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

NTP Unveils New Non-Cancer Evaluation Criteria

The National Toxicology Program (NTP) is working to bring the same rigorous standards it uses for classifying the outcomes of its cancer studies to its non-cancer studies.

Join Us at the Society for Toxicology Meeting Tuesday, March 17, 2009

1:30-2:30 PM • Room 337

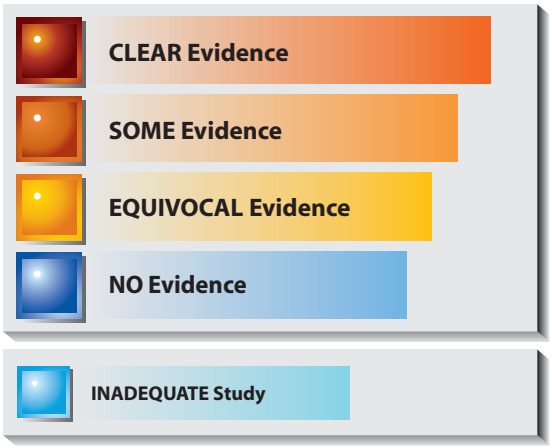
"NTP Criteria for Hazard Identification in Non-Cancer Studies" Session

NTP scientists discuss new evaluation criteria for reproductive, developmental and immunotoxicity studies.

Timeline for the Development of Non-Cancer Evaluation Criteria

- The NTP set out to develop new criteria to classify the outcomes from its non-cancer studies, including reproductive, developmental and immunotoxicity studies. These new criteria build off the classification system established in 1983 by the NTP to evaluate its cancer studies.
- In June 2008, the NTP established two working groups of the NTP Board of Scientific Counselors to provide input and help refine draft criteria developed by the NTP for evaluating its reproductive, developmental and immunotoxicity studies.
- In November 2008, the NTP Board of Scientific Counselors voted to accept and provide additional comments on the working group reports.
- In December 2008, the NTP Executive Committee accepted the working group reports and provided additional comments on the criteria.
- In February 2009, the NTP finalized the levels of evidence statements for the three sets of criteria incorporating recommendations from the Board and the Executive Committee.
- In March 2009, the NTP presents the evaluation criteria to attendees at the Society for Toxicology (SOT) 48th Annual Meeting and ToxExpo™ in Baltimore, MD.
- The NTP expects to begin applying the immunotoxicology criteria to studies for peer review by the end of 2009. The first reproductive and developmental studies featuring the new criteria could appear as early as 2010.

NTP proposes using five categories or "levels of evidence" to describe its non-cancer study findings, including reproductive, developmental and immunotoxicity studies.



"We have a desire to have uniform [non-cancer] criteria for evaluating chemicals across studies and for studies across chemicals, much as we have for the cancer bioassays," said Toxicology Branch Acting Chief and reproduction and development discipline leader Paul Foster, Ph.D.

For more information contact:

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

Dori Germolec, Ph.D.
Discipline Leader for Immunology
Toxicology Branch
National Toxicology Program • NIH/NIEHS
P.O. Box 12233 MD K2-12, Research Triangle Park, NC 27709
(919) 541-3230 • germolec@niehs.nih.gov

http://ntp.niehs.nih.gov/go/9399
for immunotoxicology criteria

http://ntp.niehs.nih.gov/go/10003
for developmental criteria

http://ntp.niehs.nih.gov/go/18711
for reproductive criteria

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
March 2009

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

Explanation of Levels of Evidence for Immune System 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 immunotoxicity, do not necessarily imply that a test article is not an immune system toxicant, but only that the test article is not an immune system toxicant under these specific conditions. Positive results demonstrating that a test article causes immunotoxicity 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 overt toxicity to non-immune organ systems.

It is critical to recognize that the "levels of evidence" statements described herein describe only immunologic **hazard**. The actual determination of **risk** to humans requires exposure data 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 immune system 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 Immune System Toxicity

Clear Evidence of Toxicity to the Immune System

- Is demonstrated by data that indicate a dose-related[1] effect (considering the magnitude of the effect and the dose-response) on more than one functional parameter and/or a disease resistance assay that is not a secondary effect of overt systemic toxicity, or
- Is demonstrated by data that indicate dose-related effects on one functional assay and additional endpoints that indicate biological plausibility.

Some Evidence of Toxicity to the Immune System

- Is demonstrated by data that indicate a dose-related effect on one functional parameter with no other supporting data, or
- Is demonstrated by data that indicate dose-related effects on multiple observational parameters without robust effects on a functional immune parameter or a disease resistance assay, or
- Is demonstrated by data that indicate effects on functional parameters or a disease resistance assay that are not-dose-related with other data providing biological plausibility.

[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

Equivocal Evidence of Toxicity to the Immune System

- Is demonstrated by data that indicate effects on functional parameters or a disease resistance assay that are not-dose-related without other data providing biological plausibility, or
- Is demonstrated by data that indicate dose-related effects on a single observational parameter without effects on a functional immune parameter or a disease resistance assay, or
- Is demonstrated by data that indicate effects on the immune system at dose(s) that produce evidence of overt systemic toxicity, or
- Is demonstrated by data that are conflicting in repeat studies.

No Evidence of Toxicity to the Immune System

- Is demonstrated by data from studies with appropriate experimental design and conduct that are interpreted as showing no evidence of biologically relevant effects on the immune system that are related to the test article.

Inadequate Study of Immune System Toxicity

- Is demonstrated by a study that, because of major design or performance flaws, cannot be used to determine the occurrence of immune system toxicity.

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 immunotoxicity studies in laboratory animals, particularly with respect to the interrelationships between endpoints, impact of the effect on immune function, relative sensitivity of endpoints and specificity of the effect. Factors to consider in selecting the level of evidence of immune system toxicity are given below:

- Immunotoxicity is defined in the context that immune responses can be enhanced or suppressed by toxicants. As such, dose-related effects consistent with immunosuppression and immunostimulation will be considered in hazard identification.
- Functional effects, as defined as an alteration in the ability of the immune system to respond to a challenge or stimulus, should usually be weighed more heavily than observational parameters such as alterations in cell counts.
- Increases in severity and/or prevalence (more individuals with the effect) 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 deficits in immune function at higher dose levels.
- Biological plausibility for immunotoxicity must be considered in the context of the nature of the response, the magnitude of the response, and the pattern of the response, as well as the current understanding of immune system structure and function.
- Insights from supportive studies (e.g., toxicokinetics, ADME, computational models, structure-activity relationships) and immunologic findings from other *in vivo* animal studies (NTP or otherwise) should be drawn upon when interpreting the biological plausibility of a change.
- The characterization of immunotoxicity must consider the impact of overt toxicity (e.g., effects on the immune system are not the direct effects of test article treatment, but are indirect effects mediated via stress and/or other dose-related responses).
- The characterization of immunotoxicity must consider the intended pharmacology of the test article. Immunotoxicity is reserved for unintended immunosuppression or immunostimulation.
- Results in one species or one sex are considered sufficient for evidence of immunotoxicity.

http://ntp.niehs.nih.gov/go/9399

Dori Germolec, Ph.D.
Discipline Leader for Immunology
Toxicology Branch • National Toxicology Program • NIH/NIEHS
P.O. Box 12233 MD K2-12, Research Triangle Park, NC 27709
(919) 541-3230 • germolec@niehs.nih.gov

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 reproductive 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

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