Conference Report

Reproductive and Developmental Toxicity Testing: From *In Vivo* to *In Vitro*

Silver Spring, Maryland, April 16, 2012

On April 16, 2012, the U.S. Food and Drug Administration (FDA) held a workshop cosponsored by the National Institute of Environmental Health Sciences, the Center for Alternatives to Animal Testing (CAAT) at the Johns Hopkins Bloomberg School of Public Health, and the Middle Atlantic Reproduction and Teratology Association to discuss emerging *in vitro* tools for predicting reproductive and developmental toxicity. This workshop, with 350 registered participants including those participating via webstream, provided an opportunity to discuss the evidence needed to evaluate and validate new test methods and to integrate these methods into regulatory decision making.

During drug development, pharmaceutical companies use a variety of *in vivo* and *in vitro* developmental and reproductive toxicology (DART) tests to predict the safety of new compounds. The results inform internal decision making and labeling, helping companies determine, for example, whether a compound would be safe for particular populations (e.g., women of childbearing age). *In vivo* testing for effects on embryofetal development in two animal species – one rodent species (usually rats or mice) and one nonrodent species (typically rabbits) is generally required by FDA to support clinical trials and labeling for use in pregnancy, explained **Ed Fisher** (FDA). Currently, companies employ *in vitro* tests using human- or animal-derived cells or cell lines or different animal models (e.g., zebrafish) only as a way to rapidly screen compounds prior to, or in conjunction with, *in vivo* testing, added **Abigail Jacobs** (FDA).

For a variety of reasons, animal toxicity is sometimes a poor predictor of human toxicity, noted **David Gerhold** (National Institutes of Health – NIH) and other participants. Validated *in vitro* tests have the potential to increase the relevance of toxicity testing for humans and to reduce, refine, or replace *in vivo* animal testing. In fact, Gerhold continued, the National Research Council (NRC) report, *Toxicity Testing in the 21st Century: A Vision and Strategy* (NRC, 2007), envisions a future in which toxicity testing relies primarily on the *in vitro* study of human-derived cells or cell lines. Furthermore, the *in vivo* DART methods used for regulatory purposes have changed little in the past few decades, explained **Jesse Goodman** (FDA), despite dramatic advances in basic scientific research. Modernizing toxicology to improve preclinical predictions of product safety, continued Goodman, is one of the priorities identified in FDA's strategic plan for regulatory science (US FDA, 2011).

But how should scientists validate emerging *in vitro* assays or batteries of tests? And what is the current status of ongoing validation efforts? Participants addressed these and related questions, focusing on the ability of *in vitro* test methods to predict *in vivo* outcomes and the potential to incorporate these new methods into regulatory decision making. Ultimately, Fisher and other participants said, to be accepted by both pharmaceutical companies and regulators, new methods must provide protection that is equivalent to, or better than, existing *in vivo* approaches.

Traditional *in vivo* toxicity testing

For decades, information on the potential adverse effects of new drugs has come from animals, said Ed Fisher. The aims of *in vivo* embryofetal developmental toxicity testing are (a) to detect adverse effects on the pregnant female and on the development of the embryo and fetus (e.g., embryofetal death, altered growth, and structural changes) consequent to exposure from implantation to the closure of the hard palate, and (b) to extrapolate the results to humans using data on likely human exposures, comparative kinetics, and mechanisms of developmental toxicity. Protocols for *in vivo* DART studies are based on safety guidelines for reproductive toxicology developed by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use.

Animals are fairly good predictors of human reproductive toxicity and teratogenicity, said Fisher and **Jeffrey Bray** (FDA), especially when multiple species are used. Fisher argued that many instances of apparent discordance between human and animal response are caused by factors such as flaws in the animal study; low exposure in humans or a lack of significant exposure in women of childbearing potential; or insufficient data on large human populations with documented exposures. In addition, human and animal responses may look more discordant overall than they truly are because of the effectiveness of *in vivo* testing early in drug development. In other words, added Bray, animal testing may be effectively preventing many harmful compounds from being tested in humans, marketed, or prescribed for women of child-bearing potential. If this is the case, then data on adverse outcomes in humans would be available only for discordant teratogens or for those with subtle effects.

