 Birth Defects Research (Part B) 92:395–403 (2011)

Review Article

Critical Evaluation of Current Developmental Toxicity Testing Strategies: A Case of Babies and Their Bathwater

Edward W. Carney,[1]* Amy L. Ellis,[2] Rochelle W. Tyl,[3] Paul M.D. Foster,[4] Anthony R. Scialli,[5] Kary Thompson,[6] and James Kim[7]

[1]Toxicology & Environmental Research and Consulting, The Dow Chemical Company, Midland, Michigan
[2]US FDA, Center for Drug Evaluation and Research, Silver Spring, Maryland
[3]Discovery and Analytical Sciences, RTI International, Research Triangle Park, North Carolina
[4]National Institute of Environmental Health Sciences—National Toxicology Program, Research Triangle Park, North Carolina
[5]Tetra Tech Sciences, Arlington, Virginia
[6]Reproductive Toxicology, Bristol Myers Squibb Company, New Brunswick, New Jersey
[7]ILSI Health and Environmental Sciences Institute (HESI), Washington, District of Columbia

This review is the second in a series of four papers emanating from a workshop entitled ''Developmental Toxicology—New Directions,'' which was sponsored by the ILSI Health and Environmental Sciences Institute's (HESI) Developmental and Reproductive Toxicology Technical Committee. The present review analyzes the strengths and weaknesses of current developmental safety testing approaches in an effort to identify those strengths that should be retained in the future versus the weaknesses that should be eliminated. Workshop participants considered the following to be key strengths of current testing approaches: the integrated biology of pregnant animal models including pharmacokinetic and pharmacodynamic processes, the ability to detect low incidence malformations as well as maternally mediated toxicity, and the long history of use coupled with extensive historical data. A number of weaknesses were related to the resource-intensive nature of developmental toxicity testing (e.g., large number of animals, high costs, low throughput, the inability to keep pace with the demand for more toxicity data). Other weaknesses included the use of very high dose levels that often far exceed human exposure levels, the confounding influence of maternal toxicity, sparse understanding of basic developmental mechanisms and genetics of standard animal models relative to mouse or lower organisms, difficulties interpreting low incidence findings, and issues surrounding the interpretation of minor skeletal variations. An appreciation of these strengths and weaknesses is critical for the design of new approaches to developmental toxicity testing in the 21st century. *Birth Defects Res (Part B)* 92:395–403, 2011. © 2011 Wiley Periodicals, Inc.

Key words: ***safety testing; maternal toxicity; pharmacokinetics; alternatives to animal testing***

INTRODUCTION

After a half-century of toxicity safety testing and a decade into the 21st century, we find ourselves in the midst of much spirited and often contentious discussion on the need for a complete makeover of toxicology safety assessment. This discussion has been catalyzed by the 2007 NAS report entitled *Toxicity Testing in the 21*st *Century: A Vision and a Strategy* (Andersen and Krewski, 2009) which envisions a not too distant future which harnesses the might of modern technology, such as high-throughput robotically controlled in vitro or cell-free assays, genomics and powerful computational biology tools, to replace or greatly reduce whole animal toxicity tests. Developmental toxicology has not been immune to this discussion. Developmental toxicity testing, which involves dosing of pregnant rats and rabbits and evaluation of fetal outcome, has changed very little since its inception. These tests have served as the principal basis for countless public health decisions regarding pharmaceuticals, pesticides, industrial chemicals, and other chemical entities. Although the incredible longevity of this system is a tribute to its designers, it is nonetheless clear that developmental toxicology must continually adapt if it is to remain vibrant and useful into the future.

The call for change is being driven by a number of factors, among them the technological explosion in systems biology, computational biology, and high-throughput assay development, as well as corresponding improvements in

*Correspondence to: Edward W. Carney, The Dow Chemical Company, 1803 Building, Midland, MI 48674.
E-mail: ecarney@dow.com
Received 17 May 2011; Accepted 3 June 2011
Published online in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/bdrb.20318

Table 1
Characteristics of Developmental Toxicity Testing: Current versus Desired Future

Current	Desired future
Pregnant animal models (predominantly rats and rabbits, some nonhuman primates)	Human cells in vitro, lower organisms, in silico models
Focus on identification of developmental effects (including low incidence malformations)	Focus on perturbation of cellular/molecular toxicity pathways
High cost per compound	Low cost per compound
Low throughput—only a small percentage of the chemical universe can be evaluated	High throughput—evaluate thousands of chemicals, as well as mixtures
Administer maximum tolerated doses	Test at concentrations more relevant to human exposure

our fundamental understanding of normal development and its control processes (Table 1). Societal pressures are also impinging on developmental toxicology, specifically with respect to the large numbers of animals used in testing, as well as the high costs and time required. Also, testing resources are concentrated on a relatively small proportion of the universe of chemical entities, while regulatory initiatives such the European Union's REACH program (Grindon et al., 2006) call for comprehensive toxicity data on thousands of chemicals, not to mention additional concerns about chemical mixtures.

