

Criteria for NTP Developmental Toxicology Studies

**Report from the
Reproductive and Developmental Criteria Working Group (RDCWG)
of the
NTP Board of Scientific Counselors**

Submitted by:
Edward Carney, Ph.D.,
George Daston, Ph.D., rapporteur

At the National Toxicology Program (NTP) Board of Scientific Counselors (BSC) meeting on June 12, 2008, Dr. Paul Foster, NTP, provided an outline of the criteria used by the NTP to describe the results of the carcinogenesis bioassay and briefly discussed the NTP's plan to develop similar criteria for reproductive and developmental toxicology studies. The NTP proposed to form working groups to formulate these criteria. Thus, the purpose of the Reproductive and Developmental Criteria Working Group (RDCWG) was to investigate the utility of having specific criteria for describing the results from individual NTP reproductive and developmental toxicology reports to indicate the strength of the evidence for their conclusions. The RDCWG was composed of 10 scientists representing academia, industry, and government. Dr. Edward Carney, The Dow Chemical Company, a member of the NTP BSC, chaired the RDCWG. Dr. Barry Delclos, National Center for Toxicological Research/NTP, Dr. Mark Cesta, National Institute of Environmental Health Sciences/NTP, and Dr. Paul Foster, Acting Branch Chief, Toxicology Branch, served as technical advisors to the RDCWG. Drs. George Daston, Procter and Gamble and Barbara Shane, NTP Executive Secretary, served as rapporteurs. Also attending the meeting was Dr. Mary Wolfe, NTP Federal Official. The full RDCWG roster is attached [Appendix A]. The RDCWG met September 11 and 12, 2008, at the Hilton Garden Inn Durham/Southpoint Hotel, 7007 Fayetteville Road, Durham, NC.

The NTP developed draft criteria for describing results of NTP reproductive and developmental studies that were modeled after the NTP criteria used to evaluate carcinogenicity studies. Dr. Foster was the lead scientist for this effort. Prior to the RDCWG meeting, the draft criteria were evaluated internally. The RDCWG was tasked to first evaluate the draft criteria for reproductive toxicology studies and then the draft criteria for developmental toxicology studies. This report addresses the revision and discussion by the RDCWG regarding the draft criteria for NTP developmental toxicology studies. A separate report was prepared to discuss the RDCWG's evaluation of the draft criteria for NTP reproductive toxicology studies.

Dr. Foster opened the meeting by providing the background for the development of the criteria by NTP. He presented information regarding NTP's developmental toxicology testing strategies and a discussion of the developmental toxicology criteria. Materials provided to the RDCWG included: the draft criteria [Attachment B], a set of case studies for testing the utility and applicability of the draft criteria for reaching conclusions on NTP developmental toxicology studies [Attachment C], a list of issues for discussion by the RDCWG [Attachment D], and the carcinogenicity criteria [Attachment E]. The RDCWG was given the following charge:

Evaluate the suitability and utility of the proposed criteria for describing the results from individual NTP developmental toxicology studies to indicate the strength of the evidence for their conclusions.

The RDCWG completed the case study exercise, deliberated on the proposed criteria, and produced the following revised criteria based on those discussions. In revising the draft criteria, the RDCWG deliberated a number of issues that are discussed below (see "RDCWG Discussion"). Their deliberations resulted in the following revised draft criteria:

EXPLANATION OF LEVELS OF EVIDENCE FOR DEVELOPMENTAL TOXICITY

The NTP describes the results of individual studies of chemical agents, and notes the strength of the evidence for conclusions regarding each study. Generally, each study is confined to a single laboratory animal species, although in some instances, multiple species may be investigated under the purview of a single study report. Negative results, in which the study animals do not exhibit evidence of developmental toxicity, do not necessarily imply that a chemical is not a developmental toxicant, but only that the chemical is not a developmental toxicant under the specific conditions of the study. Positive results demonstrating that a chemical causes developmental toxicity in laboratory animals under the conditions of the study are assumed to be relevant to humans, unless data are available which demonstrate otherwise. In addition, such positive effects should be assumed to be primary effects, unless there is clear evidence that they are secondary consequences of excessive maternal toxicity. Given that developmental events are intertwined in the reproductive process, effects on developmental toxicity may be detected in reproductive studies. Evaluation of such developmental effects should be based on the NTP Criteria for Levels of Evidence for Developmental Toxicity.