No animal species can perfectly predict the human response to drug compounds, continued Bray. Historically, rats and rabbits have been the species used most frequently for embryofetal developmental toxicity studies, and they are likely to remain the species of choice for some time. In large part, this is because extensive historical control data are available for these species. In addition, both rats and rabbits possess a suite of traits that make them good practical choices for research: short gestation, large litters, ease of breeding, low cost of housing, and ease of handling. However, explained Fisher and Bray, one also must consider similarities between the potential test species and humans in factors such as the maternal-placental-embryonic relationship; the metabolism and distribution of compounds, including transplacental transfer; pharmacodynamics (e.g., whether the target or pathway is present or relevant in that species); pharmacokinetics and metabolism; and maximum tolerated dose. Rats, rabbits, and other common species – including mice, guinea pigs, minipigs, dogs, and nonhuman primates – may differ widely in these factors.

Because rats and rabbits may not always be the most suitable species, noted Fisher and Bray, one should use a weight-of-evidence approach to guide the choice of species in DART testing. In terms of the 3Rs, choosing the most predictive or informative species could reduce the number of animals used or result in refined procedures. **Aldert Piersma** (the Dutch National Institute for Public Health and the Environment, or RIVM) suggested that the choice of species for *in vivo* testing would be aided by a database that compiles existing evidence on the most appropriate species for particular kinds of compounds. The creation of such a database would also allow an evaluation of the added value of the second species. Similar work for environmental chemicals previously led to the Organisation for Economic Co-operation and Development's guideline regarding an extended one-generation study in one species for chemicals.

Emerging *in vitro* methods and batteries

Scientists at pharmaceutical companies and research institutions are currently evaluating the ability of *in vitro* assays to predict *in vivo* developmental toxicity. The initial goal, explained **Diane Umbenhauer** (Merck) and **Robert Chapin** (Pfizer), is to provide internal guidance regarding the need for more definitive animal studies. If validated sufficiently, such tests could ultimately be used for regulatory purposes.

Whole embryo culture and embryonic stem cell tests

Umbenhauer described Merck's efforts to validate three types of assays – a rat whole embryo culture (WEC) assay, mouse embryonic stem cell test (EST), and human EST – and to assess whether these tests could be used in a complementary manner. Both the rat WEC assay and the mouse EST were originally evaluated by the European Center for the Validation of Alternative Methods (ECVAM). In its validation of the rat WEC assay, ECVAM found that the assay predicted *in vivo* toxicity with 48%-80% accuracy, depending on the prediction model used. Merck tested 57 compounds, including a 30-compound training set and a 27-compound test set. Researchers developed and tested a proportional odds model (POM), using the endpoint $IC_{MAL,\ 50}$ (the concentration at which 50% of the embryos showed malformations). For each compound, the POM produces a probability that it is a nonteratogen, weak teratogen, or strong teratogen. The WEC POM algorithm correctly identified 14 out of 15 strong teratogens (93% sensitivity, 100% specificity), but was less robust at discriminating nonteratogens from weak teratogens. In addition, the assay appears not to be sensitive to certain classes of weak teratogens – for example, the algorithm incorrectly classified nonsteroidal anti-inflammatory drugs as nonteratogens.

The mouse EST, Umbenhauer explained, is composed of three assays: cytotoxicity of the "adult cell line" mouse 3T3 fibroblasts, cytotoxicity of the mouse stem cell D3 line, and differentiation of mouse stem cells (from the D3 line) into cardiomyocytes (measured as the number of wells in a tissue culture plate that contain spontaneously beating cells). In its evaluation of the mouse EST, ECVAM found that the assay was able to predict *in vivo* developmental toxicity with 78% accuracy. However, in Merck's validation with 58 compounds, the assay's predictivity was much lower, around 60%; the rate of false negatives was 33% (7 out of 21) and the rate of false positives was 50% (8 out of 16). Further, the assay misclassified a known strong teratogen as a nonteratogen. To improve the assay, Merck scientists incorporated mRNA and miRNA expression profiling and streamlined the experimental process, assessing endpoints (cytotoxicity and total RNA) on day 5 instead of day 10. Among other advantages over the 10-day EST, the 5-day EST is based solely on molecular endpoints rather than cardiomyocyte markers (such as beating cell counts). Umbenhauer described a preliminary set of predictive genes (a combination of mRNA and miRNA) whose changes in expression best discriminate nonteratogens from weak or strong teratogens. Based on 37 compounds tested to date with the 5-day mouse EST, Merck has found an overall accuracy of 89%; the rate of false negatives was 5% (1 out of 21) and the rate of false positives was 19% (3 out of 16). Merck is now testing additional compounds to confirm this assay's performance. Merck is simultaneously developing a human EST using the H9 line, Umbenhauer added.

