

A historical perspective on the classification developed and used for chemical carcinogens by the National Toxicology Program during 1983–1992

by James Huff, PhD¹

HUFF JE. A historical perspective on the classification developed and used for chemical carcinogens by the National Toxicology Program during 1983–1992. *Scand J Work Environ Health* 1992;18 Suppl 1:74–82. To evaluate, interpret, and better communicate the data and findings from long-term chemical carcinogenesis studies on laboratory animals, the National Toxicology Program began using five categories or levels of evidence of carcinogenicity in 1983 (clear, some, equivocal, and no evidence and inadequate experiment). Through July 1991 these defined terms had been used to describe 144 chemical carcinogenesis studies comprising 530 sex-species experiments. Together with a selected descriptor of the chemically associated level of evidence for each experimental unit (male rats, female rats, male mice, female mice), mention is made of the length of the experiment, route of exposure, and the particular tumor type or types for each organ or system affected. The scientific judgements are comprised of this relevant information to inform the readers and users of these evaluations. In this paper the background rationale for the development and proper use of these categories of evidence of carcinogenicity are detailed, together with some personal reflections.

Key terms: categories of chemical carcinogens, evaluating results from long-term chemical carcinogenicity experiments, key factors in data interpretation, levels of evidence, public health, weight of the evidence.

Chemical carcinogenesis experiments on laboratory animals are conducted primarily to identify those chemicals, mixtures of chemicals, or environmental and occupational exposures that may potentially induce cancer in humans (1–4). Complementary papers have been also published on the design, conduct, evaluation strategy, and relevance to public health of long-term chemical carcinogenesis studies (5), as well as on the merit and value of using experimental findings in the risk assessment process (6).

In this paper I present some of the background that led to the development of levels of evidence of carcinogenicity and their adoption by the United States National Toxicology Program (NTP) (7–10). Data on 144 long-term chemical carcinogenesis studies (530 experiments) are used to illustrate the scientific concepts and procedures employed to evaluate and select particular levels of evidence of carcinogenicity. The evaluations cover the period from June 1983 through the July 1991 Scientific Peer Review meeting.

Other agencies (4, 11–15), states (16), consortiums (17), groups (18), and individuals (19, 20) have formulated and proposed categories of evidence for characterizing and grouping chemicals that have been studied in various carcinogenesis experiments. These are not discussed per se in this paper.

Chronology

Since 1978, when the NTP was established, until early in 1983, the results of long-term chemical carcinogenesis experiments were declared simply to be either positive or negative, that is, the chemical was carcinogenic or not carcinogenic. Beginning in June 1983, five levels of evidence of carcinogenicity were defined, established, and used to describe qualitatively the strength of the experimental evidence of carcinogenicity. This change was made because the staff of the NTP were convinced that merely labeling a chemical as either carcinogenic or not carcinogenic was not in the best interest of science or public health and in particular was not sufficiently helpful to administrators and regulatory agencies in the United States and other countries (7, 9, 10).

A major difficulty with a “black and white” appraisal system is that it does not allow for any “shades of gray” or scientific nuances that would allow a more realistic awareness of what the data and results mean. Granted, implementing a system of five categories may not seem to some as being a major change; yet the NTP staff believed that five levels of evidence permit more flexibility and better differentiation among experiments and chemicals. At the time the proposal was made, some thought that the levels of evidence approach would confuse rather than clarify the situation. Others opined that the different categories would imply quantitation. Nonetheless, the overwhelming majority of counselors and other organizations endorsed these efforts, with the attendant concepts and philosophy, and continue to do so.

¹ National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, United States.

Reprint requests to: Dr JE Huff, NIEHS, Box 12233, Research Triangle Park, NC 27709, USA.

The data evaluated in the NTP come from its own experiments. That is, the staff designs the protocols; oversees the conduct of the studies; validates the experimental results; evaluates the data; presents the findings and conclusions in public meetings before non-governmental panels of experts; prints the collections of "raw" data together with detailed literature reviews, evaluations, interpretations, and conclusions in a series of technical reports; and publishes the essence in scientific journals. Importantly these collective data allow scientific concepts to be formulated that undergird the development of research strategies and initiatives (3, 5).

The initial levels of evidence were first described and used at a public peer review meeting held at the National Institute of Environmental Health Sciences in June 1983. Five categories of interpretative conclusions were adopted for the Program's technical report series, largely to be able to maintain consistency with time and to portray better the actual "strength" of the experimental evidence (while not defining potency or implying mechanism).