Although we all can agree that developmental toxicology must embrace 21st century science and respond to 21st century societal demands, how we should implement the vision is not quite as obvious. Playing on the old adage, ''don't throw the baby out with the bathwater,'' this review is based on the proceedings of an ILSI-HESI workshop entitled ''New Directions in Developmental Toxicology.'' The present review summarizes a portion of the workshop which sought to objectively identify the strengths and weaknesses of current developmental safety testing in an effort to identify those strengths that should be retained (''the baby'') versus the weaknesses that should be relinquished (''the bathwater''). Such an analysis was then be used to guide subsequent discussions at the workshop on the design and implementation of developmental toxicity testing in the 21st century (see companion manuscripts).

EVALUATION OF CURRENT DEVELOPMENTAL TOXICITY TESTING

Time, Costs, Animal Use, and Testing Capacity

In the early 1960s, the original designers of developmental toxicity testing protocols presumably were more concerned that another ''thalidomide'' did not slip through undetected than they were about time, costs and animal use. To ensure detection of teratogenic agents they specified testing in two species (mainly rats and rabbits) and the use of large sample sizes (Table 2). When one considers preliminary dose-finding studies which are also conducted in pregnant animals of both species, the number of animals used to test just one compound

Table 2
Animal Use in Developmental Toxicity Studies

Study	No. adults[a]	No. fetuses[b]
Rat prenatal developmental toxicity probe	25	308
Rat prenatal developmental tox	100	1,260
Rabbit prenatal developmental toxicity probe	25	176
Rabbit prenatal developmental toxicity	100	720
Total	250	2,464

[a]Typical starting number assigned to yield at least 20 litters/group.
[b]Based on 90% pregnancy rate and mean litter sizes of 14 (rats) and 8 (rabbits).

amounts to about 250 adult females and almost 2500 fetuses. This constitutes approximately 22% of the total number of animals used in a standard mammalian toxicology testing package as required to register a drug or pesticide and does not include reproductive or juvenile toxicity studies, which constitute another 40% or more of the animals used in a standard mammalian toxicity testing package (Van der Jagt et al., 2004; Rovida and Hartung, 2009).

Developmental toxicity tests also take several months to complete due to the length of the gestation period and the additional time needed to process and conduct detailed evaluations of the fetal skeleton. These tests are labor intensive, particularly in the areas of fetal visceral and skeletal evaluation and analysis of fetal morphological data. Costs for a set of rat and rabbit developmental toxicity dose-finding and main studies can easily exceed $300,000 for a single compound. Developmental toxicity studies in nonhuman primates can cost more than one million dollars per study.

Laboratories that routinely conduct regulatory developmental toxicity studies are relatively few in number, being mainly confined to the large pharmaceutical and chemical company safety assessment laboratories and the larger contract research organizations. All of these factors dictate total testing capacity, which is clearly rate-limiting relative to the number of chemical entities in existence. It has been estimated that the EU REACH program alone will require an additional 14,000 developmental toxicity tests (Scialli, 2008; Rovida and Hartung, 2009), and this is just for industrial chemicals. The number of tests required under REACH could be as high or higher than the total number of developmental toxicity tests conducted on industrial chemicals over the last 50 years!

The issues of animal use, cost, and low-throughput rate are among the most problematic for the future of conventional developmental toxicity testing and clearly need to be reconsidered in terms of future sustainability. The proposed 21st century paradigms promise to solve most of these issues, offering little to no animal use, low cost, and very high throughput. However, these new approaches must be validated and applied carefully lest they generate a great deal of additional unnecessary research in animals if the results of high-throughput screens are inconclusive or generate an excess of false positives and/or false negatives. Also, several of the newer high-throughput assays utilize vertebrate whole

animal models such as zebrafish. Although a move toward these models may reduce the number of mammals used in testing, the total number of vertebrate animals could actually increase if such models come into widespread use!