Zebrafish model for developmental toxicity screening

The goal of the Zebrafish Teratogenicity Consortium, said **Belén Tornesi** (Abbot Labs), is to establish a harmonized zebrafish developmental toxicity assay rigorously tested for performance. The benefits of the zebrafish model include low husbandry and assay costs; the ability to culture embryos and larvae; rapid organogenesis; transparent embryos, which allows one to assess organ development less invasively than in many other species; developmental processes and pathways that are similar to those of other species; and a genotype and phenotype that have been extensively evaluated with regard to teratogenic mechanisms. Disadvantages to this model include the lack of a placental barrier and compound uptake.

The consortium evaluated 20 nonproprietary, chemically diverse compounds in four laboratories, explained Tornesi. After 5 days of dosing, researchers scored embryos according to morphological endpoints and growth measurements. They classified compounds as teratogens or nonteratogens using the LC_{25}/NOAEC ratio (i.e., the ratio between the concentration at which 25% of the test organisms die and the highest concentration at which no observable adverse effects occur). They found that the assay's ability to predict mammalian-based teratogenicity was only 60%-70%. The low predictivity may be due to the bioaccumulation in the embryo of some compounds and the embryo's poor uptake of others, suggested Tornesi. The consortium then modified the protocol in a number of ways, such as by lowering the highest concentration, adjusting the concentration range in cases with steep lethality curves, and making pH adjustments when necessary. When they retested the same set of 20 compounds using the optimized protocol in a single laboratory, the assay's predictivity rose to 85% (Gustafson et al., 2012). The consortium has concluded that the zebrafish model shows promise as an alternative testing method for developmental toxicology assessment. Among its next steps, said Tornesi, the consortium hopes to overcome compound solubility and uptake issues by using microinjections to dose embryos directly.

EST and zebrafish assay as part of an aggregational, gestalt-generation process

In considering a compound's potential for teratogenicity, argued Chapin, one should begin by asking two questions. First, does the target gene appear to be important in development (e.g., is it highly expressed in the embryo or placenta)? The answer to this question affects the type of assay best suited to predicting target-mediated toxicity. For example, when the target gene is primarily expressed in the placenta, the compound might not be flagged as problematic via either zebrafish or stem cell-based assays. Second, is the compound cytotoxic at very low concentrations? Simple cytotoxicity, Chapin emphasized, can provide valuable information that is often overlooked; therefore, any *in vitro* estimation of toxicity should include a measure of cytotoxicity.

Chapin described Pfizer's work merging an EST, a zebrafish assay, and other information to predict toxicity *in vivo*. Pfizer adopted the ECVAM-validated EST and added gene expression to it, then refined the model in a number of ways (e.g., by removing correlated genes and considering several different statistical models). Currently, Pfizer uses a random forest (v3) model; for each compound, the model generates probabilities that it poses a low, moderate, or high risk *in vivo*. This model has very good predictivity for compounds with low and moderate risk *in vivo*, but its predictivity is not as good for compounds with a high risk *in vivo*. Pfizer uses the zebrafish assay as well as the EST because malformations result from altered cell migrations, which are captured by the zebrafish assay but not the EST. Drug development teams must carefully weigh (a) the expression of the target gene; (b) the results of the EST (the likelihood that the compound is a developmental toxicant, based on various statistical models); (c) the results of the zebrafish assay (the LC_{25}/NOAEC ratio and the presence of malformations); (d) values indicative of cytotoxicity, such as the IC_{50}; and (e) information from DEREK, a knowledge-based expert system for the qualitative prediction of toxicity. Unfortunately, the signals are often mixed. In interpreting results of the zebrafish assay in particular, Chapin added, scientists must ensure that the fish were actually exposed to the compound. In addition, Chapin advocates placing more weight on probability outputs for the random forest model that are greater than about 80%, as such probabilities tend to be more valid.