For each experimental grouping (ie., male rats, female rats, male mice, female mice), the NTP decides on a single level of evidence of carcinogenicity selected from the following five categories: two levels indicating chemically induced carcinogenic responses (clear evidence and some evidence), one level for uncertain yet not negative findings (equivocal evidence), one for no observed response (no evidence), and one for seriously flawed experiments (inadequate experiment). These qualitative descriptors, assigned to a particular sex-species set of data, represent experienced judgements rather than placement into preestablished level-specific algorithms with strictly defined requirements. Moreover, from the perspective of the NTP and for public health purposes, these five levels seem to fit all of the situations encountered as the findings and results from the experiments are evaluated. Finally it was concluded that to divide the categories of "positive" experiments further would be more analytical or quantitative than the experiments or data would permit and might make later independent evaluations by others more difficult. However, the issue of defining and using a greater number of categories of evidence to differentiate further among the various carcinogens has gained support (17-19). Yet, these possible additions seem to accent more rote or fine tuning oriented toward regulatory audiences and may even thwart the very flexibility being sought. In addition more emphasis has been proposed for mechanistic considerations (21).

A major distinction of the NTP from that of the International Agency for Research on Cancer (IARC) (22, 23) stems from the decision not to make overall or combinational evaluations for chemicals; that is, the NTP evaluates and reports separately levels of evidence for each experimental grouping. Of course one can compare and decide qualitatively the overall strength of the evidence among studies by noting the levels of evidence and carcinogenic responses for the

individual experiments. Using extremes as examples, one could consider a chemical that causes multiple cancers in multiple organs at all exposure levels in both sexes of two or more species to be obviously more carcinogenically hazardous to humans than another chemical that induces an increase in a common cancer in a single organ in a single sex of one species.

In March 1986 the levels of evidence were reexamined by another ad hoc group of experts, again in public session, and changes were proposed and made to the definitions. No recommendations were received or voiced to alter the number of categories. In addition a change was made in the terminology from "evidence of carcinogenicity" to "evidence of carcinogenic activity" (discussed later). At that time the NTP also agreed to use the revised categories for a period of time, and, like IARC (12), to reevaluate them periodically, especially when new scientific information became available (eg, mechanism of action, cell proliferation, metabolism, and pharmacokinetics). As currently used, the levels and definitions are given separately, followed by a personal reflective opinion for each category of evidence.

NTP levels of evidence of carcinogenicity

Category 1

For "clear evidence" the NTP states in each NTP technical report on the toxicology and carcinogenesis of a substance: "Clear evidence of carcinogenic activity is demonstrated by studies that are interpreted as showing a dose-related (i) increase of malignant neoplasms, (ii) increase in a combination of malignant and benign neoplasms [of the same type in the same organ], or (iii) marked increase of benign neoplasms if there is an indication from this or other studies of the ability of such tumors to progress to malignancy."

In my opinion the concept and definition of "clear evidence" seems to be well thought-out, and the meaning reflects a chemical that causes cancer and should not be ignored. Conversely, the term "carcinogenic activity" is not defined and is confusing in regard to meaning and in comparison to the more universally used and well-understood term "carcinogenicity." The term "carcinogenic activity" was first used as a replacement for "carcinogenicity" after the March 1986 review of the levels of evidence. (This issue is discussed later in the general discussion.) Furthermore, the terms "dose-related" (perhaps logically taken to mean the more commonly used "dose-response relationship") can mean different things to different people or to groups with different backgrounds. As examples, suppose a statistical dose-response trend test shows significance for site-specific tumor response data. That is, 2-3% of the animals in the control group has this tumor, whereas 3-5% in the low-exposure group, 4-6% in the middle-exposure group,

and 15–20% in the top-exposure group have this tumor. One might ask if these experimental findings should be considered “dose-related.” The trend statistic would imply that these data would be “dose-related”; yet biologically this would not likely be the case for this set of data. Thus the argument could be made that there was no dose-response relationship, and this chemical might not qualify for the clear evidence category.

Suppose there are no statistically significant indications for the data set: 1% in controls, 1% in low-dose group, 3–4% in middle-dose group, and 3–4% in top-dose group. Should this be considered evidence of carcinogenicity? If yes, do these data represent a “dose-related” effect? Depending on the tumor type (rarely occurring versus a commonly found tumor), the occurrence of the same tumor in the other sex-species experiments, and a number of other factors (discussed later), the numerical data alone cannot be used for assigning a level of evidence in the absence of the total picture.

Therefore one of the difficulties in using this definition relates to the full meaning of “dose-related increase”; perhaps this term could be better stated. For example, hydroquinone caused a “dose-related increase” in rarely found kidney tumors in male rats (0% in controls versus 7% and 15% in two exposure groups), but, because none was malignant, the level selected was “some evidence,” even though these tumors represent, for the most part, a stage in the progression to malignancy. Conversely hexachloroethane caused this singular tumor type in two exposed groups [2% of controls versus 4% and 14% (6% were malignant) of the two exposed groups] and was judged “clear evidence”; in this case supportive evidence existed for tubular cell hyperplasia. Consistency is an additionally important issue in these evaluations. Another example is the kidney tumor response for nitrofurantoin: 6% in control male rats versus 22% and 38% in the two exposure groups. (Several extra sections of the kidney were evaluated in this study.) This finding was judged “some evidence” on the basis of the single-section original data, despite the convincing responses and dose-related increases.

