Environmental Health Perspectives
Vol. 94, pp. 265–268, 1991

Meeting Report

In Vitro Teratology

by B. A. Schwetz,* R. E. Morrissey,† F. Welsch,‡ and R. A. Kavlock§

The purpose of this conference was to reevaluate the need for and use of *in vitro* teratology assays; to examine the validation process for *in vitro* tests; and to discuss progress in the validation of *in vitro* teratology screens. Participants enthusiastically supported further development of short-term *in vivo* and *in vitro* systems both as prescreens for developmental toxicity and as experimental systems to explore mechanisms of action of toxicants. The group strongly endorsed the development of an updated reference list ("gold standard") of known developmental toxicants and nontoxicants as essential to further progress in developing and validating prescreening efforts. Independently, an expert group should further evaluate the performance characteristics for a validated prescreen. The limits of usefulness of prescreens for product development, regulatory use, and mechanistic investigations need to be clearly defined. Finally, too few *in vitro* teratology prescreens have been evaluated under multiple-laboratory conditions with common, agreed-upon test agents to draw firm conclusions regarding the merit and reproducibility of *in vitro* teratology prescreens. There was general agreement regarding the need to move several of the assays further along the validation pathway, at least using a short list of reference compounds.

Introduction

In the early 1980s, there was considerable optimism that *in vitro* assays (e.g., the Ames test) would provide valuable insights into the potential for compounds to produce toxicity. Furthermore, the Toxic Substances Control Act requirements for toxicological data placed an additional testing responsibility on manufacturers, and it appeared that *in vitro* toxicology assays might prove useful in prescreening compounds for prioritization for additional, in-depth testing. It was appropriate, therefore, that a Consensus Workshop on *In Vitro* Teratogenesis Testing (*1*) was held in 1981 to discuss *in vitro* teratology assays and how they could fit into toxicity screening and regulatory risk assessment. Validation of *in vitro* teratology prescreens was considered of high priority because of the need to characterize the toxicity of large numbers of untested chemicals and the resultant need for prioritization of chemicals as well as the stimulus to focus our animal and laboratory resources to needs of highest priority. Thus, the earlier workshop was held to review the status of *in vitro* prescreens and also to develop a list of reference chemicals for further validation efforts. The National Toxicology Program established a repository of those chemicals on the reference list.

During the intervening decade, considerable progress has been made toward development and validation of new prescreens for developmental toxicity. A conference was held at the National Institute of Environmental Health Sciences on September 21–22, 1989, to review that progress, which is summarized in this paper. The purpose of the meeting was to reevaluate the need for and use of *in vitro* teratology assays, to examine the validation process for *in vitro* tests, and to discuss progress in the validation of *in vitro* teratology screens. The conference was attended by about 100 people representing a broad spectrum of interests from university, government, industry, and contract laboratory perspectives. Speakers and topics addressed are summarized in Table 1. Points of agreement or disagreement that surfaced as important thoughts from the meeting are summarized in the following sections.

Topics of Discussion

Definitions of Terms and Clarification of Concepts

In vitro teratology assays have a dual purpose, being used in both mechanistic and screening studies. The primary focus of the discussions at this conference was on the predictiveness of the assays for the outcome of Segment II teratology studies as required by regulatory agencies.

There was considerable discussion around the topic of semantics. There was no agreement, for example, on the definitions of prescreen, screen, and definitive studies. Some participants felt that Segment II animal studies, often referred to as teratology studies, are screens for potential adverse effects in humans and that any *in vitro* assay conducted in an initial assessment mode should be considered a prescreen. Others observed that effects in humans are difficult to determine, and Segment II data may be as definitive a set of developmental toxicity data as would be available; thus, an *in vitro* assay would be a screen for these effects. The term "prescreen" will be used here for any assay used to predict the outcome of Segment II studies. Participants

*National Institute of Environmental Health Sciences, P.O. Box 12233, MD D4-02, Research Triangle Park, NC 27709.

†Current Address: Merck, Sharp & Dohme Research Laboratories, Department of Safety Assessment, Building 45-1, West Point, PA 19486.

‡Chemical Industry Institute of Toxicology, P.O. Box 12137, Research Triangle Park, NC 27709.

§U.S. Environmental Protection Agency, HERL/MD 67, Research Triangle Park, NC 27711.

