NTP
National Toxicology Program

# Criteria for Evaluation of Outcomes in Reproductive, Developmental and Immunotoxicology Studies

Paul Foster, Ph.D.

Society of Toxicology Meeting, March 17, 2009

## Background

- NTP goal: employ the same rigorous standards used historically to review carcinogenicity bioassays to NTP "non-cancer" studies.
- Efforts toward this goal:
  - Training workshops for NTP and contractor pathologists in specialized areas of toxicity (e.g., "enhanced" Immunopathology, Reproductive Pathology).
  - Establishment of Pathology Working Groups to review and agree on the diagnosis of critical lesions in NTP non-cancer studies.
  - Peer review by the NTP Board of Scientific Counselors (BSC) Technical Reports Review Subcommittee of the draft reports for multigenerational studies (e.g., ethinyl estradiol and genistein).
- Desire to have consistent criteria for the evaluation of NTP study outcomes.

## Background -2

- The NTP has long employed specific conclusion statements, that are approved by the BSC, for its "Toxicology and Carcinogenesis" studies.
- These conclusion statements represent a "level of evidence" sentence with regard to carcinogenic potential for each sex within each individual study.
  - Clear evidence
  - Some evidence
  - Equivocal evidence
  - No evidence
  - Inadequate study
- Such an approach allows for comparisons of different studies on the same test substance and for comparisons of conclusions across studies, to ensure similar criteria are employed uniformly.
- The NTP has developed guidance notes as to how these criteria should be applied.

## Conclusion Statements - Cancer Example

- Under the conditions of these 2-year drinking water studies, there was *clear evidence of carcinogenic activity* of sodium dichromate dihydrate in male and female F344/N rats based on increased incidences of squamous cell neoplasms of the oral cavity.

  There was *clear evidence of carcinogenic activity* of sodium dichromate dihydrate in male and female B6C3F1 mice based on increased incidences of neoplasms of the small intestine (duodenum, jejunum, or ileum).

## Application to other studies

- In addition to the chronic toxicity/ carcinogenicity reports that are brought to the BSC for review, more recently other study types have been reviewed (e.g. the multigeneration reproduction studies on genistein and ethinyl estradiol).
- It is the Program's intent to bring more of these large, "non-cancer" studies to the BSC in the future and it would be prudent to develop and use "levels of evidence criteria" to ensure comparability across these specific studies.
- To provide some consistency for the Program, the Board and the Public, it would be sensible to employ similar types of conclusion statements to those currently in place for cancer end points.
- NTP discipline leaders in **reproductive, developmental and immunotoxicology** have been developing such criteria and accompanying guidance documents for their application.

## Some Issues NTP Considered in Developing Draft Criteria

- Conclusions statements for NTP studies are hazard-based, not risk-based, to facilitate comparison across test substances for the same study types.
- Many of NTP's non-cancer toxicity studies include multiple (inter-related) endpoints - different from cancer studies.
- Applying the NTP cancer study "levels of evidence" approach to non-cancer studies would require some "finessing" to achieve the desired level of consistency.
- NTP staff recognized the desirability to use a graded (hazard identification) "level of evidence" scheme for expressing conclusions.
  - Any "positive" response should not result in the highest level.
  - Weight of study evidence approach.

## Some Issues NTP Considered -2

- NTP considered those endpoints that affect overall system function to merit the highest level of evidence ("clear evidence" of toxicity).
- Examples of such functional outcomes would be:
  - A positive result in a host resistance assay (not just a change in specific lymphocyte counts) for immunotoxicity.
  - A decrease in litter size (and not just a decrease in sperm count) for reproductive toxicity.
- Clear positive or negative results should be straightforward in applying the criteria. Findings at the boundaries would present more difficulty.

## Steps Taken toward Refining the Draft Criteria

- NTP conducted “in house” exercises to refine our “draft” criteria
- NTP informally shared the draft criteria with external colleagues to gain feedback for modification and improvements.
- NTP convened working groups of the BSC to provide input on the draft criteria.

## Constitution of Board Work Groups

- Comprised of stakeholders from the NTP BSC, Academia, Industry and Government.
- Practitioners
  - Experts familiar with the nuances of study types, conduct and data interpretation.
- Users of NTP study data
  - Representatives from regulatory bodies with experience in reviewing data from the specific study types.
- Some NTP staff were present at the WG meetings as technical advisors, but did not participate in the review.

## The Process

- NTP staff presented to the Work Groups (WG):
  - Study designs employed by the Program (meet or exceed EPA, OECD or FDA Guidelines).
  - Outlined a straw man of the "levels of evidence" criteria
  - Highlighted some "key issues" for guidance in how the criteria may be applied.
- WG undertook an exercise (individually) in applying the criteria to some (>15) study examples selected to explore the boundaries between levels.
- WG reviewed the exercise as a group to explore individual differences.
- WG then made adjustments, based on the review, to the draft criteria and provided edits on other "key issues" to be used in the application of the criteria.
- Prepared a Work Group report.

## Criteria Summary

- All the Work Groups enjoyed the exercises and interactions.
- All the Work Groups agreed that it was possible to apply criteria to the outcomes from NTP reproductive, developmental and immunotoxicity studies in a systematic fashion.
- The schemes and guidance notes developed (and modified) could use a similar structure to that employed in the review of cancer studies.
- The Work Groups encouraged NTP to publish in the peer-reviewed literature the “finalized” criteria.

