Original Article

A Different Approach to Validating Screening Assays for Developmental Toxicity

George P. Daston,[1]* Robert E. Chapin,[2] Anthony R. Scialli,[3] Aldert H. Piersma,[4] Edward W. Carney,[5] John M. Rogers,[6] and Jan M. Friedman[7]

[1]Procter & Gamble, Cincinnati, Ohio
[2]Pfizer Global R&D, Groton, Connecticut
[3]Tetra Tech Sciences, Arlington, Virginia
[4]RIVM, Bilthoven, The Netherlands
[5]The Dow Chemical Company, Midland, Michigan
[6]Toxicity Assessment Division, NHEERL, ORD, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina
[7]Department of Medical Genetics, University of British Columbia, and Child & Family Research Institute, Vancouver, British Columbia, Canada

BACKGROUND: There continue to be many efforts around the world to develop assays that are shorter than the traditional embryofetal developmental toxicity assay, or use fewer or no mammals, or use less compound, or have all three attributes. Each assay developer needs to test the putative assay against a set of performance standards, which traditionally has involved testing the assays against a list of compounds that are generally recognized as ''positive'' or ''negative'' in vivo. However, developmental toxicity is highly conditional, being particularly dependent on magnitude (i.e. dose) and timing of exposure, which makes it difficult to develop lists of compounds neatly assigned as developmental toxicants or not. **APPROACH:** Here we offer an alternative approach for the evaluation of developmental toxicity assays based on exposures. Exposures are classified as ''positive'' or ''negative'' in a system, depending on the compound and the internal concentration. Although this linkage to ''internal dose'' departs from the recent approaches to validation, it fits well with widely accepted principles of developmental toxicology. **CONCLUSIONS:** This paper introduces this concept, discusses some of the benefits and drawbacks of such an approach, and lays out the steps we propose to implement it for the evaluation of developmental toxicity assays. *Birth Defects Res (Part B)* 89:526–530, 2010. © 2010 Wiley-Liss, Inc.

Key words: ***in vitro methods; validation; developmental toxicity***

INTRODUCTION

The recognition of thalidomide embryopathy 50 years ago demonstrated the need for effective developmental toxicity testing of medications and other chemicals to which women may be exposed during pregnancy. As a consequence, scientists developed standard protocols for developmental toxicity testing, usually in rodents or rabbits, that were thought to be reasonably predictive of effects in human pregnancy, and regulatory authorities adopted requirements for such testing in circumstances such as the assessment of new chemicals for marketing approval.

It was clear from the beginning, however, that these mammalian in vivo assays are expensive in terms of time, requirements for skilled technical support, use of living animals, and use of chemical substances, which often are of quite limited availability early in the product development process. In addition, the scope of industrial chemical testing for developmental toxicity has greatly expanded, with conventional assays able to evaluate only a small fraction of these chemicals. Many alternative assays have, therefore, been developed that are faster and easier to perform, employ phylogenetically lower experimental organisms or in vitro systems, require smaller amounts of test materials, have automatable protocols, provide simpler and more easily reproducible readouts, or depend on computer modeling.

More than 25 years ago, a group of developmental toxicologists recognized the need for standard methods to assess the reliability of alternative assays so that their results could more easily and reliably be used to predict developmental toxicity in the standard in vivo mammalian tests. Whether the consequence of that prediction

*Correspondence to: George P. Daston, Procter and Gamble Company, Cincinnati, Ohio. E-mail: Daston.gp@pg.com

Disclaimer: The research described in this article has been reviewed and approved for publication by the Environmental Protection Agency. Approval does not necessarily signify that the contents reflect the views and policies of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