Integrated Biology of Mammalian Whole Animal Models

The great strength of mammalian whole animal models comes from their highly complex, integrated biology. This includes a complete range of pharmacokinetic processes, defined as the actions of drugs or other chemical agents within an organism over a period of time, namely absorption, distribution, metabolism, and elimination. In addition, whole animal models cover a comprehensive array of pharmacodynamic processes, defined as the cellular, molecular, biochemical, and pharmacologic changes induced by a test agent. Coordinating all of these processes, conveniently ''housed under one roof,'' so to speak, are the endocrine and neurological systems.

A major advantage of pregnant animal models is that they represent multiple individuals contained within one physiological (maternal) unit. The maternal system undergoes numerous physiological changes (e.g., increased cardiac output, hemodilution) as well as alterations in pharmacokinetic processes. In fact, these physiological changes could alter sensitivity of the pregnant animal relative to nonpregnant state. The placenta, a temporary organ formed solely for pregnancy, also carries out numerous, highly complex functions, among them toxicant transfer to the embryo/fetus (Table 3). These placental functions vary with stage of gestation, as well as across species (Carney et al., 2004). Similarly, the acquisition of metabolic capacity has its own stage- and enzyme-specific ontogeny (Rich and Boobis, 1997) such that the pharmacokinetic profile in the embryo, fetus, and/or pup is not necessarily the same, and does not necessarily mirror maternal pharmacokinetics. Ultimately, developmental toxicity is determined by a combination of pharmacokinetic conditions as present during specific critical windows of susceptibility.

Table 3
Placental Functions: An Example of Complex Biology Inherent in Pregnant Animal Models

Function	Gestational changes
Regulation of toxicant disposition to the embryo/fetus	Maternal:conceptus partitioning changes according to changes in blood flow, pH gradients, and placental type (e.g., yolk sac versus chorioallantoic placenta)
Nutrient and gas exchange	Maintains relatively anaerobic conceptus environment in early pregnancy; more aerobic as conceptus enlarges later in development
Endocrine function	In humans, syncytiotrophoblast cells produce 17β-estradiol, progesterone and chorionic gonadotropin early in the pregnancy. Humans and rodents produce placental lactogens later in pregnancy
Metabolizing enzymes	In general, expression of CYPs and other drug metabolizing enzymes increases later in pregnancy, or postnatally

Relative to in vitro systems which typically represent one or a few discrete aspects of this complex biology, another major advantage of whole animal models is the ability to assess toxicity driven by interactions between distant organs and/or tissues. For example, lipopolysaccharide was not directly toxic to cultured mouse embryos, yet it caused placental infarction and embryo death when administered to pregnant mice, perhaps due to consequences from the induction of the acute phase response resulting in increased maternal serum levels of Tumor Necrosis Factor-α, among other cytokines (Leazer et al., 2003). A related maternally mediated mode of toxic action that is relevant to a diverse range of chemicals involves a cascade initiated by maternal tissue damage and associated induction of a maternal acute phase response. In this example (Fig. 1), the initial maternal acute phase response in turn leads to release of Tumor Necrosis Factor-α, which then induces expression of the zinc binding protein, metallothionein, leading to sequestration of zinc in the maternal liver, finally resulting in an embryonic zinc deficiency and teratogenicity (Taubeneck et al., 1994). In contrast, teratogenicity does not occur following in vitro exposure of embryos to these compounds. Because many developmental toxicants act via such complex modes of action, one can see the great value in whole animal model systems, and the great challenge of assessing developmental toxicity solely via in vitro models.

Maternal Toxicity

Although the use of pregnant animal models enables detection of maternally mediated developmental toxicity, maternal toxicity also can be a major problem for data interpretation and risk assessment. In fact, approximately 75% of the chemicals evaluated in standard guideline studies show developmental toxicity only at maternally toxic dose levels, making this the most common scenario in these tests (Khera, 1984, 1985). The key question for risk assessment is whether or not the relationship between maternal and developmental toxicity is causal, but often this determination is difficult to make. The fact that maternal and developmental toxicity occur at the

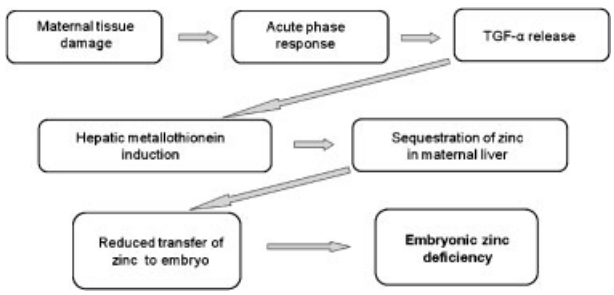

Fig. 1. Example of maternally mediated developmental toxicity. Cascade of maternal tissue damage leading to TGF-α release, maternal hepatic metallothionein induction, sequestration of circulating zinc, embryonic zinc deficiency, and developmental toxicity.

same dose level is, by itself, insufficient to establish causality. Instead, one must consider the specific nature of maternal toxicity, its severity, its timing during gestation and whether or not the maternal toxicity corresponds with the type of development effects observed (Carney, 1996). Such evaluations require a comprehensive assessment of the data coupled with a considerable degree of expert judgement.

