June 11, 2013

The National Institute of Environmental Health Sciences
Office of Health Assessment and Translation

Re: Public Comment on the "Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments"

To the National Institute of Environmental Health Sciences:

In this letter, we would like to offer public comment and observations on the "Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments," in accordance with the Request for Public Comment as posted at http://ntp.niehs.nih.gov/?objectid=960B6F03-A712-90CB-8856221E90EDA46E. For the record, we have addressed issues on systematic review, weight-of-evidence, and integration of evidence for a number of private-sector clients; our comments here, however, are our own, written without sponsorship or support of any other party.

The NRC Committee that reviewed US EPA's formaldehyde risk assessment made important criticisms about the risk assessment process in general. This included critiques of the current process' transparency and its justification for choices made in data sets and analytical approaches it chooses to rely upon. The formaldehyde review Committee set out a so-called "roadmap" for reform that called for much greater attention to creating and following a systematic process for identifying endpoints of concern and the studies that are to be relied on to characterize them. The roadmap called on official risk assessments to be much more explicit about the reasons and justifications for choices of data and analysis, and to defend its choices – including weight-of-evidence judgments – with explicitly scientific arguments.

The NTP is to be commended for taking up the challenge of forging a new and rigorous approach to systematic review of relevant studies and integration of evidence into judgments about the scientific status of hypothesized toxicity processes that might apply to human populations. We are very much in favor of systematic approaches to identifying studies, setting inclusion and exclusion criteria, and applying consistent and pre-defined means to evaluate study strengths and weaknesses irrespective of study outcome. We are concerned, however, that a system that focuses only on this aspect – and overlooks the considerable challenges of integrating the data abstracted by these processes as they are brought to bear on the human risk assessment question – risks oversimplifying the data integration and weight-of-evidence process.

The models for systematic review that are being used (such as the process for the Cochrane Collaboration) are based on settings where the studies being evaluated are *direct observations* of the question at hand (the efficiency of a treatment in clinical trials). In such settings, the main "integration" questions are about (1) internal validity of individual studies and (2) their collective consistency in directly demonstrating the effect of concern. The rationale for the approach is that, if an effect is true, it ought to be reflected consistently in all the methodologically valid direct observations of it.

In toxicology, the issue is complicated by the diversity of studies being brought together, most or all of which constitute at best *indirect observations* about other systems (in vitro, in animals, and in human populations different from those for which the assessment is targeted). The bearing of such indirect studies on the question at hand (about causal connections of exposure to an agent with human disease risk) depends on a good deal of interpretation and invocation of further assumptions and wider biological principles. The bearing, relevance, and appropriate interpretation of study results is therefore much more than simply a question of whether individual studies are conducted well or have good statistical power. A study could "score" on standardized evaluations as well-conducted by the standards for studies of its type, but the inherent limitations of studies of that type for inferring human risk potential can remain. For example, even the best epidemiological studies have the challenge of isolating causal effects from among the many possible influences of diverse and varying study populations, while even the best animal bioassays have the limit that their relevance to the possibility of parallel responses in humans needs to be inferred based on cross-species, high-to-low-dose extrapolation, and extrapolation from constant to variable exposure patterns. The statistical power to detect uncommon effects could be high in a bioassay compared to other bioassays (making it a "good, high quality" bioassay) and yet still quite low compared to the risk levels that would be of concern in the target human population. When studies disagree, it could be because of faulty or misleading results from some of them or because the effect being observed is genuinely contingent on some biological aspect that differs from one experimental system to the next.

For these reasons, it would be in error to suppose that simply conducting a systematic search for and presentation of data would by itself lead to clear interpretations and conclusions. The OHAT Approach wisely notes that the systematic review of data is not a substitute for scientific judgment, which will still be necessary. The Approach lays out some suggestions for the integration process, noting that it is a work in progress and that further thought and development will be needed. We agree that further development is needed; the current "integration" approach as suggested in the Draft Approach is largely an approximate restatement of old approaches as employed for many years by IARC and (under the 1986 Guidelines but not as much in the 2005 Guidelines) by the US EPA. A fresh approach is needed, and the advent of new kinds of data (especially mechanistic data and high-throughput *in vitro* testing) will need new approaches to continue to be relevant.