It is critical to recognize that the “levels of evidence” statements described herein only describe developmental **hazard**. The actual determination of **risk** to humans requires exposure data and other analyses that are not considered in these summary statements. This fact is particularly important to keep in mind when communicating study results to the general public.

Five categories of evidence of developmental toxicity are used in the NTP Technical Report series to summarize the strength of the evidence observed in each experiment: two categories for positive results (**clear evidence and some evidence**); one category for uncertain findings (**equivocal evidence**); one category for no observable effects (**no evidence**); and one category for experiments that cannot be evaluated because of major design or performance flaws (**inadequate study**). In addition, the study’s lowest observed adverse effect level is reported for positive results, and the highest dose level tested is reported for the **no evidence** category. Application of these criteria requires professional judgment by individuals with ample experience and an understanding of the animal models and study designs employed. For each study, conclusion statements are made using one of the following five categories to describe the findings. These categories refer to the weight of evidence of the experimental results and not to potency or mechanism.

- **Clear evidence** of developmental toxicity is demonstrated by a dose-related¹ effect on one or more of its four elements (embryo-fetal death, structural malformations, growth retardation or functional deficits) that is not secondary to excessive maternal toxicity. A statement to the effect of “This study has a lowest observed adverse effect level of

¹ The term “dose-related” describes any dose relationship, recognizing that the treatment-related responses for some endpoints may be non-monotonic due to saturation of exposure or effect, overlapping dose-response behaviors, change in developmental manifestation at difference dose levels or other phenomena.

XXXX mg/kg/d or other appropriate units (e.g. ppm in diet, mg/L in drinking water) for developmental toxicity” should accompany the evidence statement.

- **Some evidence** of developmental toxicity, relative to clear evidence, is characterized by greater uncertainties or weaker relationships with regard to dose, severity, magnitude, incidence, persistence, and/or decreased concordance among affected endpoints. A statement to the effect of “This study has a lowest observed adverse effect level of XXXX mg/kg/d or other appropriate units (e.g. ppm in diet, mg/L in drinking water) for developmental toxicity” should accompany the evidence statement, except in those instances in which the “some” classification has been based on uncertainties about the relationship that precludes confident information of the LOAEL.
- **Equivocal evidence** of developmental toxicity is demonstrated by marginal or discordant effects on developmental parameters that may or may not be related to the test article.
- **No evidence** of developmental toxicity is demonstrated by data from a well conducted, adequate study that are interpreted as showing no biologically relevant evidence of chemically-related effects on development. A statement to the effect of “This study had no observable adverse reproductive toxicity at the highest dose tested (XXXX mg/kg/d or other appropriate units (e.g. ppm in diet, mg/L in drinking water))”.
- **Inadequate study** of developmental toxicity is demonstrated by a study that, because of major design or performance flaws, cannot be used to determine the presence of developmental toxicity.

When a conclusion statement for a particular experiment is selected, consideration must be given to key factors that would extend the boundary of an individual category of evidence. Such consideration should allow for incorporation of scientific experience and current understanding of developmental toxicity studies in laboratory animals, particularly with respect to interrelationships between endpoints, impact of the change on development, relative sensitivity of endpoints, normal background incidence, and specificity of the effect. For those evaluations that may be on the borderline between two adjacent levels, some factors to consider in selecting the level of evidence of developmental toxicity are given below:

- Increases in severity and/or prevalence (more individuals and/or more litters) as a function of dose generally strengthen the level of evidence, keeping in mind that the specific manifestation may change with increasing dose. For example, malformations may be observed at a lower dose level, but higher doses may produce embryo/fetal death.
- Because of the relationship between maternal physiology and development, evidence for developmental toxicity may be greater for a selective effect on the embryo-fetus or pup, although there may be exceptions.
- Effects seen in many litters may provide stronger evidence than effects confined to one or a few litters even if the incidence within those litters is high.