Address reprint requests to B. A. Schwetz, National Institute of Environmental Health Sciences, P.O. Box 12233, MD D4-02, Research Triangle Park, NC 27709.

Table 1. Topics and speakers for current developments in *in vitro* teratology workshop and conference.

Topic	Speaker
In vitro tests and their uses	
Historical perspective—1981 Little Rock Conference	K. Smith
Overview of *in vitro* teratology tests and obstacles to their use	F. Welsch
In vitro tests: uses and needs	
Screening decisions	O. Flint
Mechanistic understanding	T. Sadler
Regulatory decisions	G. Kimmel
The validation process	
What is validation?	R. Scala
Experiences from genetic toxicology validation studies	M. Shelby
Reference list: test strategy progress update	R. Morrissey
Developmental toxicity graphical data profiles	R. Kavlock
Progress update—*in vitro* teratology test systems	
Micromass	O. Flint
Chick retina	G. Daston
Mouse ovarian tumor and human palate cell assays[a]	R. Morrissey
Hydra	M. Johnson
Drosophila	D. Lynch and R. Schuler
FETAX	J. Bantle
Whole embryo culture—mechanisms	A. Fantel
Whole embryo culture—prescreens	B. Schmidt

[a]Only two assays that have been independently evaluated.

agreed that developmental toxicity includes structural malformations, embryonic or fetal death, decrements in fetal body weight, and functional deficits following birth. However, there was no consensus as to how many or which of these types of adverse effects a prescreen must predict to be useful. It is recognized that the term "teratology" is often used interchangeably with "developmental toxicity," especially as it relates to *in vitro* studies. Furthermore, there was no agreement reached concerning what constitutes a "positive" or "negative" response in animal studies, although there are generally agreed-upon working criteria, which include the presence of a statistically significant difference from control values, the presence of a dose-response relationship, and a number of experimental design considerations including the number of animals, the dose selection, the gestational days of treatment, the use of proper statistics, and freedom from significant confounders. Thus, as for *in vivo* data, the group reinforced the importance of established criteria by which the interpretation of *in vitro* tests will be determined.

Usefulness of Prescreens

There were several points of agreement concerning *in vitro* assays: *a*) Tests give varying degrees of information about the site and mode of action of chemicals based on the level of biological organization and the design of the prescreen. Single end point prescreens give a specific answer about that end point and probably little more. Such screens may give very useful information about structure-activity relationships within a chemical group, but have limited potential for use with a wide variety of chemical structures. In contrast, whole organisms screen for more critical events (and thus are more likely to be useful for a wide variety of chemicals) but provide less specific information about the site and mechanism of action compared to a single end point prescreen.

b) While the specific uses of prescreens vary from one laboratory to another, there was agreement that prescreens have potential uses of considerable importance beyond studies of the mechanism of action. The strongest endorsement came from the use of an assay to discriminate between varying degrees of biological activity of members of a class of chemicals when groups of chemicals needed to be prioritized in rank order of potential toxicity.

c) Several prescreens predict the outcome of Segment II or similar studies with an accuracy in the range of 60 to 85%. This compares with the concordance of rat, mouse, and rabbit Segment II studies with human data of 70, 70, and 50%, respectively (*2*). No single prescreen shows 100% concordance with the outcome of these developmental toxicity studies. While different investigators involved in validation of prescreens have used different lists of reference compounds, it is clear that none of the lists were based only on Segment II data. In fact, Segment II studies have not been done for several recognized human teratogens.

d) Since several prescreens are reasonably predictive of the outcome of Segment II studies, the field would probably benefit more from further refinement and evaluation of the most promising prescreens rather than extensive searching for newer and better assays. Certain systems that were shelved for one reason or another, such as the chick embryo test, should perhaps be rereviewed to evaluate their potential usefulness.

e) Some participants felt that *in vitro* assays may be useful in special situations in which, for example, quantities of a compound or a metabolite are very limited or in studying mixtures of compounds.

Performance Characteristics of Prescreens

None of the prescreens were considered to be fully validated. Only the mouse ovarian tumor cell and human embryonic palatal mesenchymal cell assays have been evaluated in independent interlaboratory studies; thus, the level of concordance of prescreens with *in vivo* data cannot be assessed at this time.

