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

Explanation of Levels of Evidence for Reproductive 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 reproductive toxicity, do not necessarily imply that a chemical is not a reproductive toxicant, but only that the chemical is not a reproductive toxicant under these specific conditions. Positive results demonstrating that a chemical causes reproductive 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 toxicity to non-reproductive organ systems. 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 reproductive **hazard**. The actual determination of **risk** to humans requires exposure data that are not considered in these summary statements.

Five categories of evidence of reproductive 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 with and 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; if warranted, these conclusion statements should be made separately for males and females. These categories refer to the strength of the evidence of the experimental results and not to potency or mechanism.

Levels of Evidence for Evaluating Reproductive System Toxicity

- ***Clear evidence of reproductive toxicity*** is demonstrated by a dose-related[1] effect on fertility or fecundity, or by changes in multiple interrelated reproductive parameters of sufficient magnitude that by weight of evidence implies a compromise in reproductive function.
- ***Some evidence of reproductive toxicity*** is demonstrated by effects on reproductive parameters, the net impact of which is judged by weight of evidence to have potential to compromise reproductive function. Relative to clear evidence of reproductive toxicity, such effects would be characterized by greater uncertainties or weaker relationships with regard to dose, severity, magnitude, incidence, persistence and/or decreased concordance among affected endpoints.
- ***Equivocal evidence of reproductive toxicity*** is demonstrated by marginal or discordant effects on reproductive parameters that may or may not be related to the test article.
- ***No evidence of reproductive toxicity*** is demonstrated by data from a study with appropriate experimental design and conduct that are interpreted as showing no biologically relevant effects on reproductive parameters that are related to the test article.
- ***Inadequate study of reproductive toxicity*** is demonstrated by a study that, because of major design or performance flaws, cannot be used to determine the occurrence of reproductive 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 reproductive toxicity studies in laboratory animals, particularly with respect to interrelationships between endpoints, impact of the change on reproductive function, 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 reproductive 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, histological changes at a lower dose level may reflect reductions in fertility at higher dose levels.
- In general, the more animals affected, the stronger the evidence; however, effects on a small number of animals across multiple related endpoints should not be discounted, even in the absence of statistical significance for the individual end point(s). In addition, effects with low background incidence when interpreted in the context of historical controls may be biologically important.
- Consistency of effects across generations may strengthen the level of evidence. However, special care should be taken for decrements in reproductive parameters noted in the F1 generation that were not seen in the F0 generation, which may suggest developmental as well as reproductive toxicity. Alternatively, 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 the nature of the effect resulting in 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) by themselves are weaker indicators of effect than persistent changes.
- Single end point changes by themselves are weaker indicators of effect than concordant effects on multiple, interrelated end points.
- Marked changes in multiple reproductive tract endpoints without effects on integrated reproductive function (i.e. fertility and fecundity) may be sufficient to reach a conclusion of clear evidence of reproductive toxicity.
- Insights from supportive studies (e.g., toxicokinetics, ADME, computational models, structure-activity relationships) and reproductive findings from other *in vivo* animal studies (NTP or otherwise) should be drawn upon when interpreting the biological plausibility of an effect.
- New assays or 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/18711

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