Another generic issue with this category comes from a lack of discrimination between “super carcinogens” (6) and other chemical carcinogens that fit into the definition of a “dose-related increase of malignant neoplasms.” For example, a chemical like 1,2,3-trichloropropane caused multiple-site tumors in each of the four sex-species groups at each of the exposure levels from 3 to 30 mg · kg⁻¹. It is certainly a “super carcinogen” and exhibits “clear evidence.” Conversely, a chemical like d-limonene caused a dose-related increase in tubular cell tumors of the kidney (0% in controls versus 16% and 22% in exposed groups) in male rats (at 25 to 50 times the exposure concentrations used for 1,2,3-trichloropropane), and no tumors occurred in any of the other three sex-species experimental groups.

This chemical would likewise be placed into the “clear evidence” category. Obviously one must examine the total data set before equating these two chemicals for their potential of causing harm to humans. A simple scheme whereby chemicals are grouped according to the number of positive responses has now been proposed (3, 24). It too is not a fully satisfactory answer; yet it does permit some easy differentiation among chemicals with respect to “strength of the evidence” and rapid awareness of those considered the most potentially hazardous to humans.

Category 2

The NTP gives the following description for studies in category 2 (“some evidence”): “Some evidence of carcinogenic activity is demonstrated by studies that are interpreted as showing a chemically related increased incidence of neoplasms (malignant, benign, or combined) in which the strength of the response is less than that required for clear evidence.”

In my opinion the major differentiation between “clear evidence” and “some evidence” should hinge on the degree or strength of the response. Both levels represent positive evidence of carcinogenicity, but separate largely with respect to the nuances of the overall response, with apparent emphasis on dose-effect relations, “clear evidence” being the “higher degree of evidence” and “some evidence” being the “lower degree of evidence.” Yet, could “chemically related increase” for “some evidence” be taken to mean the same as “dose-related” for “clear evidence”? This distinction is not obvious and could be a source of confusion. For instance, would chemicals that produce a convincing carcinogenic response only in the top-exposure group be precluded by these definitions from being considered “clear evidence” of carcinogenicity? Perhaps a rhetorical question, but nonetheless strict interpretation would allow such an instance. Interestingly dichlorvos was given a “clear evidence” description for increases in benign tumors of the forestomach only in the top-exposure group (10% in the control group versus 12% and 36% in the two exposure groups). Conversely, “some evidence” was given to a “dose-related increase” of benign and malignant liver tumors in female mice (5% in the control group versus 29% and 24% in the groups exposed to hydroquinone). This category was chosen seemingly because the response was in mouse liver, an organ considered by some to be biologically unreliable for predicting carcinogenic effects in other organs or in other species, including humans. This reasoning is scientifically unfounded (24).

The definition for category 2 could be considered a de facto nondefinition; that is, chemicals are assigned this level of evidence almost by default. There is a positive carcinogenic response that may be “less than” clear evidence, but the lack of definition may cause one in essence to, literally, “discount” the response

because it did not “make” the more obviously positive response “clear evidence.” Thus more emphasis and attention could be given to inform the readers and users of these levels that chemicals placed into either the “clear evidence” or “some evidence” category cause cancer in animals and may pose hazards to humans. Those chemicals inducing responses showing less convincing evidence of a more uncertain nature are placed in the “equivocal evidence” category presented later.

A different problem comes from the ranges of responses that could fit adequately into either category. Certain chemicals cause overwhelming responses in the experiments, for example, many sites, supported by similar findings in the other sex-species experimental units, all or almost all dose-related, the incidence rates much higher than typically observed, exposure concentrations “close” to human exposure levels, no confounding effects from toxicity or differences in body weight or survival, and so on (6). This is generally not what occurs in these studies, and one would be naive to expect either these responses or a certain “no evidence” response to occur routinely. What actually happens more often is a considerably less overwhelming response concomitant with one or several potentially confounding factors. This argument could be used as support for the suggestion to formulate a greater number of categories of “positive” evidence to allow more distinctive differentiation among the chemical groupings. This would become a larger issue if a single or composite level of evidence were selected for all four experimental results, the way IARC or the United States Environmental Protection Agency must do in evaluating overall hazard estimates.

Category 3

“Equivocal evidence” comprises category 3, for which the NTP states: “Equivocal evidence of carcinogenic activity is demonstrated by studies that are interpreted as showing a marginal increase of neoplasms that may be chemically related.”