Developmental toxicologists have wrestled with maternal toxicity for over two decades, and many reviews and discussions of the relationship between maternal and embryo/fetal toxicities have been published, with differing conclusions (Khera, 1985; Chernoff et al., 1989; Hood, 1989; Chahoud et al., 1999; Rogers et al., 2005). This topic continues to be discussed and debated, as evidenced by a recent series of workshops on maternal toxicity and its relationship to developmental toxicity (Beyer et al., 2011). There has been speculation that some manifestations of developmental toxicity are more strongly related to maternal toxicity than others, but exact correlations have been elusive, particularly with regard to malformations. The most common effect stemming from maternal toxicity is developmental delay, often evidenced by delayed ossification and/or decreased fetal weight, especially when maternal toxicity occurs during late gestation when fetal growth is maximal.

Frequently the relationship between developmental and maternal toxicity is even more difficult to assess due to multiple mechanisms of action that need to be teased apart. For example, atrazine given to pregnant F344 rats induced full litter resorption that appeared to be mediated by decreased maternal secretion of luteinizing hormone (Narotsky et al., 2001). However, dose-related reductions in maternal body weight gain also were induced by atrazine, and it was not clear whether increased prenatal loss observed at high doses in the surviving litters was related to maternal toxicity or could have resulted from direct embryotoxicity. In humans, fetal alcohol spectrum disorders are observed most frequently in the offspring of women whose intake of alcohol would be considered toxic to both mother and offspring. The poor nutritional status of the mother (common in alcoholics) may contribute to fetotoxicity and genetic differences in metabolism among subpopulations of women have been associated with differing susceptibilities to fetal alcohol spectrum disorder, but other fetal effects are considered to be directly related to exposure to ethanol and its metabolites (Pollard, 2007; Guerri et al., 2009).

Finally, some types of maternal toxicity observed in animals are not relevant to humans. For example, pregnant rabbits exposed to some antimicrobial drugs will stop eating and spontaneously abort their litters. This inappetence is likely due to GI intolerance secondary to antimicrobial effects and not a direct effect of the drug on the developing embryo or fetus (Clark et al., 1986). Several groups have demonstrated that food restriction to levels that cause maternal body weight loss can cause spontaneous abortion in rabbits (Matsuzawa et al., 1981; Petrere et al., 1993; Cappon et al., 2005). Thus, spontaneous abortions observed in rabbits that are intolerant of an antimicrobial drug product are not considered relevant for risk assessment in humans.

Knowledge of Basic Developmental Biology of Standard Animal Models

Although we recognize the value of the complex, integrated biology afforded by pregnant whole animal models, we do not necessarily understand it very well. This lack of understanding is especially true for the rabbit, which has been a major test species underpinning more than 40 years of developmental toxicity safety assessment, yet very little is known about the basic developmental biology of rabbits. The literature on descriptive embryology of the rabbit is covered in just a few papers (Minot and Taylor, 1905; Waterman, 1943; Edwards, 1968; Pitt and Carney, 1999; Carney et al., 2007), but there is essentially nothing known about developmental lineage, developmental control genes, signaling pathways, or molecular mechanisms of developmental toxicity in this species. The amount of information is slightly greater for the rat, but still is paltry in comparison to the extensive knowledge amassed for the mouse and lower organisms such as the frog, zebrafish, Drosophila, and the nematode, *C. elegans*. This knowledge gap for mainstream developmental toxicology animal models (i.e., rat, rabbit) versus model organisms used in basic developmental biology research is becoming ever wider as we advance into the future, and should give pause for consideration. Should we devote more resources toward understanding the developmental biology of rats, rabbits, and monkeys, or should we start moving toward the mouse or even lower organisms for developmental toxicology safety evaluations?

End Points and Data Interpretation

Developmental toxicity safety assessment is mainly a descriptive science designed to detect adverse developmental outcomes, namely teratogenicity, intrauterine death, intrauterine growth retardation, and functional deficits. Evaluation of teratogenicity requires detailed examinations of fetal morphology, including external features, internal organs and tissues, and assessment of more than 200 bones of the fetal skeleton. These assessments have evolved over time, such that very subtle changes (often called variations) can be detected, in addition to frank terata.