Published online in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/bdrb.20276

was to avoid the more-toxic molecules (in pharma) or to prioritize toxic compounds for definitive testing (for environmental compounds) was not as important as the hope that there would be the best possible concordance between in vivo and in vitro. One requirement of such a method is a definitive list of compounds that are generally accepted as ''positive'' or ''negative'' developmental toxicants that can be used by developers of alternative assays to gauge the performance and validity of each assay. One of the first lists of compounds was developed by Schmid and colleagues for the validation of rat whole embryo culture (reviewed by Webster et al., 1997); however, this list was criticized because the ''negative'' compounds were inactive chemicals that would be nontoxic in any test system (developmental or not) and the ''positive'' compounds were antimetabolites that would interfere with any biologic system. The ''Smith list'' of 47 compounds (Smith et al., 1983) was subsequently developed but was recognized as having similar limitations. New information about the compounds on the Smith list and new appreciations about the role of maternal toxicity in producing embryofetal malformations prompted a re-evaluation of the Smith list in 1991, but the group of experts assembled to complete this task could not decide which in vivo effects an alternative test should be asked to predict (Schwetz, 1992). Instead, this group identified a list of in vivo test outputs (e.g. adult/developmental [A/D] ratio, slope of the dose–response curve) that might be important to predict.

In the 1990s, the European Centre for the Validation of Alternative Methods (ECVAM) provided a forum in which new alternative assays could be vetted and evaluated. Because reproduction and developmental studies use so many animals and because minimizing animal use was a key driver of the ECVAM process, one of the areas of focus was embryofetal toxicity. To enable progress in this area, a list of compounds was generated (Brown, 2002). This list was used in the evaluation of three in vitro assays for developmental toxicity: the micromass assay, the rat whole embryo culture assay, and the embryonic stem cell test (Genschow et al., 2002). Compounds were classified as strong, weak, or nonteratogens, and the different in vitro tests successfully predicted group status about 80% of the time. It was, then, a big surprise that the embryonic stem cell test underperformed spectacularly when it was later evaluated with a different set of chemicals with known in vivo activities (Marx-Stoelting et al., 2009). As even more ambitious in vitro programs such as ToxCast and Tox21 (at least a part of which is intended to predict developmental toxicity) ramp up, there will be an even more pressing need for appropriate assay validation.

This experience underlines the need for rigorous validation of any surrogate developmental toxicity assay against the known in vivo activity of numerous compounds. Because the record of success in developing in vitro and other simpler assays that accurately predict the results of standard in vivo mammalian developmental toxicity tests is disappointing, new assays will continue to be developed, which means that the need for a ''gold standard'' validation list is stronger than ever.

It was with such motivators and this checkered past in mind that the current group of toxicologists and teratologists assembled and began a new approach, mindful of the substantial expertise that has already been brought to bear on this problem by giants in the field. A summary of our approach, an explanation of its rationale, and a proposal for the implementation of a system based on these concepts are outlined below. This project is being coordinated and administered by the Developmental and Reproductive Toxicology Technical Committee (DART) of the ILSI Health and Environmental Sciences Institute (HESI).

LIMITATIONS OF PREVIOUS APPROACHES

Previous efforts in this area appear to have foundered on two critical issues. The first is maternal toxicity. While it is accepted that fetal malformations in the absence of maternal toxicity present a clear hazard, what should be done with compounds that produce relatively modest developmental effects only at doses sufficient to cause significant maternal toxicity? In in vivo studies, these maternal effects would be exposure-limiting, but in in vitro assays, very high chemical concentrations could be easily achieved, resulting in ''positive'' tests under exposure conditions that would never arise in the real world.

Schmid's approach, still carried forward by some embryo culturists, was to consider growth of the cultured embryo to be the dose limiting factor in place of the missing maternal component (Webster et al., 1997). Under this scheme, malformations occurring in the absence of growth impairment are considered to indicate teratogencity. However, as Webster et al. (1997) pointed out, in vivo experience does not predict such clear distinctions between growth impairment and malformations.

The second problematic issue is also related to dose. Paracelsus is paraphrased as having said, ''The dose makes the poison.'' This principle was restated for teratology by Karnofsky, who maintained that with the right dose, the right species, and the right timing, any compound could be shown to be a teratogen. Similarly, at a low enough exposure level, any compound would be a nonteratogen. A low enough dose of thalidomide is without effect, even in a sensitive species.