In my view chemicals in this category are closer to “evidence of carcinogenicity” than they are to “no evidence.” Typically, increases in tumors or the occurrence of a few rarely occurring neoplasms are observed in a chemically exposed group, but there is no real confidence that the response is truly “evidence of carcinogenic activity” (eg, a few benign papillomas in the forestomach in a gavage experiment) or that the response is absolutely associated with chemical exposure (eg, two or three tumors of even a relatively uncommon tumor in one exposure group). Other results that may be labeled “equivocal” include responses in the middle- or low-exposure group and not in the high-exposure group, a marginal cluster of benign lesions with questionable malignant potential, an increase in comparison with the status of controls that is possibly within the historical control rate range (this latter compari-

son is often cited to propose discounting or ignoring an effect that may be dose-related and increased compared with that of the concurrent control group), an increase in a common tumor with no supporting evidence in the other sex or species, no scientific rationale indicating the existence of the chemical in the affected organ, or a statistically marginal (and biologically implausible) effect.

Basically, the strength of “equivocal evidence” is considered insufficient to permit a conclusion of a positive association between the toxicologic response and the chemical; yet some correspondence exists that obviates placement into the “no evidence” category. There may have been some confounding factors that could have had an impact on the results — exposures were too low, causing the inherent property of carcinogenic activity to have been underestimated, or exposures were too high, causing debilitation and shortened survival. Thus the findings are considered to be uncertain, and one needs to decide if another experiment is needed for the best interests of public health. The possibility of another experiment should receive more attention, especially for those chemicals inducing single sex-species positive responses when the other three sex-species units exhibit “no evidence.” See the preceding discussion as well.

As an example of a chemical being associated with a tumorigenic response in one organ of one sex of one species, hydrochlorothiazide might serve to illustrate a chemical “causing” liver tumors in mice (males only) that was “down-graded” from “some evidence” (NTP recommendation) to “equivocal evidence” by the NTP public peer review panel. The age-adjusted tumor rates were 16% for the controls versus 23% for the low-dose group and 44% for the top-exposure group. These biologic and statistical data appear to support “dose-related” increases in neoplasms, rather than a “marginal increase,” and perhaps other more discriminatory studies may be warranted.

Category 4

Category 4 is the “no evidence” category. For it NTP states: “No evidence of carcinogenic activity is demonstrated by studies that are interpreted as showing no chemically related increases in malignant or benign neoplasms [or a combination of malignant and benign neoplasms].”

With the suggested clarifying addition in brackets, I believe this definition is appropriate and adequate. Notice the deliberate avoidance of the words “negative evidence,” which can be taken to mean a protective or chemotherapeutic outcome [decreases may happen for nonneoplastic or toxic effects as seen for ascorbic acid (26), for tumors as observed for 4-hexylresorcinol (27), and for d,l-amphetamine sulfate (28)]. Furthermore, “no evidence” should not be considered equivalent to guaranteed safety under any circumstances, largely because of the relative insensitivity of

the experiment (small numbers of animals), and the specific conditions under which the experiments were conducted (eg, inbred rats and hybrid mice, and no other species or strains). With particular emphasis and caution, one must be certain that the animals were adequately challenged with optimum exposure concentrations and that chemically or environmentally ("diseases") associated early mortality did not preclude the detection of late appearing or rarely occurring tumors. Importantly, one must recall that the duration of these studies is only 24 months (not "life-span").

Regardless, this level of evidence should give reasonable but not absolute comfort about the reduced or absent potential for causing cancer in humans. However, other toxic effects must be considered from a public health aspect [eg, toluene does not cause cancer in laboratory animals but is a neurotoxicant in animals and humans (29)].

Category 5

The last category (category 5) is that of "inadequate study." With respect to it, NTP states: "Inadequate study [or experiment] of carcinogenic activity is demonstrated by studies [or experiments] that, because of major qualitative or quantitative limitations, cannot be interpreted as valid for showing either the presence or absence of carcinogenic activity."

Hopefully this category will be seldom used. In cases whereby the response was obviously positive, then the use of "inadequate" will likely not be an issue even if the experiment had some serious flaws, the rationale being that carcinogenic responses under almost any circumstances cannot on a scientific basis or in good public health conscience be ignored or discounted. And one should consider additional studies to determine better the full scope of carcinogenicity.

One example of an experiment that was perhaps mistakenly considered "inadequate" was the male mice portion of the study done with 1,3-dichloropropene (Telone II). The rationale was that one-half of the control animals at one year died of suppurative inflammation of the heart, and thereby made proper contemporary comparisons difficult. Nonetheless the exposed male mice exhibited increases in tumors similar to those seen in female mice (urinary bladder, lung and forestomach) and to one seen in male and female rats (forestomach). This supportive evidence might have been used more fully in the selection of a category of evidence.

Selecting levels of evidence

To decide on a particular level of evidence entails factors in addition to knowledge and experience. First, one must not be influenced by the societal value or the economic importance of a chemical, that is, whether the chemical being evaluated is considered essential to the economy (eg, unleaded gasoline), is associated with a common everyday exposure (eg, fluoride in the drink-

ing water and in oral hygiene consumer products), or is a laboratory curiosity (to which exposure is almost nonexistent) should have no impact whatsoever on the scientific evaluation. Other persons are responsible for considering and dealing with the questions of benefits versus risks and with the attendant societal and political issues (30).