The descriptive nature of these fetal examinations brings with it some critical challenges confronting us as we design for the future. One is that the evaluation criteria and nomenclature for fetal morphology have been difficult to standardize across different laboratories. Although this problem would seem to be easily remedied, it has been difficult because individual laboratories have built up large volumes of historical data based on their own criteria, and they also may use different animal strains and evaluate fetuses on different days of gestation. Fetal examinations also are very time consuming and labor intensive, and require a significant investment in examiner training in fetal morphology, coupled with extensive proficiency testing.

One issue with skeletal evaluation is the interpretation of minor skeletal variations and their impact on risk assessment. This issue was the subject of a previous ILSI-HESI expert panel project, the outcome of which is reviewed in Daston and Seed (2007). Depending on the laboratory's evaluation scheme, a large number of individual skeletal variations often are recorded and

some occur at a very high incidence (sometimes >80%), even in control animals. Many laboratories distinguish between several subtly different degrees of ossification of individual bones, leading to a large volume of statistical analyses and evaluation of corresponding historical control data (reviewed in (Carney and Kimmel, 2007). Although the practice of recording minor skeletal variations was established many years ago, we have since learned that the skeletal system possesses an extensive capacity to remodel during postnatal development, and current evidence indicates that many of the minor skeletal variations present in the term fetus are no longer evident postnatally (Nishimura et al., 1982; Wilson et al., 1985; Collins et al., 1987; Marr et al., 1992; Price et al., 1996). Thus, minor skeletal variations, particularly findings such as wavy ribs and minor delays in ossification are generally not considered adverse in and of themselves (Carney and Kimmel, 2007). With the advent of new technology such as micro-CT image analysis of the fetal skeletal (Winkelmann and Wise, 2009), we may be able to dispense with the recording of some of these minor changes and replace them with more clinically relevant and readily interpretable measures such as bone density and/or quantitative measures of overall degree of ossification. The digital nature of micro-CT makes such quantitative measures possible, and also may foster more objective and consistent evaluations. The throughput of this technology is improving greatly, making it a realistic option for many laboratories. However, additional research in this area will be necessary in order to gain regulatory acceptance.

The interpretation of fetal malformations can also be a challenge, particularly when faced with a low incidence of a particular malformation occurring in the high-dose group only. As highlighted by Palmer many years ago, "because low rates of malformation are the rule, one faces the recurring nightmare of deciding whether one or two malformations are related to treatment or accidental" (Palmer, 1968). Currently there are few options for resolving these issues, which is of particular concern given the enormous impact on regulation of the chemical as well as the potential labeling of the compound as a teratogen. In some cases, the studies have been repeated using extremely large sample sizes, but this is obviously problematic in terms of animal use, costs, and time. Mechanistic studies are another option, although these may only be possible if higher doses can be used to increase the incidence. As shown in Figure 2, statistics often are of limited help in resolving these uncertainties, as very large numbers of offspring are needed to achieve the statistical power needed to detect an increase in low incidence malformations. To overcome some of these statistical limitations, historical control data are considered in judging whether or not a low incidence finding seen in a treated group might have been a chance occurrence. However, historical control data should be used judiciously and within a reasonable time frame, as drift in the background incidence can occur over time, as can sudden spikes in the incidence of a particular effect.

As we consider testing approaches for the future, we must acknowledge that the embryological basis for many fetal morphological changes, as well as their low incidence, can make them very difficult to detect using high-throughput approaches. Take for example retroesophageal subclavian artery, a rare malformation in which the right subclavian artery originates along the descending aorta, rather than at the innominate, causing it to pass dorsal to the esophagus instead of ventrally as the normal vessel does. This malformation can have serious functional consequences such as chronic dysphagia and respiratory infections due to constriction of the esophagus and trachea (Smith et al., 1979). However, the biochemical and cellular structure of the vessel is apparently normal, and the only problem is that it has been routed differently. If such a malformation was induced by a particular drug or chemical, it is difficult to envision how a high-throughput gene-based assay would detect a perturbation of this nature.