- Concordant effects (syndromic) may strengthen the evidence of developmental toxicity. Single endpoint changes by themselves may be weaker indicators of effect than concordant effects on multiple endpoints related by a common mechanism.
- In order to be assigned a level of “clear evidence” the endpoint(s) evaluated should normally show a statistical increase in the deficit, or syndrome, on a litter basis.
- In general, the more animals affected, the stronger the evidence; however, effects in a small number of animals across multiple, related endpoints should not be discounted, even in the absence of statistical significance for the individual endpoint(s). In addition, rare malformations with low incidence should be interpreted in the context of historical controls and may be biologically important.
- Consistency of effects across generations in a multi-generational study strengthens the level of evidence. However, if effects are observed in the F1 generation but not in the F2 generation (or the effects occur at a lesser frequency in the F2 generation), this may be due to survivor selection (i.e., if the effect is incompatible with successful reproduction, then the affected individuals will not produce offspring).
- Transient changes (e.g., pup weight decrements, reduced ossification in fetuses) by themselves may be weaker indicators of an effect than persistent changes.
- Insights from supportive studies (e.g., toxicokinetics, ADME, computational models, structure-activity relationships) and developmental findings from other *in vivo* animal studies (NTP or otherwise) should be drawn upon when interpreting the biological plausibility of an effect.
- Uncertainty about the presence of developmental toxicity in one study may be lessened by effects (even if not identical) that are observed in a second species.
- The studies should be well designed and be of adequate experimental design and statistical power.
- New technical approaches and highly sensitive techniques need to be appropriately characterized to build confidence in their utility, and their usefulness as indicators of effect is increased if they can be associated with changes in traditional endpoints.

Ancillary recommendations

For **Some/Equivocal Evidence** calls or discordant effects, it may be important to convey whether additional studies are needed to clarify the effects. If additional studies are recommended, then the specific area of concern and recommended approach should be articulated.

RDCWG Discussion

The RDCWG deliberated a number of issues in determining what revisions were needed to the draft criteria and the factors that should be considered when determining the appropriate level of evidence for describing a study's results. This was a rich discussion that resulted in agreement by the Working Group on the levels of evidence, and on the bullet items listed in the sections above entitled "Other Key Points for Consideration" and "Ancillary Recommendations."

As described earlier, the process by which the RDCWG addressed its charge began with a draft ("straw man") criteria document provided by NTP. The RDCWG agreed with the general approach of structuring the proposed criteria after the carcinogenicity criteria and agreed that the proposed number of categories was appropriate. There was significant discussion about the hazard-based nature of the summary statements, with concerns expressed about the general public's tendency to view hazard as synonymous with risk. As such, the RDCWG asks NTP and other end users of the criteria documents to use adequate caution when using the criteria and summary statements to communicate to the general public. Along similar lines, there was a vibrant discussion on whether or not to include some indication as to the dose level required to elicit adverse reproductive effects, as many felt that this information is fundamental to the characterization of a chemical's potential hazard. The RDCWG recognized the need to communicate this dose level information in a simple, coherent manner, leading to the recommendation that a short statement declaring the LOAEL (for "clear evidence" or "some evidence" categories) or NOAEL (for "no evidence category") accompany the summary statement.