Second, the initial question one needs to consider is whether the chemical in a particular experiment causes a chemical-induced response. Could the response have possibly occurred if the animals did not receive the chemical? Hence the importance of the neoplastic findings in the concurrent and historical control groups. Only after one decides that the chemical indeed caused the neoplastic response should one begin to think about the most appropriate and proper level of evidence. In the NTP program it was decided that only two levels of evidence — clear and some — were needed for categorizing positive responses. Thus the subsequent consideration and scientific debate on a particular set of "positive" data takes place to determine the most appropriate level or "degree" of evidence (clear or some).

Third, another issue one must ignore in deciding the level of evidence concerns the eventual perceived potential impact the evaluations will have on the public, on the scientific and environmental communities, or on the regulator or the regulated. In my experience this point (and the other two already mentioned) can be an almost insurmountable (and most often a "subconscious") burden on some people, and it strengthens the need for a team-evaluation approach with a subsequent independent and balanced peer review panel of experts. The panel members must be selected for their expertise and by a careful and judicious selection process to have representation from all viewpoints from which to judge an NTP evaluation objectively. After all, these long-term carcinogenesis studies (showing either evidence of carcinogenicity or no evidence of carcinogenicity) do have impact on public, environmental, and occupational health, and at times on economics. And if these studies are judged to be adequately designed, conducted, and evaluated, there is little possibility that most chemicals will ever be subjected to another long-term carcinogenesis study.

There are some exceptions. An economically important chemical could be studied again to confirm the reported results or to determine those experimental exposure concentrations that, under the conditions of the study, do not cause any observable carcinogenic effects. A chemical considered to be carcinogenic to humans and reportedly not carcinogenic in laboratory animals will likely be evaluated again, perhaps with other species or strains or at a higher exposure concentration, largely to determine why there are incongruent results (benzene, eg) (25). This concept may be a valid reason for placing more concern on other chemicals not showing any response in animals. Typically however "no evidence" chemicals rarely receive much

attention for further long-term carcinogenesis studies. Certain chemicals may be restudied for mechanistic purposes however (1,3-butadiene and methylene chloride, eg) (31).

One group of chemicals that should be given more attention for reexamination is that of chemicals which cause cancer in a single site in one sex of one species, the rationale being that for some the results may not be "logical" and readily explainable and one would like to know with more certainty if the response is "real" and thus repeatable. If the NTP data base of 313 chemicals studied adequately in each of the four experiments is examined, one finds that 38 chemicals (12%) induced cancer in one sex of one species, and another 56 chemicals (18%) caused cancer in two of four experiments. Each of these would need to be examined to evaluate those causing single site cancer or, for the latter grouping, those producing perhaps single (but different) site tumors in both sexes of a single species.

Fourth, in some instances a chemical that is placed into the "clear evidence" category may induce other exposure-associated tumors that would not by themselves constitute "clear evidence," and these are often labeled as tumors that "may be related" to the chemical exposure. This is often taken to mean "equivocal evidence." Yet in this instance the induced increase may in fact be part of the totality of "clear evidence" or might stand independently as "some evidence," and this information would not be obvious. Likewise for "some evidence" and "may be related" effects for the same chemical. In my view, tumors that are caused by the chemical could be more inclusively and appropriately grouped or listed under the highest evidence category and should not be separated or split into different levels for the same experiment. For illustration, the organ sites exhibiting chemically associated carcinogenicity could be listed in order of significance, or one could specify in the summary conclusion statement "clear evidence of carcinogenicity based primarily on carcinomas of the lung, liver, and kidney, as well as on carcinomas of the skin and preputial gland." This then serves to convey a more straightforward message than one may conclude from reading that "carcinomas of the skin and preputial gland may have been related to chemical exposure."

General opinion and discussion

Since June 1983 the levels of evidence of carcinogenicity have served constructively both the NTP and the public. After gaining several years' experience with the initial categories, changes were clearly needed to allow some refinements and perhaps altered emphasis in definitions. In March 1986 these objectives were largely met. From my perspective, however, replacing the universally used term "carcinogenicity" with the words "carcinogenic activity" for the categories of evidence caused confusion and might even appear capri-

cious. This concern comes principally from the difficulty of defining or attempting to explain adequately the later terminology, or for even describing satisfactorily a basis for the change.

First, the term carcinogenic activity, when used to describe results from a long-term carcinogenicity study on rodents, is confusing because it can be taken to mean an inherent property of the particular chemical that may or may not be manifest in experimental studies using rodents (regardless of whether the results show a carcinogenic response or not). As an example, short-term *in vitro* (or *in vivo*) mutagenicity tests give results that permit one to use the term carcinogenic activity with scientific credibility. These two words imply a potential inherent property of the chemical, but also the direct requirement for a particular system in which to express this inherent property. Carcinogenic activity can also be taken to imply mechanism, whereby the chemical may be acting through some contributory or indirect fashion rather than being a direct-acting carcinogen. Since these studies do not reveal mechanisms of the carcinogenesis process or ultimate responses, it cannot be assumed that a chemical may possess some property or another that will only influence cancer by some indirect means. The major end point in long-term chemical-exposure studies on rodents is cancer or no cancer. Thus the study is designed to determine carcinogenesis and not just an inherent property that might contribute to that process.