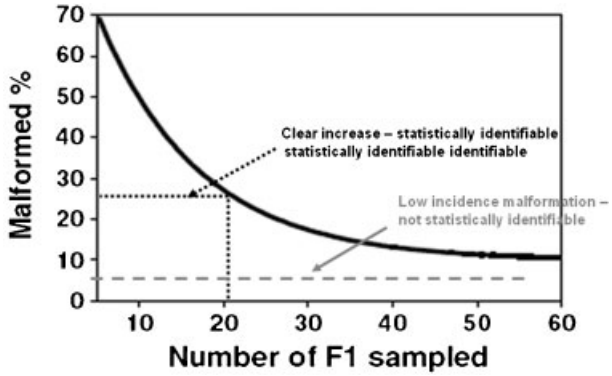

Fig. 2. Comparison of sample size needed to detect a high incidence malformation (black line, short dashes) versus a low incidence malformation (gray line, long dashes). Detection of low incidence malformations requires large numbers of offspring, often more than are available in a standard developmental toxicity study. Adapted from (Hotchkiss et al., 2008). Used with permission of Oxford University Press.

Testing in Multiple Species

Developmental toxicology testing is routinely conducted in two species; a rodent and a nonrodent. The basis for this practice derives from thalidomide, to which rabbits were the only test species exhibiting a teratogenic response, thus leading to inclusion of the rabbit as a preferred test species in addition to the standard workhorse of toxicology, the rat. To increase the chances of detecting developmental toxicity, both the rat and rabbit are generally required for developmental toxicity testing of industrial and agricultural chemicals and small molecule pharmaceuticals, although a different nonrodent species, including nonhuman primates, may be selected if deemed more appropriate based on pharmacokinetic (including placental transfer) or pharmacodynamic similarities to humans. For biological pharmaceuticals, the nonhuman primate may be the only animal model relevant to humans if other species do not express the pharmacologic target. The most common nonhuman primate used for developmental toxicity testing of pharmaceuticals is the cynomolgus monkey (*Macaca fascicularis*) and the number of monkey developmental toxicity studies has increased dramatically in recent years, to the point where it is exerting a strain on the availability of these animals for research. In contrast, the use of nonhuman primates for industrial and agricultural compounds is extremely rare. Guidelines for the conduct of reproduction toxicity studies of small

molecule and biotechnology-derived pharmaceuticals are available from the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH, 1997, 2005). Guidelines for the conduct of reproduction studies for food additives are available from the US Food and Drug Administration and the European Union (FDA, 2000; ECSCF, 2001).

Nonhuman primates are not the ideal species for developmental toxicity testing, although their use may be necessary in particular situations as discussed previously. These animals have low fertility rates (40–70%) and high spontaneous abortion rates (10–30%) relative to other species used in developmental toxicity testing (Martin et al., 2009), making data from studies in nonhuman primates difficult to interpret. These problems are exacerbated by small group sizes, due to the expense of purchasing and maintaining these animals. As a result of this variability combined with low sample sizes, statistical power is extremely weak. There has been discussion regarding means to increase the amount of information that can be obtained from studies in nonhuman primates. Some have proposed conducting microscopic evaluation of the offspring in addition to the external, visceral and skeletal examinations. Another suggestion is that offspring be allowed to reach a postnatal age that would permit behavioral testing if there are questions/concerns regarding a test article's effects on the developing nervous system (Stewart, 2009). The trend toward increased nonhuman primate studies deserves serious reconsideration owing to animal welfare concerns, the high cost of such studies, high variability, and low statistical power. Despite the fact that nonhuman primates are phylogenetically most similar to humans, very little is actually known about the basic developmental biology of these species and the assumption that they accurately model the human may not always be true.

Interspecies Concordance and Extrapolation to Humans

The ability of animal models to predict the human response is a fundamental assumption in developmental toxicity and risk assessment, yet varying degrees of discordance among species are very common in actual practice. Discordance may be manifest in various shades of gray, such as different types of effects across two positively responding species, similar effects but varying degrees of sensitivity, or situations in which a compound is positive in one species and negative in another. Accordingly, some have argued that animal-based toxicity testing is fundamentally flawed, reasoning that if species as similar as rats and rabbits show discordance, how can the tests possibly predict what will happen in humans (Bailey et al., 2005)?

Even if one rejects this position, the challenge of interspecies discordance should at least prompt us to ask why it exists, which could lead to improvements in animal to human extrapolations. Discordance in developmental toxicity testing certainly seems to conflict with the widely held dogma stating that the basic events in embryo development are highly conserved across species, even for species as disparate as fruit flies, frogs, mice, and humans. This degree of conservation mainly applies to the most fundamental processes in embryogenesis, such as establishment of the general body plan, pattern formation, cellular induction, and regulation of differentiation via signaling pathways.