Metabolomics as a high-throughput screen

In a poster presented at the workshop, **Helena Hogberg** (Johns Hopkins University) and colleagues described an *in vitro* approach using metabolomics and transcriptomics to identify pathways of developmental neurotoxicity (Hogberg et al., 2012). They exposed a three-dimensional rat primary neuronal organotypic model to three concentrations of lead chloride or control (untreated). They quantitatively measured genes expressed in different cell types and performed mass spectrometry-based metabolomics measurements, then looked for associations between the metabolites that changed after lead chloride exposure and changes in gene expression. In particular, oleamide, a metabolite that increased with exposure to lead chloride, is hydrolyzed by the enzyme FAAH, which was downregulated at the mRNA level in the exposed samples. In addition, several metabolites in the pathway of the neuronal-specific metabolite NAA decreased significantly after lead chloride exposure. (In humans, decreased NAA levels have been associated with neuronal or axonal loss and compromised neuronal metabolism.) The mRNA levels of MBP and NF-200 were significantly downregulated after lead chloride exposure, indicating neuronal or axonal loss. Hogberg et al. concluded that omics approaches can contribute to the identification of pathways of toxicity.

Optimizing *in vitro* assays

The importance of dose

The safety of most compounds for humans depends on many factors, including the dose, or concentration. Therefore, said **George Daston** (Proctor and Gamble), the dose at which a compound has adverse effects – in an *in vivo* or *in vitro* assay – is essential to the interpretation of whether that compound will be problematic in

humans. With *in vivo* developmental toxicity studies, one always evaluates dose-response, determining the effects of a compound, or lack thereof, at different doses. This allows one to compare the dose-response in animals with likely human exposure to determine the risk of a particular compound or drug for humans. In contrast, when designing *in vitro* assays or interpreting their results, researchers tend to neglect concentration.

To address this problem, Daston argued for exposure-based validation of *in vitro* assays. In a conventional validation, researchers select a list of compounds and classify them into "positive" and "negative" developmental toxicants based on *in vivo* data. In an exposure-based validation, researchers use pharmacokinetic information about each compound to define "positive" as a particular chemical at the exposure at which that chemical is active (expected to produce an effect *in vivo*) and "negative" as a particular chemical at the exposure at which it is inactive (not expected to produce an effect *in vivo*). Using this approach, a chemical can be its own control. In one exposure-based validation of a chick embryo neural retina cells assay, Daston and colleagues found, using conventional criteria, an accuracy of 82%; using exposure-based criteria, accuracy rose to 93% (Daston et al., 1995). Daston suggested that published data from other *in vitro* assay validations should be reanalyzed using exposure-based criteria; such a reanalysis might find increased concordance between *in vitro* and *in vivo* results.

Predictive models and the importance of rigor

To reliably use the results of *in vitro* assays to predict the safety of new compounds, explained **Kjell Johnson** (Arbor Analytics), one must carefully develop and evaluate predictive models. The concept is simple: one builds a predictive model using existing data, and then uses the model to predict the safety of new compounds. In practice, however, building and interpreting such models can be complicated by data limitations, the trade-offs inherent in choosing a model, and the interpretation of results (Kuhn and Johnson, in press).

Ideally, the data for predictive models would come from a controlled experiment with many samples that are balanced across the response (i.e., a balance of compounds with low, moderate, and high risk of teratogenicity) and with more samples than predictors, said Johnson. But real data are typically messier: often they are not derived from controlled experiments, predictors (e.g., genes) generally outnumber samples (compounds), predictors are measured with error, and the data set tends to include missing values. Some of these problems can be overcome. For example, one should design experiments with predictive modeling in mind to ensure the collection of appropriate data. Further, collaboration among companies and researchers can increase the number of samples.