Beyond these general issues, much of the discussion was intended to refine the specific wording of the criteria and was driven by the case studies. These case studies were provided by both NTP and members of the RDCWG and were purposely designed to reside in the "transition zones" between categories. The case studies initially were reviewed and scored separately by each RDCWG member and the results tallied so the group could view the degree of concordance (or lack thereof). The ensuing discussions revealed the thought process behind each member's score, and proved quite constructive in refining the criteria so that the boundaries between categories were as clear as reasonably possible.

Given the nature of fetal morphology data, the RDCWG agreed that study interpretation and application of the criteria require a strong working knowledge of fetal morphological evaluation in order to distinguish between effects of varying severity. In particular, the difference between "clear evidence" and "some evidence" will often be influenced by severity of effect. Some skeletal variants fit a general fingerprint indicative of a slight developmental delay, whereas other effects may signal something more serious, particularly if the same structures are malformed in fetuses at higher dose levels. Similarly, expert knowledge is required to judge the plausibility of relationships between fetal and maternal effects. For example, it is plausible that a decrease in maternal weights in the last week of gestation could have caused a decrease in fetal body weight, but is highly unlikely to have caused an increased incidence of cervical ribs.

As the interpretation of complex data always carries with it some degree of judgement, it is recommended that NTP develop some additional examples by which the criteria were applied in order to accumulate some “case history”. These examples could be developed over time, and would supplement the specific criteria and considerations adopted in the criteria document.

Appendix A

NTP Board of Scientific Counselors Reproductive and Developmental Criteria Working Group

Working Group Members

Kim Boekelheide, M.D., Ph.D.
Professor, Division of Biology and Medicine
Brown University
70 Ship Street
(Chestnut Street Loading Dock)
Providence, RI 02903

Tracie Bunton, D.V.M., Ph.D., D.A.C.V.P.
Pharmaceutical Toxicology and Pathology
Consulting
EICARTE LLC
c/o 150 Irishtown Road
Fairfield, PA 17320

Edward Carney, Ph.D. (Chair)
Technical Leader, Development Reproductive &
General Toxicology
The Dow Chemical Company
Building 1803
Midland, MI 48674

Robert Chapin, Ph.D.
Pfizer
Eastern Point Road, Bldg 274
Groton, CT 06340

George Daston, Ph.D.
Miami Valley laboratories
The Procter and Gamble Company
11810 E. Miami River Rd.
Cincinnati, OH 45253

James M. Donald, Ph.D.
Chief, Reproductive Toxicology and
Epidemiology Section
Reproductive and Cancer Hazard Assessment
Branch
Office of Environmental Health Hazard
Assessment
1001 I Street, P.O. Box 4010, MS 12B
Sacramento, CA 95812

L. Earl Gray, Ph.D.
USEPA
NHEERL, Reproductive Toxicology Division
Endocrinology Branch
EB (MD-72)
Research Triangle Park, NC 27711

Barry McIntyre, Ph.D., D.A.B.T.
Reproductive Toxicology
Safety Evaluation Center
Schering-Plough Research Institute
556 Morris Avenue, Bldg. 12
Summit, NJ 07901

Kenneth M. Portier, Ph.D.
Director of Statistics
Statistics and Evaluation Center
Research Department
American Cancer Society
250 Williams Street, Suite 600
Atlanta, GA 30303

Shelley Tyl, Ph.D.
Center for Life Sciences and Toxicology
RTI International
Hermann Laboratory Building, Room 124
3040 Cornwallis Road
Research Triangle Park, NC 27709

Technical Advisors

Mark Cesta, D.V.M., D.A.C.V.P.
Cellular and Molecular Pathology Branch
National Toxicology Program
National Institute of Environmental
Health Sciences
P.O. Box 12233, MD B3-06
Research Triangle Park, NC 27709

Barry Delclos, Ph.D.
Department of Biochemical Toxicology
United States Food and Drug Administration
National Center for Toxicological Research
3900 NCTR Road HFT 110
Jefferson, AR 72079