Second, from my observations, scientists from both the research and regulatory communities of the United States, and in other countries as well, thought this change was material and significant. A conspicuous policy definitional change such as this could have had a remarkable impact on how one used data, results, and evaluations from these studies. Moreover, no distinction between carcinogenicity and "carcinogenic activity" has been made. Fortunately, most seem to have either ignored or not noticed any change or simply decided to view the change in wordage as no change at all. That is, many decided that "carcinogenic activity" simply or likely meant the same as carcinogenicity. If that were the case, one might then wonder why a change in terminology was made. In essence, therefore, positive results from long-term chemical carcinogenesis experiments conducted on rodents mean that the study has demonstrated the carcinogenicity of the chemical, not that the chemical has carcinogenic activity in rodents.

Other considerations for evaluating carcinogenicity studies

The most significant advancement to come out of the March 1986 meeting was the development of a list of those items and concepts that one ordinarily would include, perhaps intuitively by those experienced in making these interpretations, in the evaluation process leading to a conclusion about a particular experiment

or data set. Thus for those persons perhaps only tangentially involved in this field or not yet fully experienced, NTP staff compiled a collection of key factors that were considered to define more clearly the boundary of an individual category of evidence. These considerations include: (i) adequacy of the experimental design and conduct, (ii) tumor types and organ and tissue sites, (iii) multiplicity in site-specific neoplasms, (iv) occurrence of common versus uncommon neoplasms, (v) progression (or lack thereof) from benign to malignant neoplasms, as well as from preneoplastic to neoplastic lesions, (vi) the combination of benign and malignant tumor incidences known or thought to represent stages of progression in the same organ or tissue, (vii) latency in tumor induction, (viii) metastases, (ix) supporting information from proliferative lesions (hyperplasia) in the same site of neoplasms or from other experiments (same lesion in another sex or species), (x) presence or absence of dose relationships, (xi) statistical significance of the observed tumor increase, (xii) the concurrent control tumor incidence, as well as the historical control rate and variability for a specific neoplasm, (xiii) survival-adjusted analyses and false positive or false negative concerns, (xiv) structure-activity correlations, especially for other chemicals of the same class, and (xv) genetic toxicology.

Of course there could easily be more items of importance that one could add to this list of key factors, but these are the main ones that should always be considered when long-term chemical carcinogenesis data and results are evaluated. Proposed mechanisms of action are not yet considered to be understood well enough to become significant factors in the evaluations, but certainly all relevant biologic information (proto-oncogene activation and tumor suppressor gene inactivation, pharmacology and pharmacokinetics, hormonal influences, deoxyribonucleic acid damage and repair, cell proliferation, biologic verisimilitude, among others) should be considered when a single level of evidence of carcinogenicity is decided on.

Even when one follows all these guidelines, however, the most important attribute for evaluating long-term chemical carcinogenesis studies is knowledgeable and objective scientific staff. The staff should have direct experience in the design, conduct, data collection, results interpretation, and evaluation of experiments. Because these studies and interpretations do have impact on public health policy, peer review beyond that used successfully by scientific journals is essential.

Peer review

Scientific and public peer review remains fundamental to good research, be it basic, developmental, or applied. Peer review provides indispensable objective (or balanced) criticism and public exposure of new findings and ideas and maintains a critical oversight of the quality of science. Furthermore, public trust in science-

based governmental decisions is gained and maintained if research conducted and supported or used by the government utilizes a system of unbiased and balanced peer review. In the particular case of the NTP draft technical reports on the toxicology and carcinogenesis studies of chemicals, public meetings to accomplish peer review of the results and proposed evaluations are held approximately three times yearly in Research Triangle Park, North Carolina. In advance of these meetings, draft reports are also made available to anyone upon request, and anyone can attend the meetings.

Panel members are nominated and selected, and also serve, as independent scientists, not as representatives of any institution, company, public interest group, or governmental agency. In this capacity, panel members have the following five major responsibilities: (i) to ascertain that all relevant literature data have been adequately cited and interpreted, (ii) to determine whether the design and conditions of the NTP studies were appropriate, (iii) to ensure that the technical report presents the experimental results and conclusions fully and clearly, (iv) to judge the significance of the experimental results by scientific criteria, and (v) to assess the evaluation of the evidence of carcinogenicity and other observed toxic responses. Included as an important part of the printed technical report are renditions of the key debates that took place, as well as a summary of the votes of the individual panel members on particular levels of evidence.