When we apply study results as evidence about potential human risk in the target population, we are in effect proposing a *generalization* of the causative processes seen in the source study that should also apply in some way to the target human population -- we are proposing that the causes responsible for the study's outcome could also apply in some relevant way to possible causation of adverse effects in humans because they constitute general properties of the interaction of the agent with biological living systems. As such, this generalization ought to apply throughout the body of studies wherever it is in principle observable. This makes other observations in different systems (other animal studies, or human compared to animal studies, or mechanistic studies that should illuminate the action of the hypothesized common causal processes) useful as evidence about the existence and properties of the asserted causal basis that makes each individual study constitute evidence. Of course, all causal processes are not universal, and there may be reasons why an effect seen in one study is not manifested in a different species or when a different experimental design is employed. But an important (and often overlooked) aspect of weight of evidence is to recognize when the explanations for such inconsistencies is largely *ad hoc*, that is, based solely on the existence of the difference that the explanation is introduced to explain without any

independent evidence for it. Such accommodation of particulars that would otherwise be seen as evidence against the generality of the causal processes being evaluated does not mean that the *ad hoc* explanations are false, but it does mean that the overall weight-of-evidence is lessened because the data at hand do not differentiate as well between explanations that invoke the hypothesized causal process and those that explain the array of results otherwise. In short, data support a hypothesis not just by being consistent with it, but by being inconsistent with rival explanations.

For these reasons, simple weight-of-evidence or evidence-integration schemes that count up "supportive" studies or have rules by which studies with "better" quality trump contradictory studies will miss the important aspect of working through the logic of why the particular results should be considered to constitute evidence. The issues of general applicability of causal processes and their consistency in operation (and the validity of reasons for any inconsistencies in a way that leaves the inference about the target population intact) need to be evaluated.

In particular, the approach outlined in the OHAT evidence integration process, which comes to an overall conclusion about how "animal studies" collectively indicate a potential human risk (combining all the animal evidence, despite its diversity, into one conclusion on their joint bearing), is suspect. It is important to work through what mechanistic studies and animal results have to say about the understanding of a scientific plausibility of patterns seen in human studies. That is, the different streams of data should inform the interpretations and understanding of the phenomena seen in other streams, and one should not just combine "conclusions" drawn from each stream independently.

One must grant that it is a challenge to find a way to work through the foregoing kind of analysis of weight of evidence, coming up with an approach that is not formulaic or merely procedural, one that gives full attention to the complexities of the underlying scientific arguments and how the available data bear on them, and yet one that does not simply invoke "professional judgment" in a way that produces untransparent declarations of conclusions that cannot be held to a standard of rigor or consistency with other cases. But that is the challenge that risk assessment now faces; it is the challenge posed by the NAS formaldehyde committee's "roadmap"; and it is the challenge that faces OHAT as it works on how to define and implement the process of evidence integration. We recognize that the draft process acknowledges that further work on this aspect is needed, and we urge that such work proceed, with care not to finalize a process too quickly before a sound and well conceived process is articulated.

Even though our comments have focused on the evidence integration process, it is important to note that the diversity of studies that need to be brought together also affects the earlier part of the process, where studies are included or excluded, their data abstracted, and their designs and quality evaluated. When the studies to be considered are all of highly standardized design – and when only studies of such design are to be included – it is an approachable task to name the inclusion/exclusion criteria beforehand, as well as to make standardized data-extraction protocols and to measure and report study quality along predefined and objective evaluation criteria. But full toxicological evaluations need to consider many kinds of studies, often ones that do not follow standardized designs (but must nonetheless be held to standards of rigor and ability to avoid extraneous interference with the potential causal processes being investigated). The diversity of study types can be considerable, and this will only increase as new-technology testing methods come into prominence. Importantly, useful and informative data often come from observations

that were not the primary purpose of the studies that contain them, and particularly if these observations show no effects on the ancillary endpoints, they may not be noted in publication titles, keywords, or abstracts. Thus, a search process that relies only on these may miss important data, and may systematically overlook null findings that should bear on the evaluations being conducted. Moreover, a diversity of data and study designs complicates the specification of data inclusion/exclusion criteria and standardize data-extraction processes. As the later evidence-integration step proceeds, it may become evident that aspects of the available studies that were not deemed primary are after all relevant to the evaluation of toxicological hypotheses. For all these reasons, a process that is flexible enough to recognize that things that were first deemed irrelevant might come to be seen as relevant, or aspects of data not seen as critical may become important, such that data extraction and recording standards might need to be altered. It is notable that the Draft OHAT Approach recognizes this and has provisions for revisions to *a priori* protocols as becomes warranted (as long as the changes are justified and recorded). We only note that, based on experience, it may well be frequent that such mid-course modifications are needed, and it would be a mistake to adhere too rigidly to initial criteria in the name of avoiding biases when in fact biases could be created by too rigidly keeping to a process that has more conventional thinking imbedded than was at first realized.

We look forward to a process of open scientific discussion as OHAT develops and implements its Draft Approach. A successful process for this will be a great service to toxicology and risk assessment generally.

Sincerely,

GRADIENT

[Redacted]

Lorenz R. Rhomberg, Ph.D., FATS
Principal

[Redacted]

Julie E. Goodman, Ph.D., DABT
Principal