Paul Foster, Ph.D.
Toxicology Branch
National Toxicology Program
National Institute of Environmental Health Sciences
P.O. Box 12233, MD EC-34
Research Triangle Park, NC 27709

NTP Executive Secretary

Barbara Shane, Ph.D., D.A.B.T.
Office of Liaison, Policy, and Review
National Institute of Environmental Health Sciences
P.O. Box 12233, MD A3-01
Research Triangle Park, NC 27709

NTP Federal Official

Mary S. Wolfe, Ph.D.
Deputy Program Director for Policy
Director, NTP Office of Liaison, Policy, and Review
National Toxicology Program
National Institute of Environmental Health Sciences
P.O. Box 12233, MD EC-31
Research Triangle Park, NC 27709

Appendix B

Levels of Evidence Criteria for Developmental Toxicity

1. Clear Evidence of Developmental Toxicity

Demonstrated by the results of a study or studies, in one or more species, that indicate a clear treatment-related effect in one or more of the four elements of developmental toxicity (embryo-fetal death, structural malformations, growth retardation or functional deficits) that is not secondary to overt systemic toxicity.

Concordant effects in multiple endpoints that indicate biological plausibility of the response would also provide clear evidence of developmental toxicity.

In order to be assigned a level of “clear evidence” the end point(s) evaluated should normally show a statistical increase in the deficit, or syndrome, on a litter basis.

2. Some Evidence of Developmental Toxicity

Demonstrated by a study or studies indicating a treatment-related increase in deficits of developmental parameters in which the strength of response, incidence, or biological plausibility are insufficient for clear evidence.

The presence of developmental toxicity that is only significant on a fetal, and not litter basis, would be assigned the level of “some evidence”.

3. Equivocal Evidence of Developmental Toxicity

Demonstrated by a study or studies that are interpreted as showing marginal deficits in developmental parameters that may or may not be chemically related.

4. No Evidence of Developmental Toxicity

Demonstrated by a well-conducted study or studies that are interpreted as showing no biologically relevant evidence of chemically related deficits in developmental toxicity parameters.

5. Inadequate Study of Developmental Toxicity

Demonstrated by a study that because of major qualitative or quantitative limitations cannot be interpreted as valid for showing the presence or absence of developmental toxicity. A study may be deemed inadequate if it produced neither developmental nor systemic toxicity (unless tested at an NTP limit dose level).

Other Key Points for Consideration

- The relationship of maternal to developmental toxicity. (Concern may be greater for a selective effect on the embryo-fetus or pup – but not always!)
- The severity of responses. (Is death worse than malformation, or growth retardation? Current dogma would allocate equal weight to any of the four manifestations of developmental toxicity; e.g., embryo-fetal death is equivalent to growth retardation.)
- Incidence of responses (on a litter and individual fetus basis).
- Differences between effects seen at high incidence in few litters and low(er) incidence in many litters.
- Confounding of effects in a continuum (e.g., fetal death can mask malformation production). “Odd” dose-response relationships e.g. delays in growth that may lead to malformations and perhaps lead to fetal death that are biologically plausible.
- The need for well designed and conducted studies of adequate experimental power.
- Concordance of responses. (For example, were some skeletal variants seen in the presence of a fetal weight reduction?)
- Low incidences of rare effects. (How should we handle biological versus statistical significance?)
- Appropriate use of historical control data to understand background control incidence of developmental effects.
- Known structure-activity relationships.
- Effects in one species are sufficient for a conclusion. Is confidence raised by effects in multiple species?

Appendix C

Case Study Exercise for Developmental Toxicology Studies

Introduction and General Points

The “Levels of Evidence” criteria are loosely based on those used in the NTP toxicology and carcinogenicity reports. All three sets of draft criteria for our non-cancer toxicities that are being evaluated in BSC work groups (reproductive and developmental toxicity and immunotoxicity) have wording about “concordance of end points and biological plausibility,” because we are dealing with multiple end points in these toxicities where some are redundant and/or should be linked. Our approach is essentially a weight of evidence type approach – the greater the weight of evidence, the more likely a more severe conclusion will be reached.