Metabolic activation has not been uniformly achieved in the prescreens available at this time. In some cases the metabolic capabilities of the cells, organs, or organisms are not characterized. Whether a mammalian liver-derived 9000*g* supernatant (S-9) fraction or microsomes from cells that are known to have metabolic capability is the most appropriate source of exogenous metabolic activity is a topic for further research. Would a metabolizing system be needed only for compounds showing no activity in assays?

The criteria for interpreting the results of prescreens are specific for each test, but there appear to be no universal criteria for interpreting the results of *in vitro* tests. The minimal study design needs to be established prior to any validation study. What will constitute a positive or negative effect? What is the confidence in a positive or (especially) negative outcome?

The end point concordance (defined as the ability to recapitulate *in vivo* data for either developmental toxicants or nondevelopmental toxicants) and the accuracy (defined as the overall ability to obtain the same outcome as *in vivo* tests) of *in vitro* data for *in vivo* findings is uncertain; concordance may not necessarily be a requirement for a valid test (see "Validation"). Since concordance between Segment II results and effects in humans is variable depending on the species and test compounds, concordance is perhaps a moot point for prescreens.

The level of tolerance for false positives or false negatives was not agreed upon. The was no consensus about the minimum accuracy for a test to be useful. These levels may vary, depending on the purpose for which the assay is conducted.

Validation

Several speakers agreed that there is a general path to assay acceptability. In general, it is necessary to define the model and its significance (and limitations), characterize the test system, and conduct studies with a limited number of compounds. This is followed by an evaluation of the developed test, with intra- and interlaboratory standardization and testing with a larger number of coded compounds. Validated tests have broad acceptance, potential credibility for regulatory purposes, and have been used to test many agents. Studies to understand the scientific nature of the tests may be ongoing during any or all of these stages of validation.

Some prescreens have been developed to an extent where they warrant further validation. None of the prescreens was considered fully validated for broad prescreening purposes of compounds with unknown developmental toxicity. Some prescreens are sufficiently promising that they definitely warrant further validation efforts.

Development and refinement of systems for investigation of mechanisms of toxicity will take its own course. The exploration of systems for mechanistic research is independent of the validation considerations that would be imposed on tests used as prescreens.

There was no agreement reached about the criteria for and the approach to validation. If a prescreen is based on a highly specific mechanistic event, the fact that the event is measured in that test was considered to be sufficient validation by some investigators. Others considered that any prescreen must be able to properly distinguish between chemicals that are known to produce developmental toxicity and those that lack such properties and to distinguish selective embryotoxicity from general toxicity to be considered valid. Relevant concentrations and end points should be used.

It was generally agreed that testing protocols must be standardized to conduct validation studies. Further, it was suggested that investigators test compounds without knowledge of their identity and that outcomes be decided prior to decoding the data. The evaluation criteria must be clearly stated before the validation study is initiated. Interlaboratory concordance is essential; the assay must produce equivalent results in independent laboratories.

Future Efforts and Considerations

Consideration should be given to a validation approach of parallel testing with prescreens as chemicals are tested in Segment II studies. As organizations test specific chemicals in Segment II developmental toxicity screens, they should consider simultaneously testing the same substances in *in vitro* tests as prescreens to compare the outcome with that of Segment II studies. Publishing such comparisons would vastly expedite the evaluation of these tests, especially if industry-wide coordination was achieved.

Perhaps the most commonly agreed upon point of the whole workshop was the need for a new reference list of positive and negative agents as regards mammalian developmental toxicity. There was agreement that a new list was desirable; the chief criticism of the only published list (*3*), which has become the basis for preliminary validations, was that the assessment of maternal toxicity was not complete according to present day standards. It was pointed out that several known human teratogens act only at levels that produce obvious/overt maternal toxicity. Thus, excessive concern over maternal toxicity may not be warranted, as it is impossible at the present time to ascertain a causal relationship between maternal and developmental toxicity. Another criticism of that and other proposed lists is that there may be too many chemicals included that act by similar mechanisms. The primary purpose of a new reference list would be to help focus further validation efforts and research to develop new test systems. Criteria must be established such that chemicals included on the list would permit a rigorous evaluation of any prescreen in subsequent validation studies. A variety of compounds should be considered for the list, representing gradations of developmental toxicity. Criteria for selection of chemicals for the reference list might include, but not necessarily be limited to: *a*) the weight of evidence for effects in animals and humans, *b*) the mechanism of toxicity, chemical structure/function/class, *c*) the adult to developmental toxicity (A/D) ratio in developmental toxicity studies, and *d*) the selection of chemicals that are structurally and configurationally closely related (e.g., enantiomers) known to be either positive or negative in *in vivo* screens.