''Exposure'' consists of administered dose, route, and timing. We now recognize that the internal dose (e.g. plasma concentration) is a useful way to express exposure, because it bypasses to some extent questions of route and interspecies pharmacokinetic differences. We recommend the use of internal concentration (in maternal blood or the conceptus itself) at a critical time of gestation as the in vivo exposure metric of choice for comparison to concentrations used in a candidate alternative test.

DEFINITION OF DEVELOPMENTAL TOXICANT

Previous efforts to develop validation lists have categorized chemicals as ''positive'' or ''negative,'' ''teratogens'' or ''nonteratogens.'' In some instances, additional descriptors have been added (e.g. strong and weak teratogens). We believe that identification of teratogenicity is so conditional that the classification of *compounds* rather than *exposures* as developmental toxicants or teratogens is meaningless. For example, caffeine is teratogenic in rats at a gavage dose of 100 mg/kg bw/day but not at 25 mg/kg bw 4 times daily or at 1 mg/kg bw/day (reviewed by Christian and Brent, 2001). Should caffeine be rated as a teratogen or a nonteratogen?

For our approach, then, we define a developmental toxicant as ''An exposure (agent at a stated internal dose with stated timing) to the developing organism that leads to a permanent adverse effect.''

We recognize that our ultimate goal is to minimize developmental toxicity in humans, and for an exposure to be called a human developmental toxicant, this definition requires human data. In practice, we rely on data from laboratory animals (rodents and rabbits, principally) to be surrogates for human data. Such reliance assumes a concordance of the target (that it is similarly expressed in these species, at approximately the same stage in development, with roughly the same function) and assumes a reasonable concordance of chemical kinetics. We also take the position that compound-induced effects are of concern to the degree that they represent a permanent adverse change. A delay in ossification or the occurrence of small membranous ventricular septal defects (Solomon et al., 1997) or wavy ribs (Nishimura et al., 1982), which resolve spontaneously after birth, would not be enough to classify an exposure as developmentally toxic.

Low birth weight deserves special attention as to whether it should be used as an endpoint for classifying an agent as a developmental toxicant. In humans, low birth weight has a clinical definition and is considered to be of concern because it is often associated with persistent delays or deficits in function. Decreased fetal weight in animal studies has also been considered to be an indicator of developmental toxicity because (1) it appears to be analogous to low birth weight; (2) it is considered by some to be an integrator of effects on pathways that control growth; and (3) pragmatically, because it is a continuous variable, it is often the most sensitive developmental endpoint evaluated (Schwetz and Harris, 1993). However, decreased fetal weight can also be secondary to maternal toxicity; in such instances decreased fetal weight may not be an indicator that the test agent adversely affects the embryo. In vivo toxicity testing paradigms are insufficient to determine whether decreased fetal weight is a primary developmental effect, so it is unreasonable to expect a less-integrated alternative method to do so. In the face of this uncertainty, it is not reasonable to use decreased fetal weight as a criterion for classifying an agent as a developmental toxicant for the purpose of evaluating the performance of alternative methods.

Conversely, we define a developmental nontoxicant as an exposure (compound, concentration, time of exposure) that does not cause permanent adverse effects. We understand that confidence in noneffects is limited by sample size, adequate dosing range, and other aspects of study design. We note, however, that our emphasis on internal concentration as a key component in distinguishing a toxic exposure from a nontoxic exposure permits the same compound to be considered as a positive developmental toxicant at one concentration and as a negative developmental toxicant at another concentration. This feature captures an all-important aspect of real-world toxicology: the dose–response relationship. It also provides an important practical advantage in creating a gold-standard list of developmental exposures. The number of test exposures on the validation list can be increased by establishing both ''negative'' and ''positive'' exposure conditions for each compound for which adequate data are available.