Also, note that for reproductive toxicity and immunotoxicity, we are proposing an effect on **integrated function** to meet the criteria for a “clear evidence” designation. This approach is more difficult to apply for developmental toxicity. Here the conclusion statements are proposed to be based on effects on one or more of the four components of developmental toxicity (i.e., embryo-fetal death, structural malformation, growth retardation, or functional deficit) in a **litter-based** analysis for studies involving pre-natal necropsy and, if appropriate, for those studies evaluating post-natal developmental effects.

Note that the “key considerations/points” are outlined separately. You may wish to look at these first to aid you in how to consider data and put the draft criteria into context.

In the case studies, the descriptions are purposely short (to generate some discussion) and provided in a series of bullets. For the purpose of this developmental toxicity exercise, please assume that these data are from studies that meet (or exceed) the current EPA guidelines for multigenerational or developmental toxicity studies (control plus three treated dose groups).

As a rule of thumb for these studies, a decrease in terminal body weight greater than 10% between a treated group and controls exceeds the normal amount of systemic toxicity expected at the highest dose level – but beware – if, for example, a test article produces fetal death or a marked effect on fertility, then the maternal body weights could be reduced by >10%. This would still be a meaningful toxicological effect, but not necessarily one resulting as a consequence of selecting too high a dose level for the dam.

If an effect is noted in the bullets, please assume it is statistically significant and dose related (unless this is specifically noted otherwise). If effects are not specifically noted please assume they were not significantly different from controls (and not missing!). For the developmental toxicity case studies, F= significant on a fetal basis and L= significant on a litter basis.

Case Study # 1

- No effect on final body weight for dams (8% decrease at top dose level)
- Increase in post-implantation loss (L)
- Decreased fetal weight (L)
- Increase in specific skeletal variants (F)

Case Study # 2

- Increase in skeletal and visceral variants (F)
- Small, but not statistically significant increase in the number of fetuses with a rare malformation (mid and high dose groups)
- Fetal death (L) at mid and top dose levels in the presence of some (10%) decrease in terminal maternal body weight

Case Study # 3

- Increase in number of skeletal variants (L)
- Small decrease in fetal weight (L) in presence of (a) a modest maternal weight decrement (< 7 %) or (b) without any maternal effects

Case Study # 4

- Increase in overall malformation incidence, but not dose-related
- Decrease in live fetuses at top dose level (L)
- Fetal weight decrease (L)
- Increase in some skeletal variants (L)
- 5% decrease in maternal body weight at top dose level

Case Study # 5

- 13% decrease in terminal maternal body weight at top dose level
- Decreased fetal survival (L)
- Increase in malformations and decrease in fetal weight (L) at top dose level
- Some skeletal and visceral variants increased (F) at lower dose levels

Case Study # 6

- 15% decrease in terminal maternal body weight at top dose level
- Small (max 8%), but significant effects on fetal weight and increase in skeletal variants (L)

Case Study # 7

- Delay in male puberty (3 days) at top dose in presence of a 13% decrease in body weight at weaning
- No effects on AGD

Case Study # 8

- Significantly reduced AGD on PND 1 (L; when corrected for birth weight and litter size)
- Small increase (up to 4 per male) in nipple retention (L; all in male offspring), that (a) is not present at adulthood or (b) is present at adulthood

Case Study # 9

- Small (15%), statistically significant decrement in motor activity (L) in presence of 5% body weight loss at top dose level
- No changes in histopathology noted

Case Study # 10

- Increase in AGD in females (L)
- Small advancement (2 days) in vaginal opening (puberty) (L) at top dose level
- No significant effect on body weight

Case Study # 11

- 15% decrease in terminal maternal body weight at top dose level
- Decrease in fetal weight (L) and skeletal variants (L) at top dose level