## Post Working Group Activities

- Working group reports approved (with additional comments) by the NTP Board of Scientific Councilors (November 2008) and encouraged the Program to keep these criteria "evergreen" and modify, as appropriate, based on future experience in their application.
- Draft criteria reviewed and approved by the NTP Executive Committee (EC, December 2008) with some suggestions for improvement.
- Draft criteria reviewed and revised by NTP staff
  - Addressing BSC and EC comments
  - Consistency between disciplines and harmonization of language
- Presentation of Criteria to SOT attendees (March 2009)

## Implications for Adoption of the New Criteria

- More consistency in the conclusions from NTP studies of reproductive, developmental and immunotoxicity.
- There have not been previous attempts to develop such criteria for these study types.
- Potential for the studies to be noted as “authoritative” by certain regulatory bodies (e.g., Prop 65 – California OEHHA) like the cancer studies.
- Requisite expertise on the NTP BSC (or BSC sub-committees) for review of studies.
- Potential adoption by other groups.

NTP
National Toxicology Program

# NTP Levels of Evidence Criteria for Reproductive Toxicology Studies

Paul Foster, PhD

Society of Toxicology Meeting, March 17, 2009

Toxicology Branch

## Reproductive and Developmental Toxicity Criteria Work group

- Edward Carney (Chair, BSC member)
  - Dow Chemical
- Tracie Bunton (BSC member)
  - EICARTE LLC
- Kenneth Portier (BSC member)
  - American Cancer Society
- Kim Boekelheide (rapporteur reprotox)
  - Brown University
- Robert Chapin
  - Pfizer
- George Daston (rapporteur dev tox)
  - Proctor & Gamble
- James Donald
  - OEHHA
- Earl Gray
  - USEPA
- Barry McIntyre
  - Schering Plough
- Rochelle Tyl
  - RTI International
- Barry Delclos*
  - NCTR, FDA
- Mark Cesta*
  - NTP
- Paul Foster*
  - NTP

*Technical Advisors

## Study types employed by NTP to assess Reproductive Toxicity

- Multigenerational Reproduction Studies
  - EPA/OECD type studies
  - Reproductive Assessment by Continuous Breeding
- Specific Transgenerational studies (commence with timed pregnant animals)
  - Assessments of immune, neurological and reproductive function.
  - Including proposed alternate to the ILSI/ACPA/OECD extended one generation study (Poster 1433, March 18th).
- Supplementary information (e.g., organ weights and histopathology of the reproductive organs) from standard NTP toxicity studies.

## Introductory Comments -1

- It is critical to recognize that the "levels of evidence" statements only describe reproductive **hazard**. The 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, 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 Reproductive Toxicity - 1

- **Clear Evidence of Reproductive Toxicity**
  - Demonstrated by dose-related[1] effects 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.

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

# Levels of Evidence for Reproductive Toxicity - 2

## Some Evidence of Reproductive Toxicity

- 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.

# Levels of Evidence for Reproductive Toxicity - 3

- **Equivocal Evidence of Reproductive Toxicity**
  - 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**
  - 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**
  - Demonstrated by a study that, because of major design or performance flaws, cannot be used to determine the occurrence of reproductive toxicity.

## Key points to consider with the Levels of Evidence criteria

- 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 reproductive toxicity studies in laboratory animals,
  - interrelationships between end points,
  - 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.

## Other Key Points -2

- 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 strengthens the level of evidence.
  - Special care should be taken for decrements in reproductive parameters noted in the $F_1$ generation that were not seen in the $F_0$ generation, which may suggest developmental as well as reproductive toxicity.
  - Alternatively, if effects are observed in the $F_1$ generation but not in the $F_2$ generation (or the effects occur at a lesser frequency in the $F_2$ 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).

## Other Key Points -3

- 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.

NTP
National Toxicology Program

# NTP Levels of Evidence Criteria for Developmental Toxicology Studies

Paul Foster, PhD

Society of Toxicology Meeting, March 17, 2009

## Study types employed by NTP for Assessment of Developmental Toxicity

- EPA/OECD prenatal developmental toxicity studies
- Post-natal developmental toxicity studies
  - Developmental Neurotoxicity
  - Developmental Immunotoxicity
  - Developmental Reproductive toxicity
- Specific Transgenerational studies (commence with timed pregnant animals)

## Levels of Evidence for Developmental Toxicity - 1

- **Clear Evidence of Developmental Toxicity**
  - Demonstrated by dose-related[1] effects 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.
  - [1]The term "dose-related" describes any dose 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, change in manifestation of the effect at different dose levels, or other phenomena.

## Levels of Evidence for Developmental Toxicity - 2

- **Some Evidence of Developmental Toxicity**
  - 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.

## Levels of Evidence for Developmental Toxicity - 3

- **Equivocal Evidence of Developmental Toxicity**
  - 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**
  - 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**
  - Demonstrated by a study that, because of major design or performance flaws, cannot be used to determine the occurrence of developmental toxicity.

## Key points to consider with the Levels of Evidence criteria

- 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.

## Other Key Points -2

- 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 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 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 $F_1$ generation but not in the $F_2$ generation (or the effects occur at a lesser frequency in the $F_2$ 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).

## Other Key Points -3

- 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.