepting a Plausible endpoint as a potential human risk entails accepting these additional conditions as being true. The final category, Unlikely, is intended to hold hypotheses that cannot be disproved but that have little to speak for them other than their possibility. To accept them would require accepting a number of further assumptions about why apparently refuting observations should not be taken at face value.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

Completing Table 2 results in an array of endpoint-specific weight-of-evidence conclusions. The advantage of such an array is that it makes explicit that any overall conclusion about "carcinogenicity" not-otherwise-specified is based on a compilation over such individual arguments. It may be of use to establish a further step, in which the array of endpoint-specific conclusions are combined into an overall conclusion about the likelihood that there is some carcinogenic process among those suggested by the array of studies that operates in the target human population. There is no necessary reason to suppose that just one of the tumor types constitutes an indicator of human risk; several could actually operate, although the likelihood of this is a function of the likelihood that each endpoint is relevant.

F. Advantages of Hypothesis-Based Weight of Evidence for Hazard Identification

The process outlined above has the advantages that it shows which endpoints are most compelling, and it forces one to set out the logic and reasoning by which one concludes that an endpoint can be extrapolated to humans. The strengths and weaknesses of such arguments are laid out, and this allows debate about conclusions that can focus on particular data and their interpretation, with the consequences of these for the overall conclusion being evident. It frames the weight-of-evidence classifications as scientific statements in the sense that they are elaborated and justified hypotheses about generalizations that can be tested by identifying potentially observable consequences of those generalizations. The process outlined does not blend all evidence into a difficult-to-interpret overall statement about the compound's "carcinogenicity" divorced from how that carcinogenicity is expected to be expressed. Finally, as will be elaborated upon below, it informs the quantitative dose-response analysis, including the choice of datasets to represent a compound's potential human cancer risks and the models and assumptions to be used in characterizing potency.

IV. CHALLENGES IN QUANTITATIVE DOSE-RESPONSE ANALYSIS AND IN CHARACTERIZING ITS UNCERTAINTY

Dose-response analysis is part of quantitative risk assessment. The goal is to express how the probability or magnitude of risk is expected to vary as a function of the level of

exposure to a compound. The typical approach has been to define an "upper bound" potency that is unlikely to be exceeded but might, depending on circumstances, substantially overstate actual risks. Accordingly, the tendency has been to choose the "most sensitive" dataset (the one implying highest risks), with only secondary regard as to which particular cancer endpoint constitutes the most compelling possible source of human risk. Risks are extrapolated to low doses using an assumption of linearity, unless data suffice to show that a nonlinear process prevails.

For a variety of reasons, this focus on upper bounds and most-sensitive datasets is no longer considered sufficient. Risk management analyses increasingly seek to balance central estimates of risk reduction with central estimates of the costs of control. This requires better understanding of the uncertainty in potency estimation, a sense of the distribution of tenable potency estimates (including upper bounds but also other estimates), and a sense of what particular actual effects are expected to be changed in frequency when a population's exposures change. In short, a much fuller characterization of the uncertainty in potency estimation is sought.

A. Some Current Problems in Dose-Response Analysis

At the beginning of the dose-response analysis step, there are many choices of datasets that could be chosen, representing different species and different endpoints. There are also choices of dose measure, assumptions about cross-species toxicological equivalence (dose-scaling) and alternative mathematical dose-response models. A number of approaches to defining statistical distributions for potency estimates have been examined; discussing these is beyond the scope of the present essay, but it should be noted that these statistical methods focus most naturally on the quantitative uncertainties inherent in analysis of a single dataset by a single dose-response model. They consider factors such as the uncertainty in fitted model parameters and the consequent uncertainties in extrapolating to low doses as well as the uncertainties in dose measures and their equivalence across species. Such methods become much more awkward when they are applied to the more *qualitative* uncertainties: choice

among alternative mathematical formulations of the dose-response relationship, choice of linear or nonlinear approaches, and most especially choice of the dataset (and hence, choice of the sex, species, and tumor endpoint) that is taken to represent the carcinogenicity of the agent in question.

This awkwardness stems from three problems. First, the alternatives are qualitative choices that do not easily lend themselves to descriptions by statistical distributions. One can use expert elicitation to assign probabilities to alternatives (e.g., Evans et al., 1994), but this raises transparency issues and the regulatory process has been uncomfortable with it.