Random selection from the universe of chemicals was not only considered unnecessary but probably inadvisable. Selection for and inclusion on the list must also take into account the cost and avilability of the chemicals as well as the ability to work safely with the substance in the laboratory.

An expert group should be asked to develop such a list, and consideration should be given to designating a subset of compounds that might be used as a short list to quickly evaluate the potential of a prescreen to produce results in concordance with the reference chemicals of established developmental toxicity. Concordance of test outcome with the short list may then be used as a criterion to proceed with full-scale validation. There was a suggestion that human developmental toxicants may be useful as positive compounds on a short list. Graphical activity profiles for developmental toxicity end point described at the workshop (*4*) may be very useful in summarizing data for many compounds.

Experience from genetic toxicology validation studies designed to predict the outcome of carcinogenicity bioassays leads one to believe that adding additional assays (i.e., creating a battery of tests) may not improve predictive ability of short-term assays and may raise costs to unacceptable levels that are not competitive with conventional *in vivo* screening. Based on the experience in genetic toxicology, it is clear that validation studies must be well designed and managed. It is desirable to standardize the protocol, use random chemical selection from a reference list, include developmental/nondevelopmental toxicants, do the testing blind, conduct intra- and interlaboratory comparisons, and analyze the resulting data uniformly and objectively. It is best to know the mechanistic link to *in vivo* results, to understand the test's role (confirmatory, part of a battery, etc.), statistical methods, and minimal criteria (number of doses, conditions,

duration of exposure, etc.). If a negative result will not have the same impact as a positive one, then there is reason to question the rationale for the entire prescreening approach.

It is necessary to consider the route of exposure and other factors that may influence developmental toxicity, e.g., formaldehyde injected into conceptuses *in utero* may produce very different effects than when entering the body by ingestion. Thus, pharmacokinetics *in vivo* may provide guidance for the relevance of *in vitro* concentrations.

Summary and Recommendations

Participants in the workshop demonstrated considerable enthusiasm for further development and validation of short-term and *in vitro* prescreens both for developmental toxicity testing and as experimental systems to explore mechanisms of action of toxicants. Factors that were considered important for future advances in this field include: *a*) development of a new reference list. The list proposed by Smith et al. (*3*) needs to be updated according to currently available data. The list needs to be expanded or modified according to selection criteria which must be agreed upon by a knowledgeable review panel. *b*) An expert group should further evaluate performance requirements for a prescreen to be considered validated. This group should define what kind of performance (concordance) criteria are considered minimal performance for an assay. *c*) The role for *in vitro* teratology prescreening systems must be clearly defined (either for product development, regulatory use, or mechanistic studies). *d*) Too few *in vitro* teratology prescreens have been evaluated under multiple-laboratory trials with common, agreed-upon test agents to draw firm conclusions regarding the reproducibility of *in vitro* teratology prescreens. There is a need to move several of these assays further along the validation pathway, at least using a short list of reference compounds.

REFERENCES

1. Kimmel, G. L., Smith, K., Kochhar, D. M., and Pratt, R. M. Overview of in vitro teratogenicity testing: aspects of validation and application to screening. Teratog. Carcinog. Mutagen. 2: 221–229 (1982).
2. Schardein, J. L., Schwetz, B. A., and Kenel, M. F. Species sensitivities and prediction of teratogenic potential. Environ. Health Perspect. 61: 55–67 (1985).
3. Smith, M. K., Kimmel, G. L., Kochhar, D. M., Shepard, T. H., Spielberg, S. P., and Wilson, J. G. A selection of candidate compounds for in vitro teratogenesis test validation. Teratog. Carcinog. Mutagen. 3: 461–480 (1983).
4. Kavlock, R. J., Greene, J. A., Kimmel, G. L., Morrissey, R. E., Owens, E., Rogers, J. M., Sadler, T. W., Stack, F., Waters, M. D., and Welsch, F. Activity profiles of developmental toxicity: design considerations and pilot implementation. Teratology 43: 159–185 (1991).