Since their adoption by the NTP, from June 1983 through July 1991, the levels of evidence have been used to evaluate 144 long-term chemical carcinogenesis studies comprised of 530 individual experiments. The independent peer review panel concurred on 506 of the premeeting levels of evidence recommended by NTP staff. Roughly one-half of the proposed changes were for decreases in the level of evidence, and one-half were for increases in the level of evidence. None involved a shift of more than one level of evidence.

Results extrapolation

Even though it is the judgement of the NTP staff, and the subsequent ad hoc peer review panel, that the experiments were designed and conducted adequately and showed "no evidence of carcinogenicity," the evaluations (in which the laboratory animals did not have a greater incidence of site-specific neoplasms than did concurrent control animals) do not mean necessarily or absolutely that a chemical is not a carcinogen in as much as the experiments were conducted under a limited set of conditions. Nonetheless, the National Institute of Environmental Health Sciences and the NTP endorse the concept that, under the conditions of exposing both sexes of two rodent species to relatively high exposure levels for extended periods of times (at least two years), it is scientifically reasonable, for practical purposes, to regard those chemicals that did not induce any neoplasia as presenting little (but not the

absence of) potential carcinogenic hazard to humans. However, one must remain mindful and receptive to new data or scientific advances that could later influence this interpretation.

Positive carcinogenicity results in these experiments demonstrate that a "chemical" is carcinogenic for laboratory animals under the conditions of the study, and, depending on the overall weight of the carcinogenicity evidence, indicate that exposure to the "chemical" should be regarded for prudent public health and scientific purposes as being a likely carcinogenic hazard to humans.

Concluding remarks

Levels of evidence of carcinogenicity as introduced and used by IARC and subsequently adopted and adapted by the NTP have stimulated constructive and valuable dialogue among the various factions having intense interest in these evaluative interpretations. The process of evaluating experimental carcinogenesis studies is difficult and at times arduous. The public peer reviews and scientific debates on the conclusions made from the experimental results are both stimulating and necessary. These public forums permit the NTP to maintain objectivity and to contribute to overall public health. Because much scientific knowledge has been gained since early 1986 when the current definitions were endorsed and have been used by the NTP, the time for a reevaluation seems propitious and desirable.

Acknowledgments

For reviewing this manuscript, offering valued suggestions, and helping to maintain my scientific objectivity over the years, I thank Dr J Bucher, Dr L Hart, Dr J Haseman, and Dr R Melnick. Furthermore, I am indeed fortunate and privileged to have had the opportunity to study, learn, and work at IARC in the monograph program during 1977–1979; this was a period when much of the effort was being directed towards categories of evidence. Obviously, many of the ideas generated during the initiation and adoption of the IARC "categories" were mimicked when the NTP "levels" were formulated. I am likewise honored to have been part of the National Toxicology Program for the 10-year period 1979–1989, during which the program enjoyed considerable scientific and intellectual growth.

References

- Tomatis L, Aitio A, Wilbourn J, Shuker L. Human carcinogens so far identified. *Jpn J Cancer Res* 1989;80:795–807.
- Huff JE, Rall DP. Relevance to humans of carcinogenesis results from laboratory animal toxicology studies. In: Last JM, Wallace RB, ed. *Maxcy-Rosenau-Last's public health & preventive medicine*. 13th ed. Norwalk, CT: Appleton-Century-Crofts, 1992, 433–40&453–7.
- Huff JE, Haseman JK, Rall DP. Scientific concepts, value, & significance of chemical carcinogenesis studies. *Annu Rev Pharmacol Toxicol* 1991;31:621–52.
- International Agency for Research on Cancer (IARC). Preamble. In: Coffee, tea, mate, methylxanthines, and methylglyoxal. Lyon: IARC, 1991:13–33. (IARC monographs on the evaluation of carcinogenic risks to humans; vol 51.)
- Huff JE. Design strategies, results and evaluations of long-term chemical carcinogenesis studies. *Scand J Work Environ Health* 1992;18 Suppl 1:31–7.
- Huff JE, Hoel DG. Perspective and overview of the concepts and value of hazard identification as the initial phase of risk assessment for cancer and human health. *Scand J Work Environ Health* 1992;18 Suppl 1:83–9.
- Huff JE, Moore JA. Carcinogenesis studies & experimental data interpretation/evaluation at the National Toxicology Program. In: Järvisalo J, Pfaffi P Vainio H, ed. *Industrial hazards of plastics & synthetic elastomers*. New York, NY: Liss, 1984:43–64.
- Maronpot RR. Considerations in the evaluation and interpretation of long-term animal bioassays for carcinogenicity. In: Milman HA, Weisburger EK, ed. *Handbook of carcinogen testing*. Park Ridge, NJ: Noyes Publications, 1985:372–82.
- Rall DP. National Toxicology Program (NTP): levels of evidence of carcinogenicity used to describe evaluative conclusions for NTP long-term toxicology and carcinogenesis studies; request for comments. *Fed Regist* 1986;51:2579–82.
- Rall DP. National Toxicology Program (NTP): notice of modifications in the levels of evidence of carcinogenicity used to describe evaluative conclusions for NTP long-term toxicology & carcinogenesis studies. *Fed Regist* 1986;51:11843–4.
- International Agency for Research on Cancer (IARC). Preamble: IARC monograph programme on the evaluation of the carcinogenic risk of chemicals to humans. In: IARC. IARC internal technical report no 77/002. Lyon: IARC, 1977.
- International Agency for Research on Cancer (IARC). Overall evaluations of carcinogenicity: an updating of IARC monographs volumes 1 to 42. Lyon: IARC, 1987. (IARC monographs on the evaluation of carcinogenic risks to humans; suppl 7.)
- Environmental Protection Agency. Guidelines for carcinogen risk assessment. *Fed Regist* 1986;51:33992–4054.
- Occupational Safety and Health Administration. Identification, classification and regulation of potential occupational carcinogens. *Fed Regist* 1980;45:5001–296.
- Office of Science and Technology, Interagency Staff Group on Chemical Carcinogenesis, Executive Office of the President. Chemical carcinogens: a review of the science and its associated principles. *Fed Regist* 1985; 50:10371–442. (Also published in *Environ Health Perspect* 1986;67:201–82.)
- Department of Health Services, Health and Welfare Agency, State of California. Guidelines for chemical carcinogen risk assessments & their scientific rationale. Berkeley, CA: Department of Health Services, Health and Welfare Agency, State of California, 1985.
- American Conference of Governmental Industrial Hygienists. Guidelines for the classification of occupational carcinogens. *Appl Occup Environ Hyg* 1991;6: 616–20.
- Ashby J, Doerr NG, Flamm FG, Harris JE, Hughes DH, Johannsen FR, et al. A scheme for classifying carcinogens. *Regul Toxicol Pharmacol* 1990;12:270–95.
- Griesemer RA, Cueto C. Toward a classification scheme for degrees of experimental evidence for the carcinogenicity of chemicals for animals. In: Montesano R, Bartsch H, Tomatis L, ed. *Molecular and cellular aspects*