The typical approach to building a predictive model, Johnson explained, is to randomly split the original data into a training set and a test set; one builds the model using the training set, then evaluates model performance using the test set. However, such an approach will not work with models that have tuning parameters, such as partial least squares, neural networks, random forests, k-nearest neighbors, and naïve Bayes models. To determine the optimal value of a tuning parameter for such a model, one may still split the data into training and test sets, but one then selects the tuning parameter values and, for each parameter value, re-samples the data using a bootstrapping approach and fits the model. One evaluates the model many times for each value of the tuning parameter, calculates the resampling performance, determines the best parameter values, and fits the model to all training data. Then one applies the final model to the test data and evaluates its performance. In some cases, one can use the original data rather than splitting it into training and test sets; this can be helpful when one is working with a small number of compounds because splitting the data results in a loss of information.

The choice of model, said Johnson, depends on the data. Models vary in their ability to handle data issues such as missing data or descriptors that outnumber the samples; they also vary in speed, performance (accuracy), interpretability, and robustness. No single model will perform the best across a range of problems; therefore, one should build several different kinds of models for every problem and use a resampling approach to identify the optimal model. Over time, one should revisit and rebuild the models as more information becomes available. Regardless of the problem or the model, cautioned Johnson, one must avoid extrapolating beyond the range of data on which the model was built.

Status of collaborative efforts to develop and validate *in vitro* assays

Collaborations among researchers from pharmaceutical companies, federal agencies, and academic institutions provide forums for the exchange of data and ideas; such efforts also can promote and facilitate the rigorous testing of *in vitro* methods and the dissemination of results. Some workshop participants, including Tornesi and Daston, are members of such collaborative efforts. Participants provided updates of several large-scale collaborative initiatives related to the development and validation of *in vitro* assays.

Tox21 consortium

In response to the NRC's report, *Toxicity Testing in the 21st Century: A Vision and a Strategy*, said Gerhold, the U.S. Environmental Protection Agency (EPA), FDA, and NIH (the National Toxicology Program and the NIH Chemical Genomics Center) initiated the Tox21 Consortium in 2008. This unique partnership also coordinates with a number of other groups and initiatives, such as the E.U. Joint Research Centre and the Interagency Coordinating Committee on the Validation of Alternative Methods. Using quantitative high-throughput screening methods, Tox21 is prioritizing compounds for more extensive toxicological evaluation. The consortium also aims to identify mechanisms of compound-induced biological activity as a way to characterize toxicity and disease pathways, facilitate cross-species extrapolation, and provide input to models for low-dose extrapolation. Ultimately, Tox21 intends to develop (a) the infrastructure to support the basic and applied research needed to develop tests and pathway models and to make all data and results available to the scien-

tific community; (b) a comprehensive suite of *in vitro* tests, based primarily on human-derived cells, cell lines, or components; (c) computational models of toxicity pathways to support the application of *in vitro* test results in hazard characterization and risk assessment; (d) targeted animal tests to complement *in vitro* tests; (e) appropriate validation of tests and test strategies; and (f) evidence that the toxicity pathway approach is adequately predictive of adverse health outcomes to use it in decision making.

Tox21 screens compounds at multiple concentrations and generates robust activity profiles for all compounds with low rates of false positives and false negatives, Gerhold noted. The majority of assays are target-specific, focusing on cellular stress responses and nuclear receptor responses; other assays evaluate phenotypic endpoints (e.g., cytotoxicity and apoptosis), cell signaling, drug metabolism, and genetic variation. In Phase 1, Tox21 screened approximately 2,800 compounds (additional compounds, mostly pesticides, have been screened through EPA's separate ToxCast program). Currently in Phase 2, Tox21 is screening a library of more than 10,000 compounds (including industrial chemicals, pesticides, and food and drug components), a process now facilitated by a dedicated toxicology robot. Tox21 partners are selecting assays based on Phase 1 experience, information from *in vivo* toxicological investigations, the advice of basic researchers, and maps of disease-associated pathways. Researchers will begin with a focus on receptor activation or inhibition and the induction of stress response pathways; later, they will address other disease-associated pathways and move to high-throughput gene array assays. Tox21 is constructing predictive models based on assay results that correlate with developmental toxicities in rodent or human studies.

Human on a Chip

NIH is collaborating with the Defense Advanced Research Projects Agency, with guidance from FDA, on a project called "Human on a Chip." This project, explained Gerhold, aims to develop a tissue chip that mimics human physiology; it will serve as an extremely efficient preclinical screen for safe and effective drugs.