Second, the sources of information for describing and characterizing the uncertainties reside not in the datasets themselves (as they do, say, for dose-response model parameter estimation), but in the wider external context about biological relevance, bearing, and the likelihood that the responses being analyzed can be taken to represent a real risk process in the target human population based on presumption of commonality of underlying causes. Whether to apply a linear or a nonlinear approach to a dataset is informed very little by the tumor-incidence data themselves and mostly by other information on mode of action, and evidence regarding the understanding of dose-dependency in key events, all of which depend on wider scientific understanding gleaned from hypothesis-driven analysis of whole bodies of relevant studies.

Finally, if in the Hazard Identification step the evidence regarding the agent's carcinogenicity has been blended across all potential endpoints and datasets (as has been the practice), how is any single dataset – which necessarily represents a particular endpoint and a particular rationale for its application to humans – able to represent the whole? If for instance, a compound is judged possibly carcinogenic to humans (and hence in need of dose-response analysis) based on mouse liver tumors and rat kidney tumors and possible rat mononuclear cell leukemias, how is this generalized and provisionally accepted carcinogenic ability to be quantitatively characterized by a single dataset on one of those endpoints? Under the "upper bound" mandate, one could imagine an argument (albeit a somewhat tortuous one) that the

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

actual process generating human cancer risk (if there is one) could be as potent as the most sensitive animal dataset would indicate. But if a statistical distribution of estimates, or a central estimate, is desired, it is clear that the analysis has to consider the information from all the candidate datasets simultaneously, somehow weighted by the likelihood that each dataset is indicative of an actual parallel human risk.

The progress of moving beyond upper-bound carcinogenic potency analysis and of improving the characterization of uncertainty in potency has been hampered by difficulty with handling these qualitative questions. Yet such questions cannot be ignored, since they probably constitute much larger sources of uncertainty than the statistical, curve-fitting and parameter estimation uncertainty that is more readily analyzed. What is needed is a means to array the quantitative consequences of the choices of datasets and rationales for their use in human quantitative risk estimation that makes evident the consequences of choices of datasets and models and gives a means to rate the alternatives in terms of there relative plausibility as measures of potential human risk.

B. Hypothesis-Based Weight of Evidence as an Approach to Uncertainty Analysis of Dose-Response Analysis

The hypothesis-based approach to weight of evidence in the Hazard Identification step lends itself to addressing the problems set out above. The very difficulties of characterizing the plausibility and credence we should put in each choice of dataset and dose-response analysis for projecting risk to humans are the factors that have been examined and characterized in the evaluation of hypothesized bases for invoking studies as qualitative sources of potential human risk.

Each endpoint-specific hypothesis, embodying a rationale for the comparability of some data on that endpoint and the risk process in humans, and each one evaluated against all the pertinent data to judge the degree that it should be considered a compelling argument for human risk, corresponds to a quantitative analysis that uses those data and a dose-response model appropriate for that rationale.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

Thus, the hypothesis-based approach provides the structure for and the basis for evaluating qualitatively alternative approaches to the dose-response analysis. Each one can have its statistical uncertainty that emerges from parameter fitting and low-dose extrapolation further characterized, but by "pulling out" the qualitative questions, these further uncertainty analyses are less problematic.

For each endpoint in Table 2, there is one particular quantitative analysis implied (or perhaps a few), and each such analysis has its numerical estimates of possible potency. The overall uncertainty in carcinogenic potency can then be characterized by this array of qualified possibilities. The advantage that this has over a simple range of possibilities is that each possibility comes embedded in a lot of analysis of the likelihood of its applicability and the relative credence to be given alternatives is readily apparent. It may be, for instance, that the highest potency estimate comes from and endpoint and dataset that is judged as only marginally compelling, and a lesser potency alternative may have more credibility in arising from a hypothesis about human risk that is deemed quite likely to apply in fact.

In order for the advantages of this approach to be realized, it would be necessary to maintain a tighter connection between the analyses of the Hazard Identification step and the Dose-Response Analysis step. Indeed, the potency-estimation results should play some role in feeding back on the evaluation of potential hazards, because the observed lack of an expected effect in a study could be attributed to insufficient statistical power to detect the effect, an argument the reasonableness of which depends on potency. So, rather than throw away all the insights into how the evidence supports or does not support particular ideas about risks operating in the target human population, it would seem well for dose-response analysis to use the fact that these considerations have been analyzed and systematically laid out.

V. DEPARTING FROM DEFAULT ASSUMPTIONS

A perennial question in risk assessment is, "How much evidence is needed does to depart from a default assumption?" This can also be seen as a weight-of-evidence question.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

As has been argued throughout, data become evidence only in the context of being used to evaluate some hypothesis, and so one must frame the default-replacement question as one about hypotheses in order to make any headway on this matter.