- of carcinogen screening tests. Lyon: International Agency for Research on Cancer, 1980:259—81. (IARC scientific publications; no 27.)
20. Chu KC, Cueto C, Ward J. Factors in the evaluation of 200 National Cancer Institute carcinogen bioassays. *J Toxicol Environ Health* 1981;8:251—80.
 21. Vainio H, Magee P, McGregor D, McMichael AJ, ed. Mechanisms of carcinogenesis in risk evaluation, Lyon: International Agency for Research on Cancer. In press. (IARC scientific publications; no 116.)
 22. Vainio H, Wilbourn J. Identification of carcinogens within the IARC monograph program. *Scand J Work Environ Health* 1992;18 Suppl 1:64—73.
 23. Vainio H, Coleman M, Wilbourn J. Carcinogenicity evaluations and on-going studies: the IARC databases, *Environ Health Perspect* 1991;96:5—9.
 24. Huff JE, Haseman JK. Long-term chemical carcinogenesis experiments for identifying potential human cancer hazards: collective data base of the National Cancer Institute & National Toxicology Program (1976—1991). *Environ Health Perspect* 1991;96:23—31.
 25. Huff JE, Haseman JK, DeMarini DM, Eustis S, Maronpot RR, Peters AC, et al. Multiple-site carcinogenicity of benzene in Fischer 344 rats and B6C3F1 mice. *Environ Health Perspect* 1989;82:125—63. (Also published in NTP technical report; no 289.)
 26. Douglas JF, Huff JE, Peters AC. No evidence of carcinogenicity for L-ascorbic acid (vitamin C) in rodents. *J Toxicol Environ Health* 1984;14:605—9. (Also published in NTP technical report; no 247.)
 27. Chhabra RS, Huff JE, Haseman JK, Hall A, Baskin G, Cowan M. Inhibition of some spontaneous tumors by 4-hexylresorcinol in F344/N rats and B6C3F1 mice. *Fundam Appl Toxicol* 1988;11:685—90. (Also published in NTP technical report; no 330.)
 28. Dunnick JK, Eustis SL. Decreases in spontaneous tumors in rats and mice after treatment with amphetamine. *Toxicology* 1991;67:325—32. (Also published in NTP technical report; no 387.)
 29. Huff JE. Toxicology & carcinogenesis studies of toluene (CAS No. 108—88—3) in F344/N rats & B6C3F1 mice (inhalation studies). Research Triangle Park, NC: National Toxicology Program, 1990. (NTP technical report; no 371.)
 30. Rall DP. Issues in the determination of acceptable risk. *Ann Ny Acad Sci* 1981;363:139—44.
 31. Melnick RL, Huff JE, Chou BJ, Miller RA. Carcinogenicity of 1,3-butadiene in C57Bl/6 X C3H F₁ mice at low exposure concentrations. *Cancer Res* 1990;50: 6592—9.