ILSI–HESI DART Technical Committee

The International Life Sciences Institute (ILSI) – Health and Environmental Sciences Institute (HESI) DART Technical Committee, which includes representatives from 33 companies, universities, and federal agencies, aims to advance DART research and develop consensus on the appropriate use of experimental toxicity data for human risk assessment, said **Bruce Beyer** (Sanofi). The committee is addressing challenges in the areas of animal use and welfare, alternatives to animal models, stem cell technology, risk assessment for sensitive or vulnerable populations, improved testing and assessment strategies, and improved biomonitoring through biomarkers. Beyer provided an update of four ongoing projects of relevance to this workshop.

– Alternative developmental toxicity assays: Committee members are exploring the utility of currently available zebrafish and stem cell-based assays as developmental toxicity screens. The zebrafish research is expected to be published in 2012 and presented at the 2012 Teratology Society meeting through the zebrafish consortium effects that Tornesi presented. The stem cell work, which will be based on a gap analysis of current data sets, may not be completed until 2013. Ultimately, explained Beyer, the committee hopes to help determine whether nonmammalian assays can be used to postpone or remove the requirement for a second mammalian species in developmental toxicity testing.
– Consensus list of developmental toxicants: In this ongoing project, the committee has identified an exposure-based validation list for developmental toxicants, using the procedure described separately by Daston (2010). The proposed list of exposures (each of which is a chemical-concentration combination) should provide a definitive list of developmental toxicant concentrations to be used by assay developers and regulators to validate alternative assays, said Beyer and Daston.
– Rodent vs. non-rodent second species: The relative value of rodents vs. non-rodents in developmental toxicity signal detection and the influence on human risk assessment is unknown, said Beyer. In this project, the committee is conducting a survey of pharmaceutical companies to collect data on (a) the strength of the developmental toxicity signal in each species, (b) the putative safety margin against human therapeutic dose and exposure in each species, and (c) the pharmacologic relevance of each species. This project should help gauge the potential risks of collecting embryofetal developmental data from only one species prior to Phase 3 trials and specific circumstances in which a second species does or does not add value.
– Testicular toxicity: This project, which should be completed by September 2012, is expected to raise awareness of the need for *in vitro* testis models and to promote research on this topic, Beyer continued. A joint workshop with CAAT in fall 2011 furthered these discussions. The committee published a survey manuscript on this area of research in 2011 (Sasaki et al., 2011).

IQ preclinical safety leadership group

The International Consortium for Innovation and Quality in Pharmaceutical Development (IQ), a pharmaceutical initiative with 28 member companies, formed the Preclinical Safety Leadership Group (PSLG) in 2010, said **Maryellen Mcnerney** (Bristol-Myers Squibb). A working group was convened by the PSLG to assess the ability of *in vitro* teratogenicity assays to predict the findings of *in vivo* embryofetal developmental studies. This working group has designed a survey requesting all compound assessments from member companies for which data, including proprietary data, from both *in vivo* and *in vitro* assays are available. This will be the largest survey of its kind to date, asserted Mcnerney. Surveyed companies will be asked to identify unique aspects of their *in vitro* assay designs. The working group will distribute the survey and analyze responses in 2012 and will hold a face-to-face meeting in early 2013 to review the results and discuss the next steps.

CAAT initiatives

A poster presented at the workshop described CAAT's ongoing activities (CAAT, 2012). The center's collaborative efforts include the transatlantic think tank for toxicology (t^4), a collabo-

ration with the Doerenkamp-Zbinden Foundation that aims to promote the efficient implementation of the recommendations of *Toxicity Testing in the 21st Century*. The members of t4 include leaders in the fields of evidence-based toxicology and alternatives. In addition, CAAT serves as the secretariat of the Evidence-Based Toxicology Collaboration, which aims to foster the development of a process, based on evidence-based medicine, for the quality assurance of new toxicity tests for the assessment of safety in humans and the environment. CAAT is also the secretariat of the Refinement Working Group, sponsored by the Klingenstein Foundation, which brings together top industry and academic researchers to develop new approaches to refinement, especially for pharmaceutical development and testing.