Case study #12

- 6% decrease in body weight at top dose
- Decrease in number of fertile pairs (top dose F1 only)
- Increase in liver weight (5% max)
- Increase in malformations of the prostate (L) and cryptorchidism (L)
- Decrease in sperm count in F1 only
- Delays in PPS
- Decrease in testis weight (F1 only)
- Testis histology (F1 only)
- Low incidence of uterus unicornis (2 at top dose, 0 at mid dose, and 1 at low dose)

Case study #13

- Decrease in dam body weight gain (11% in top dose) on GD6-9 and 10-12. No effect on terminal dam body weight
- Increase in sternebral skeletal variants (L) and skeletal variants of the front and hind paws (F)
- No effects on fetal body weight
- Two rare malformations (of the jaw in different litters) in the top dose (historical control rate 0/3000)

Case study # 14

- Maternal toxicity at top dose: a 10-15% decrease in maternal body weight gain
- Developmental toxicity at top dose: statistically significant increases (L) in select skeletal variations (shortened 13th rib, rudimentary cervical ribs, incomplete ossification of the skull and sternebrae)
- Mid and low dose levels: no maternal or developmental effects

Case study # 15

- Maternal toxicity at top dose: body weight *loss* on GD 6-8, with body weight gains for GD 6-15 slightly decreased
- Developmental toxicity at top dose: decreased fetal body weight, increases in two minor skeletal variants, and delayed ossification of the axial skeleton. Increased incidence of microphthalmia (4.4% and 19% incidences in fetuses and litters, respectively) that was slightly outside of the historical control range (note: F344 rats have a higher background incidence relative to SD rats, and have shown sporadic clusters of this malformation). No other malformations were observed.
- No maternal or developmental effects at low and mid dose levels

- Repeat study (same dose levels) in F344 rats 22 years later showed nearly identical maternal toxicity, but only two cases of microphthalmia in high dose group vs. no cases in the control group. No historical data are available due to discontinuation of this strain for developmental toxicity.

Appendix D

Issues for Discussion with NTP BSC Reproductive and Developmental Criteria Working Group

1. Conclusions statements for NTP studies are hazard and not risk-based, to facilitate comparison across chemicals using the same study types. These conclusion statements are voted upon by the NTP Board of Scientific Counselors (BSC) in its advisory role to the NTP Executive Committee, which contains representatives from our sister regulatory agencies that can use this information in quantitative risk assessment decisions.
2. It would be helpful if we could model conclusion criteria for non-cancer studies based on that currently employed for the NTP carcinogenicity studies (attached), to generate some consistency in approach and wording for both the BSC and the public.
3. NTP staff recognizes that for many of the non-cancer toxicity studies, we are dealing with multiple (inter-related) end points very different from cancer studies. Thus, the NTP cancer study approach to levels of evidence in drawing study conclusions will require some “finessing” to achieve the desired level of consistency.
4. NTP staff also recognizes the desirability to use a graded (hazard identification) conclusion scheme, such that a single positive finding does not necessarily result in the highest level of conclusion. We have considered those end points that affect overall function to merit the highest level of conclusion (clear evidence of toxicity). So, there may be a statistically significant, dose-related decrease in some end point (for example, sperm count in a reproduction study), but without a concomitant effect on animal function (e.g., fertility or litter size parameters), it would not merit the clear evidence category.

Appendix E

EXPLANATION OF LEVELS OF EVIDENCE OF CARCINOGENIC ACTIVITY

The National Toxicology Program describes the results of individual experiments on a chemical agent and notes the strength of the evidence for conclusions regarding each study. Negative results, in which the study animals do not have a greater incidence of neoplasia than control animals, do not necessarily mean that a chemical is not a carcinogen, inasmuch as the experiments are conducted under a limited set of conditions. Positive results demonstrate that a chemical is carcinogenic for laboratory animals under the conditions of the study and indicate that exposure to the chemical has the potential for hazard to humans. Other organizations, such as the International Agency for Research on Cancer, assign a strength of evidence for conclusions based on an examination of all available evidence, including animal studies such as those conducted by the NTP, epidemiologic studies, and estimates of exposure. Thus, the actual determination of risk to humans from chemicals found to be carcinogenic in laboratory animals requires a wider analysis that extends beyond the purview of these studies.