Our approach does not fit well with current classification schemes that categorize compounds as toxic to reproduction without regard to exposure level. We do not believe that hazard-based classification systems (i.e. those based on compound independent of exposure concentration) are helpful or tenable. Any system that ignores the importance of dose steps away from toxicological realities. Also, our approach may not fit well with the *current* use of alternative tests in pharmaceutical development, where a series of compounds is tested to identify which of them appears least toxic. Under the existing paradigm, the anticipated internal exposure level in humans (or rats) is predicted based on the binding to the targeted receptor, and reasonable guesses are made about the concentrations at which testing will be performed. Such testing is likely to identify the least *potent* of the compounds and not necessarily the compound with lowest potential for developmental toxicity at clinically useful dose levels. A more useful approach, and one that is consistent with the definition of developmental toxicity proposed here, would balance binding at the receptor with ability or potency to induce developmental toxicity.

APPROACHES TO VALIDATION

We propose to create a list of exposures (compounds at specific concentrations) at which developmental toxicity is expected and a list of exposures at which developmental toxicity is not expected. The exposures will be selected based on clear in vivo data in commonly used experimental animals or in humans. Internal effective concentrations in experimental animals will be compared to human data when available. In some instances, the data will be sufficient to use two concentrations of the same compound, one as a ''positive'' and one as a ''negative.'' In other instances, we may have confidence in the data for a compound only at one end of the dose–response curve.

There will be at least two ways to validate an assay using our list of ''gold standard'' developmental toxicity exposures. One method is to use a candidate alternative test to predict exposures as toxic or nontoxic and tally up the percent that match our list. This method is amenable to the application of Cooper statistics and to support decision-making based on a simple, dichotomous classification. A more quantitative approach is to use the candidate test to construct dose–response curves and calculate IC_{50} values or another suitable metric from the curve. This method can be used to rank potency of the exposures and to make other predictions about exposure level that can be checked against our list. In such a scenario, a candidate assay might be correct for 85% of the positives, but report 35% of the negatives as having toxicity within range of the named concentration at which no effect is expected. The high false-positive rate could then be the target of further improvement efforts.

CONSTRUCTION OF A ''GOLD STANDARD'' LIST OF EXPOSURES

How, then, would a list of such exposures be assembled? Because a positive (or negative) exposure must be specified in terms of an internal concentration of a compound, the list will be based on studies that link an internal dose metric (levels of active agent in maternal blood or the embryo) to an adverse outcome. The entries in the list can, therefore, only include compounds that have been subjected to mammalian developmental toxicity studies, or those for which there are human data, with adequate measures of internal dose. This requirement will likely bias the initial list toward pharmaceuticals, where the relationship between administered dose and consequent blood level is generally available. However, we hope that our validation list will prompt investigators to take some well-recognized teratogenic or nonteratogenic exposures involving nonpharmaceuticals and measure internal concentrations, which would increase the range of exposures that could be included on this list. We consider the list to be a work in progress because it will become more robust as more diverse exposures are added.

There are a number of potential pharmacokinetic parameters that could be used as the basis for setting a concentration to test in vitro, the most obvious being peak concentration and AUC (area under the time–concentration curve). We believe that the default choice should be peak concentration. Peak concentration is clearly the important parameter in modes of action involving interference with receptor function or inhibition of enzymes. Given that developmental toxicity can be produced by as little as a single exposure at a critical point in development, it is far more likely to be dependent on peak concentration. We recognize that AUC has been found to be more predictive than peak concentration for some chemicals; however, many alternative assays use static concentrations of test chemical, making AUC a simple product of peak concentration over assay time.

Although we believe that it is important to have actual pharmacokinetic data for the initial set of validation chemicals, it may eventually be possible to estimate internal concentrations based on physical chemical and or acute toxicological characteristics of the test agent. Retrospective comparisons of effective concentrations in vitro with in vivo results (benchmark concentration vs. benchmark dose) suggest that this approach may be reasonable if the in vivo study designs are of good quality (Janer et al., 2008). There have been attempts to use estimates of maximum achievable concentrations, estimated from acute toxicity data, to set upper limits for testing compounds in vitro that could be reapplied for this purpose (Daston et al., 1995).