Application of *in vitro* assays to regulatory decision making

Can we forgo or delay in vivo *testing in a second species today?*

Participants considered a number of issues and questions related to the use of *in vitro* assays for regulatory decision making. In particular, participants focused on whether regulators and pharmaceutical companies would be willing to rely on *in vivo* testing in a single species coupled with a battery of *in vitro* tests, forgoing *in vivo* testing in a second species. Piersma proposed that a reasonable approach would begin with *in vitro* assays capable of catching most dangerous compounds, such that only the truly promising compounds would be run through *in vivo* assays. One would use a second species only if significant uncertainty about the compound's safety remains. Ed Fisher countered that such an approach may work for chemicals such as pesticides, but for drugs the goal is very different – one needs to characterize the drug's effects for labeling purposes.

Several participants – including Ed Fisher, **Ben Fisher** (FDA), and Mcnerney – argued that, except in limited circumstances, it is too early to replace *in vivo* testing in a second species with an *in vitro* battery. Evidence is not yet sufficient to be certain that currently available *in vitro* models capture all potential mechanisms, continued Ed Fisher. Bray agreed that the use of two *in vivo* species should remain the default for DART testing for now, but he argued for a weight-of-evidence approach. Participants proposed a number of specific circumstances in which it might be acceptable to forgo *in vivo* testing in a second species.

- For biologics, said Bray, only nonhuman primates are relevant, so testing in a single species is already acceptable for such products.
- For aromatase inhibitors, suggested Bray, *in vivo* DART testing is not necessary at all because previous evidence has clearly established that such compounds are deleterious to the mammalian embryo, often at doses that are very low compared to human exposures.
- **Amy Ellis** (FDA) noted that some companies test antimicrobials in rabbits, even though rabbits are poorly suited to the study of this class of compounds. In such cases, a combination of *in vitro* assays and testing in one *in vivo* species might be more informative.
- Mcnerney noted that pharmaceutical companies sometimes use two species even when one would predict, by virtue of the mechanism of action, that a substance is teratogenic. She suggested that, in such cases, forgoing the second species might be valid.

Rather than forgoing the second species, some participants proposed that it might be more acceptable to *delay* testing in a second species until after clinical trials. Such a strategy might reduce the number of animals used because some compounds will fail during clinical trials, obviating the need for further *in vivo* testing. One participant suggested that postponing testing in a second species might be scientifically relevant for small molecules; for such compounds, one might not be able to identify the metabolites in humans or determine the therapeutic concentration prior to clinical trials. Bray and other participants noted that the patient population and other factors should be considered. For example, this approach might not be suitable for compounds intended primarily or only for women of childbearing potential.

Companies wishing to use *in vitro* assays to replace or delay *in vivo* testing in a second species should clearly articulate the rationale and support for the proposed replacement, advised Bray; FDA reviewers should be receptive and should determine whether the replacement makes sense.

What will it take to begin replacing some in vivo *testing with* in vitro *assays?*

For *in vitro* assays to become an acceptable replacement for some animal testing, several participants said, researchers and regulators must (a) further refine *in vitro* tests and batteries, (b) gain a greater understanding of mechanisms and toxicity pathways, and (c) develop a strategy for determining the circumstances under which *in vitro* assays may be an appropriate substitute for animal testing.

Thomas Hartung (CAAT) argued that some of the *in vitro* assays discussed at the workshop were introduced more than four decades ago and have remained largely unchanged. Among the shortcomings of current *in vitro* tests, **Deborah Hansen** (FDA) pointed out, most currently available assays consider only the early stages of development; she urged researchers to develop assays that consider later points in development. As they endeavor to improve *in vitro* toxicity testing, **Jennifer Sasaki** (Alkermes) noted, researchers must determine whether to focus on animal-derived cells or cell lines, for which a great deal of *in vivo* data are available, or human-derived materials, which should be more relevant. Hartung urged a greater focus on integrated batteries of tests rather than the standalone *in vitro* assays that constituted the primary focus of this workshop. An assessment of the true value of *in vivo* testing – for example, by the Evidence-Based Toxicology process – would also be useful, Hartung said.