On the other hand, pharmacokinetics and in particular, maternal metabolism can vary widely between species and are likely to drive interspecies discordance. If test animal and human metabolic processes for the test chemical are different, then false positives or false negatives could result, depending on whether toxicity is caused by the parent compound or a metabolite. Placental anatomy and physiology also vary greatly between conventional test species and humans. In fact, rats, mice, and rabbits utilize two very different types of placentae—the inverted visceral yolk sac placenta which is extremely important in early pregnancy, as well as a chorioallantoic placenta which does not become functional until mid-pregnancy. In contrast, humans only utilize a chorioallantoic type of placenta throughout most of gestation (Georgiades et al., 2002).

An example of pharmacokinetics driving interspecies discordance in response is ethylene glycol (EG), to which rats and mice are susceptible species, whereas rabbits show no evidence of developmental toxicity (Carney et al., 2008). This species difference appears to be due to differences in rates of hepatic metabolism of EG to the proximate toxicant, glycolic acid, as well as to a much more limited transfer and/or uptake of glycolic acid in the rabbit embryo relative to the rat. As shown in Figure 3, equivalent doses of EG given to pregnant rats and rabbits result in nearly identical peak blood levels of parent EG, but a 10-fold difference in levels of glycolic acid in the embryo.

Thalidomide is an example in which interspecies discordance may be due to pharmacodynamic differences, rather than pharmacokinetics. Biochemical differences in redox regulation and the ability to cope with thalidomide-induced oxidative stress are apparent, leading to misregulation of the NF-kB/FGF-10 signaling

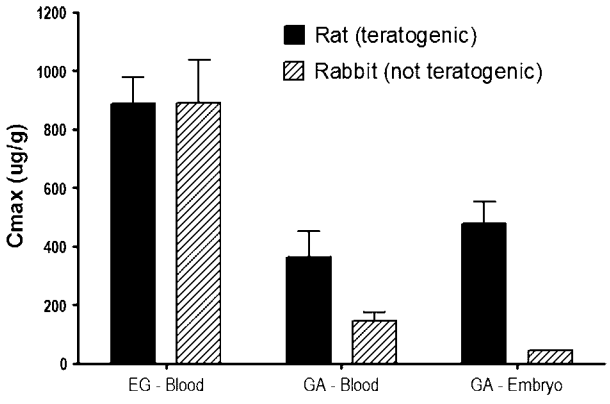

Fig. 3. Ethylene glycol (EG): an example of species discordance driven by differences in pharmacokinetics. Equivalent doses of EG given to pregnant rats and rabbits result in equivalent peak concentrations of parent compound in maternal blood, but much lower concentrations of the teratogenic metabolite (glycolic acid; GA) in maternal blood and embryo of rabbits (the insensitive species) relative to rats (the sensitive species). These differences are due to the lower rate of metabolism of EG to GA in rabbits, as well as limited transfer via the yolk sac placenta of rabbits. From (Carney et al., 2008). Used with permission of Wiley-Liss.

pathway in the developing limb bud of the rabbit, but not the rat (Hansen and Harris, 2004). In any event, given that the ability to use animal data to estimate risks to human development is a fundamental premise of toxicology, it seems that we need a much greater mechanistic understanding of the causes for interspecies discordance and what it might mean for human development.

Dosimetry Context

The administration of test material at maximum tolerated doses via the gavage route of exposure has been the default practice in developmental toxicology for several decades. This practice makes perfect sense for pharmaceuticals which are given orally at therapeutically active dose levels in humans. For industrial and agricultural compounds found in the environment, typical human exposures tend to occur at much lower levels, often orders of magnitude lower, than the doses used in test animals. Furthermore, human exposure to these compounds is more likely to occur via dermal and/or multiple routes of exposure, and also is more likely to be spread over time rather than as a bolus. Nonetheless, most testing guidelines for chemicals still specify gavage administration of maximally tolerated dose levels as the default, even in cases where human exposures are known to be orders of magnitude lower.

In recent years, advances in toxicokinetics, dose–response modeling and mode of action research have demonstrated that high doses can overwhelm normal detoxification processes, often leading to major shifts in compound metabolism and/or saturation of toxicokinetic processes such as renal clearance. In the case of many industrial and agricultural chemicals, these unique high-dose responses are unlikely to occur at lower doses characteristic of human exposure (Holsapple and Wallace, 2008), rendering the data of questionable relevance for human risk assessment. In fact, a 1999 Society of Toxicology Task Force stated that ''*The relevance of using doses that are many multiples of conceivable human exposures … is, at most, quite dubious … the predicted risks may have little or no relationship to risk in the real world*.'' (Conolly et al., 1999). The negative consequences of this practice include unnecessary animal use, cost, and effort as the irrelevant findings generated are further investigated. This testing scheme also can lead to the erroneous classification and labeling of compounds which pose little risk to humans.

