National Toxicology Program
U.S. Department of Health and Human Services

Explanation of Levels of Evidence for Developmental Toxicity

The NTP describes the results of individual studies of chemical agents and other test articles, 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 test article is not a developmental toxicant, but only that the test article is not a developmental toxicant under the specific conditions of the study. Positive results demonstrating that a test article 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 describe only developmental **hazard**. The actual determination of **risk** to humans requires exposure data that are not considered in these summary statements.

Five categories of evidence of developmental toxicity are used 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)**. 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 strength of the evidence of the experimental results and not to potency or mechanism.

Levels of Evidence for Evaluating Developmental System Toxicity

- ***Clear evidence of developmental toxicity*** is demonstrated by data that indicate a dose-related[1] effect on one or more of its four elements (embryo-fetal death, structural malformations, growth retardation or functional deficits) that is not secondary to overt maternal toxicity.
- ***Some evidence of developmental toxicity*** is demonstrated by dose-related effects on one or more of its four elements (embryo-fetal death, structural malformations, growth retardation or functional deficits), but where there are greater uncertainties or weaker relationships with regard to dose, severity, magnitude, incidence, persistence, and/or decreased concordance among affected end points.
- ***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 study with appropriate experimental design and conduct that are interpreted as showing no biologically relevant effects on developmental parameters that are related to the test article.
- ***Inadequate study of developmental toxicity*** is demonstrated by a study that, because of major design or performance flaws, cannot be used to determine the occurrence of developmental toxicity.

[1] The term "dose-related" describes any dose-response relationship, recognizing that the test article-related responses for some endpoints may be non-monotonic due to saturation of exposure or effect, overlapping dose-response behaviors, changes in immunologic manifestations at different dose levels or other phenomena.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

March 2009

When a conclusion statement for a particular study is selected, consideration must be given to key factors that would support the selection 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 end points, impact of the change on development, relative sensitivity of end points, 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 affected litters) as a function of dose generally strengthen the level of evidence, keeping in mind that the specific manifestation may be different with increasing dose. For example, malformations may be observed at a lower dose level, but higher doses may produce embryo-fetal death.
- 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.
- Because of the complex relationship between maternal physiology and development, evidence for developmental toxicity may be greater for a selective effect on the embryo-fetus or pup.
- Concordant effects (syndromic) may strengthen the evidence of developmental toxicity. Single end point changes by themselves may be weaker indicators of effect than concordant effects on multiple end points related by a common process or mechanism.
- 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.
- In general, the more animals affected, the stronger the evidence; however, effects in a small number of animals across multiple, related end points should not be discounted, even in the absence of statistical significance for the individual end point(s). In addition, rare malformations with low incidence, when interpreted in the context of historical controls, may be biologically important.
- Consistency of effects across generations in a multi-generational study may strengthen 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 for resistance to the effect (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.
- Uncertainty about the occurrence of developmental toxicity in one study may be lessened by effects (even if not identical) that are observed in a second species.
- 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.
- New assays and techniques need to be appropriately characterized to build confidence in their utility: their usefulness as indicators of effect is increased if they can be associated with changes in traditional end points.

http://ntp.niehs.nih.gov/go/10003

Paul M. Foster, Ph.D.
Discipline Leader for Reproduction and Development
Acting Chief • Toxicology Branch • National Toxicology Program • NIH/NIEHS
P.O. Box 12233 MD K2-12, Research Triangle Park, NC 27709
(919) 541-2513 • foster2@niehs.nih.gov