The analysis that is a candidate for replacing the default – a chemical-specific factor that addresses the issue covered by the default assumption – is based on a hypothesis that the basis for calculating the factor indeed applies. Thus, it can be evaluated for how compelling a case is made by articulating the hypothesized basis and testing that hypothesis against all pertinent data, just as with the hazard identification application.

But the default assumption is also a hypothesis, albeit a more generalized one, and its plausibility also needs to be assessed to see whether the chemical-specific analysis is notably more compelling in view of the evidence at hand. To do such analysis, it will be necessary to do a better job of articulating default assumptions in terms of the particular presumptions that they make about the underlying basis for the factors they address. It would appear valuable that a program of investigating the scientific basis for defaults be undertaken, and that the often nebulous ideas about what distribution of possibilities the default is meant to cover be made more explicit.

Broadly speaking, there are two kinds of bases for defaults. One is to invoke some proposed overarching principle that is presumed to rule the factor in question. An example is the use of interspecies allometry to arrive at the ¾-power of body weight approach to scaling equivalent doses across species; the presumption is that differing proportionality between body mass and the rates of uptake, processing, metabolism, and excretion across essentially similar mammals is the key to understanding toxicologically equivalent doses. One could imagine testing this hypothesized basis for dose equivalence against data that could address whether this is indeed the dominating factor. A second example is the use of low-dose linear extrapolation on the grounds that direct-acting genotoxic carcinogens should increase the rate of key somatic mutations in proportion to internal concentration owing to the law of mass action.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

The second kind of default is based on the distribution of cases among many previously experienced. Uncertainty factors in non-cancer risk assessment are examples. The values chosen are justified as big enough that the extrapolation using them is likely to constitute an upper bound, in view of our general experience in how much the factors covered (cross-species equivalence, interhuman variability, chronic-to-subchronic extrapolation, etc.) actually vary in our past experience. Doing a better job of articulating the bases for defaults and documenting the actual extent of variation would help in evaluating the defaults as competing hypotheses for bases of extrapolation, and such work would be useful for defining uncertainty distributions for more quantitative approaches.

VI. CONCLUSION

The theme throughout these comments is that, in order to weigh evidence, one needs to make specific hypotheses about proposed bases for generalizing toxicological phenomena, and it is these generalizations that one uses to apply indirect data (on animals, in vitro, in high doses or special human populations) to the evaluation of whether and how much risk may occur in the target human population. The approach should be to test these hypothesized bases against all the data that can illuminate how compelling is the scientific case being made.

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

Table 1

Kidney tumors secondary to high-dose cytotoxicity due to reactive thiol

PREDICTED	SUPPORTING	SUPPORTING WITH SIGNIFICANT *AD HOC* ASSUMPTIONS	NEITHER SUPPORTING NOR REFUTING	REFUTING UNLESS ACCEPT SIGNIFICANT AD HOC EXPLANATIONS	REFUTING
	Bioassay Rats w kidney tumors have kidney cytotoxicity	Some Strains of Rat have fewer tumors and females have fewer (*sensitivity differences? metabolic differences?*) Human studies w kidney tumors only at very high exposures (*high enough for cytotoxicity?*) Humans seem to produce reactive thiol (*enough?*)		Bioassay Mice have kidney toxicity but no kidney tumors (*less sensitive to cytotoxicity? why?*) Most human studies are w/o kidney tumors, but have lower exposures (l*ow enough to avoid cytotoxicity?*)	

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

Table 2

Reasoned Ranking:

▪ Compelling	
▪ Likely	
▪ Plausible	Liver, Kidney
▪ Unlikely	MCL, Esophageal

Aerospace Industries Association of America, Inc.
1000 Wilson Boulevard, Suite 1700 Arlington, VA 22209-3901 (703) 358-1000 www.aia-aerospace.org

REFERENCES

EPA. 1986. Guidelines for Carcinogen Risk Assessment EPA/630/R-00/004, Sep 1986.

EPA. 2005. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001F, Mar 2005.

Evans JS, Gray GM, Sielken R, Smith AE, and Graham JD. 1994. Use of probabilistic expert judgment in uncertainty analysis of carcinogenic potency. Regulatory Toxicology and Pharmacology. 1994; 20:15-36.

Weed DL. 2005. Weight of evidence: a review of concept and methods. Risk Analysis 25:1545-1557.