Five categories of evidence of carcinogenic activity are used in the Technical Report series to summarize the strength of the evidence observed in each experiment: two categories for positive results (**clear evidence and some evidence**); one category for uncertain findings (**equivocal evidence**); one category for no observable effects (**no evidence**); and one category for experiments that cannot be evaluated because of major flaws (**inadequate study**). These categories of interpretative conclusions were first adopted in June 1983 and then revised in March 1986 for use in the Technical Report series to incorporate more specifically the concept of actual weight of evidence of carcinogenic activity. For each separate experiment (male rats, female rats, male mice, female mice), one of the following five categories is selected to describe the findings. These categories refer to the strength of the experimental evidence and not to potency or mechanism.

- **Clear evidence** of carcinogenic activity is demonstrated by studies that are interpreted as showing a dose-related (i) increase of malignant neoplasms, (ii) increase of a combination of malignant and benign neoplasms, 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.
- **Some evidence** of carcinogenic activity is demonstrated by studies that are interpreted as showing a chemical-related increased incidence of neoplasms (malignant, benign, or combined) in which the strength of the response is less than that required for clear evidence.
- **Equivocal evidence** of carcinogenic activity is demonstrated by studies that are interpreted as showing a marginal increase of neoplasms that may be chemical related.
- **No evidence** of carcinogenic activity is demonstrated by studies that are interpreted as showing no chemical-related increases in malignant or benign neoplasms.
- **Inadequate study** of carcinogenic activity is demonstrated by studies that, because of major qualitative or quantitative limitations, cannot be interpreted as valid for showing either the presence or absence of carcinogenic activity.

For studies showing multiple chemical-related neoplastic effects that if considered individually would be assigned to different levels of evidence categories, the following convention has been adopted to convey completely the study results. In a study with clear evidence of carcinogenic activity at some tissue sites, other responses that alone might be deemed some evidence are indicated as “were also related” to chemical exposure. In studies with clear or some evidence of carcinogenic activity, other responses that alone might be termed equivocal evidence are indicated as “may have been” related to chemical exposure.

When a conclusion statement for a particular experiment is selected, consideration must be given to key factors that would extend the actual boundary of an individual category of evidence. Such consideration should allow for incorporation of scientific experience and current understanding of long-term carcinogenesis studies in laboratory animals, especially for those evaluations that may be on the borderline between two adjacent levels.

These considerations should include:

- adequacy of the experimental design and conduct;
- occurrence of common versus uncommon neoplasia;
- progression (or lack thereof) from benign to malignant neoplasia as well as from preneoplastic to neoplastic lesions;
- some benign neoplasms have the capacity to regress but others (of the same morphologic type) progress. At present, it is impossible to identify the difference. Therefore, where progression is known to be a possibility, the most prudent course is to assume that benign neoplasms of those types have the potential to become malignant;
- combining benign and malignant tumor incidence known or thought to represent stages of progression in the same organ or tissue;
- latency in tumor induction;
- multiplicity in site-specific neoplasia;
- metastases;
- supporting information from proliferative lesions (hyperplasia) in the same site of neoplasia or in other experiments (same lesion in another sex or species);
- presence or absence of dose relationships;
- statistical significance of the observed tumor increase;
- concurrent control tumor incidence as well as the historical control rate and variability for a specific neoplasm;
- survival-adjusted analyses and false positive or false negative concerns;
- structure-activity correlations; and
- in some cases, genetic toxicology.