A ''negative'' for our purposes is an exposure at the maternal Maximally Tolerated Dose that produces no adverse effects in the offspring, or a lower and ineffective dose of a compound that is toxic at higher concentrations. Because sample size and assay sensitivity influence identification of adverse effects at the putative NOAEL (No Observed Adverse Effect Level), we propose to use a benchmark dose approach to selecting a negative exposure. We believe that the lower bound of the 95% confidence interval around a 5% benchmark dose for malformations or embryolethality should provide a very reasonable estimate of a nontoxic exposure. The benchmark dose concept is discussed on the EPA web site (http://www.epa.gov/ncea/bmds/index.html).

Our list is intended to reflect whole animal or human exposure that can confidently be linked to developmental toxicity or lack of developmental toxicity. This list may need to be adjusted for individual alternative assays for which the concentrations we indicate cannot be directly tested. For example, it may be that a standard dilution of the concentrations we propose would be most appropriate for one particular alternative assay but impossible to administer for technical reasons in another assay. We anticipate that some exposures on our list would be selected by the developer of an assay as a calibration set and used to adjust assay conditions or interpretation. Once an assay is calibrated to a representative subset of the exposures on our list, the remainder could be used for validation.

ADVANTAGES OF THIS APPROACH

We see a number of advantages to this method:

A. Foremost, this approach is built on what we all, as working teratologists, know to be true: that developmental toxicity is conditional and depends on the exposure conditions as well as on the compound. Internal dose is a more consistent metric than administered dose, sidestepping to some extent interstudy differences in route or interspecies differences in metabolism.

B. By using only permanent effects as the basis for considering an exposure to be developmentally toxic, we emphasize effects that alter fetal organization. Thus, a decrease in fetal weight (which may or may not be due to disruption of organizational aspects of embryonic development and which may be transient) is not considered developmental toxicity for purposes of developing our list. However, permanent impairment of an organism's growth potential that has been demonstrated in experimental studies would be considered developmental toxicity. We fully recognize that for other purposes (e.g. risk assessment), decreased fetal weight or other fetal manifestations that may not affect the long-term structure, function, or viability of the offspring may be appropriate endpoints, but our intention is to use unequivocal organizational changes, particularly structural malformations, as endpoints so that our ''positive'' exposures are unequivocally positive.

C. Our approach does not use indices such as the ratio of doses that affect the adult to the dose that affects the developing organism (the A/D ratio) (Johnson et al., 1982). Although some investigators have found these indices to be useful, they have been shown not to be generalizable, at least not in any quantitative way (Daston et al., 1991). We believe a simpler approach based on identifying concentrations at which responses should be expected to be positive or negative provides a cleaner and more interpretable basis for evaluating alternative assays.

DISADVANTAGES OF THIS APPROACH

We recognize that our approach has some drawbacks.

A. It is not clear that our list of unequivocal positive and negative exposures will be more successful in avoiding some criticisms that were aimed at previous validation lists, particularly the selection of cytotoxic exposures as positives and excessively bland exposures as negatives. We plan to include as wide a range of chemicals among our gold standard exposures as possible, but we will be limited to those for which adequate in vivo mammalian developmental toxicity studies with internal dose information are available.
B. We are also aware that some test systems may have trouble using our specific concentrations in their individual assay conditions, hence our suggestion to use some of our exposures for calibration.
C. The advantage of our list in avoiding questions of maternal toxicity is also a disadvantage: if we inadvertently include an exposure for which maternal toxicity is the mechanism of developmental toxicity in vivo (e.g. uterine ischemia) we cannot expect an alternative test system to make the same call.
D. Inattention to active metabolites unless those metabolites have been well characterized is recognized as a limitation of any alternative assay system (Smith et al., 1983; Brown, 2002; Verwei et al., 2006). The burden here is larger and less certain when alternative assays are used for unknowns. Metabolizing systems have been incorporated into some assays in an attempt to overcome this problem. Ideally, our list will include only exposures for which we believe no further metabolism is needed for activity, but our knowledge is likely to be less than perfect in that regard.
E. A related issue is that of an unknown degree of protein binding. This approach shares this shortcoming with nearly all other in vitro approaches.