Tom Flynn (FDA) noted that many of the questions raised at this workshop were also discussed 32 years ago at the first U.S. conference on alternative assays for teratogenicity; most of the questions remain unanswered. **Catherine Willett** (Hu-

mane Society of the United States) agreed that the application of *in vitro* assays in the pharmaceutical industry has remained unchanged for decades. To make progress, Flynn argued, the scientific community would have to indicate its acceptance of the goal of using *in vitro* assays to replace some animal testing, rather than only as prescreens. Willett proposed that the goal should be to characterize more thoroughly the chemical-biological activities that underlie the adverse outcome. This kind of information will help convert *in vitro* assays from prescreening tools into a means to predict the compound's effect in the whole organism. The adverse outcome pathway paradigm is one way to organize information and to quantitatively link mechanisms to outcomes; this paradigm could be part of a long-term strategy for improving the utility and applicability of *in vitro* tests. In the short term, Willett added, the scientific community should establish databases compiling relevant information. Piersma agreed that an adverse outcome pathway approach might be informative. Mcnerney noted, however, that the adverse outcome pathway paradigm does not always reflect what happens when one actually administers the drug.

As a strategy for weighing the need for testing in a second species, Gerhold proposed that one could categorize compounds as high-risk or low-risk for teratogenicity, based on what is known about the drug, whether it will be taken short-term or chronically, and the intended patient population. Separate rules would apply to low-risk vs. high-risk drugs: regulators would continue to require *in vivo* testing in two species for high-risk drugs but could allow companies to forgo or delay a second species for low-risk drugs.

Suzanne Fitzpatrick (FDA) and Bray said that having *in vivo* data from pharmaceutical companies on failed compounds – and the compounds themselves – would be quite valuable to advancing the regulatory use of *in vitro* tests. FDA scientists could then conduct *in vitro* tests on those same compounds; a comparison of the *in vivo* and *in vitro* results would demonstrate the true predictivity of *in vitro* assays. Toxicokinetic data would be especially useful, added Bray. Chapin suggested that a safe harbor agreement might encourage pharmaceutical companies to provide *in vivo* data for failed compounds.

Conclusion

The opportunity and imperative now exist to transform preclinical toxicology from an empirical animal-based exercise to a predictive mechanism-based science, asserted Gerhold. Fitzpatrick said that FDA is assessing new ways to develop more predictive models to determine product safety. This workshop, she said, will be one of many opportunities to discuss collaboration in the development of new and exciting assays.

References

CAAT (2012). Center for Alternatives to Animal Testing: Innovation, education, and implementation of 21st century humane science. Poster presented at the workshop Reproductive and Developmental Toxicology Testing: from in Vivo to in Vitro, Silver Spring, MD, April 16.

Daston, G. P., Baines, D., Elmore, E., et al. (1995). Evaluation of chick embryo neural retina cell culture as a screen for developmental toxicants. *Toxicol Sci 26*, 203-210.

Daston, G. P., Chapin, R. E. Scialli, A. R., et al. (2010). A different approach to validating screening assays for developmental toxicity. *Birth Defects Res Part B 89*, 526-530.

Gustafson, A.-L., Stedman, D. B., Ball, J., et al. (2012). Interlaboratory assessment of a harmonized zebrafish developmental toxicology assay – Progress report on phase I. *Reprod Toxicol 33*, 155-164.

Hogberg, H. T., Welles, H., Zhao, L., et al. (2012). DNTox-21c Identification of pathways of developmental neurotoxicity for high-throughput testing by metabolomics. Poster presented at the workshop Reproductive and Developmental Toxicology Testing: from in Vivo to in Vitro, Silver Spring, MD, April 16.

Kuhn, M. and Johnson, K. (in press). *Applied Predictive Modeling: A Step-by-Step Guide*.

National Research Council (2007). *Toxicity Testing in the 21st Century: A Vision and Strategy*. Washington, DC: National Academies Press.

Sasaki, J. C., Chapin, R. E., Hall, D. G., et al. (2011). Incidence and nature of testicular toxicity findings in pharmaceutical development. *Birth Defects Res Part B 92*, 511-525.

US FDA – U.S. Food and Drug Administration (2011). *Advancing Regulatory Science at FDA*. Silver Spring, MD: U.S. Food and Drug Administration.

Elizabeth Stallman Brown, Abigail Jacobs*, and Suzanne Fitzpatrick*

* The views presented in this paper represent those of the individual scientists and not necessarily their respective organizations