OPERATIONAL CONSIDERATIONS

As the process of developing a standard method to assess the reliability of alternative developmental toxicity assays on which we have embarked moves forward, we anticipate the following steps:

1. We will host a meeting of outside experts who will be asked to review and critique our plan. We look forward to a lively dialogue and many useful suggestions.
2. The creation of the list itself will largely be a literature search to identify developmental toxicity studies using common species (rat, mouse, and rabbit) in which a reliable measure of internal dose is available. To be selected, studies must include either clearly permanent adverse developmental outcomes (e.g. structural malformations or fetal death) or a lack of any developmental outcome (permanent or not) in a study of adequate size with dosing to the maternal Maximally Tolerated Dose. Studies should include sufficient information for BMD modeling.
3. Selection of ''positive'' and ''negative'' exposures. The selections will be reviewed by experts in the field. Our goal is to use unequivocal positives and negatives to provide a fair test of alternative systems.
4. There are a number of published validation exercises of alternative development toxicity assays, some of which will likely include exposures that are in our list. We will examine how those members of our list appear to have done in the previous studies.
5. We will publish our rationale, methods, and list in the open literature, again looking forward to many lively interactions and useful comments from the scientific community.

REFERENCES

Brown NA. 2002. Selection of test chemicals for the ECVAM International validation study on in vitro embryotoxicity tests. Altern Lab Anim 30:177–198.

Christian MS, Brent RL. 2001. Teratogen update: evaluation of the reproductive and developmental risks of caffeine. Teratology 64:51–78.

Daston GP, Rogers JM, Versteeg DJ, et al. 1991. Interspecies comparisons of A/D ratios: A/D ratios are not constant across species. Fundam Appl Toxicol 17:696–722.

Daston GP, Baines D, Elmore E, et al. 1995. Evaluation of chick embryo neural retinal cell culture as a screen for developmental toxicants. Fundam Appl Toxicol 26:203–210.

Genschow E, Spielmann H, Scholz G, et al. 2002. The ECVAM international validation study on in vitro embryotoxicity tests: results of the definitive phase and evaluation of prediction models. European Centre for the Validation of Alternative Methods. Altern Lab Anim 30:151–176.

Johnson EM, Gorman RM, Gabel BE, George ME. 1982. The hydra attenuata system for detection of teratogenic hazards. Teratog Carcinog Mutagen 2:263–276.

Janer G, Verhoef A, Gilsing HD, Piersma AH. 2008. Use of the rat postimplantation embryo culture to assess the embryotoxic potency within a chemical category and to identify toxic metabolites. Toxicol In Vitro 22:1797–1805.

Marx-Stoelting P, Adriaens E, Ahr HJ, et al. 2009. A review of the implementation of the embryonic stem cell test (EST). The report and recommendations of an ECVAM/ReProTect Workshop. Altern Lab Anim 37:313–328.

Nishimura M, Iizuka M, Iwaki S, Kast A. 1982. Repairability of drug-induced ''wavy ribs'' in rat offspring. Arzneimittelforschung 32: 1518–1522.

Schwetz BA. 1992. Criteria for judging the relative toxicity of chemicals from developmental toxicity data: a workshop summary. Teratology 45:337–339.

Schwetz BA, Harris MW. 1993. Developmental toxicology: status of the field and contribution of the National Toxicology Program. Environ Health Perspect 100:269–282.

Smith MK, Kimmel GL, Kochhar DM, et al. 1983. A selection of candidate compounds for in vitro teratogenesis test validation. Teratog Carcinog Mutagen 3:461–480.

Solomon HM, Wier PJ, Fish CJ, et al. 1997. Spontaneous and induced alterations in the cardiac membranous ventricular septum of fetal, weanling, and adult rats. Teratology 55:185–194.

Verwei M, van Burgsteden JA, Krul CAM, et al. 2006. Prediction of in vivo embryotoxic effect levels with a combination of in vitro studies and PBPK modeling. Toxicol Lett 165:79–87.

Webster WS, Brown-Woodman PDC, Ritchie HE. 1997. A review of the contribution of whole embryo culture to the determination of hazard and risk in teratogenicity testing. Int J Dev Biol 41:329–335.