Alternative approaches to dosing that increase relevance to humans are emerging, in part due to advances in human biomonitoring and exposure modeling, as well as the incorporation of pharmacokinetics into developmental toxicity testing. Although the incorporation of pharmacokinetics has been standard for pharmaceutical compounds, only recently have pharmacokinetics begun to be included in testing of agricultural and industrial chemicals. A recent approach to dose setting that marries human exposure data and pharmacokinetics is called the Kinetically Derived Maximum dose (KMD), and is an alternative to the classical maximum-tolerated dose (Saghir et al., 2009). In the Kinetically Derived Maximum dose, the top dose level is defined as the highest dose level showing linear kinetics and is set at or near the point at which kinetics shift from linear to nonlinear. This approach is valid for compounds to which human exposure is known to be very low, such that a transition to nonlinear kinetics (e.g., driven by saturation of metabolism or renal clearance) is not plausible in humans. Thus, the testing is still conducted at high multiples of human exposure in an effort to ensure detection power, but the extreme high-dose range, that usually generates effects that are not relevant to humans, is avoided.

DISCUSSION

Based on the previous evaluation, the aspects of current practice which are critical to retain (i.e., the ''baby''), versus those which can we begin to dispense with (the ''bathwater'') can be summarized as follows (Table 4).

Future Strategies

As the remaining presentations in this workshop were devoted to future strategies and refinements, these will only be mentioned briefly here. However, it is clear that a combination of both short- and long-term strategies need to be pursued. In the short run, we can take advantage of technologies and approaches that already are available, such as enhanced study designs which incorporate pharmacokinetics and mechanistic end points, the utilization of exposure and dosimetry information to set dose levels and choose routes of exposure which are

Table 4
Summary of Strengths and Weaknesses of Current Developmental Toxicity Safety Assessment

Strengths	Weaknesses
Integrated, high complex biology of current models, which includes a full range of pharmacokinetic and pharmacodynamic processes, as well as a maximal capacity for interactions	Large numbers of animals required
The ability to detect low incidence malformations	High cost per compound (>$100,000 per study)
The ability to detect maternally mediated effects	Long time to evaluate each compound
Phylogenetic similarity of mammalian models to human	Capacity gap—cannot keep pace with increasing demands to evaluate existing and new chemicals, as well as mixtures
Long history of use and extensive historical data	Maternal toxicity—can confound data interpretation
Present testing models are considered the most definitive currently available	Fundamental knowledge of developmental biology for current animal models (e.g., rat, rabbit, monkey) is sparse relative to mouse or lower organisms
Every very chemical or drug known to be teratogenic in humans, with possible two exceptions, is also teratogenic in one or more laboratory species (Schardein, 2000)	Uncertainty regarding interpretation of low incidence findings
	Large amount of effort placed on the evaluation of minor skeletal variations with little impact on risk assessment
	Use of high doses that sometimes far exceed human exposure levels

more relevant to humans, and increased use of early stage in vitro and/or in silico screening tools. In addition, tiered strategies which consider exposure data, chemical class, and structure–activity relationship information can be used to limit the need for animal testing, and to get the most information out of the testing that is done. Some of the time-honored practices of traditional skeletal examination also warrant reconsideration, particularly considering their utility for human risk assessment versus the labor-intensive nature of these evaluations.

In the longer term, the scientific, technological, and societal drivers for change are steering us toward entirely new paradigms such as the pathways-based system for evaluating chemicals as envisioned by NAS (NAS, 2007). One key point that should be evident from the present discussion is that any new paradigm we adopt should not simply be a quicker, cheaper version of the old one. For example, we already know that a high-throughput in vitro system is unlikely to detect the vast majority of known malformations due to their very low incidences even following teratogen exposure, particular those malformations like retroesophageal subclavian artery which involve subtle spatial alterations. Similarly, how would an in vitro test or a fish embryo assay identify relevant maternally mediated developmental effects?

Therefore, the goal of new testing paradigms should not be the detection of specific effects or hazards, but instead, to gather a body of information that can better predict risk to humans. The NAS vision calls for integration of pathways-based assay results with toxicokinetics and human exposure modeling to constitute a risk-based rather than hazard-based evaluation system. This risk context is ultimately what is needed to manage chemicals safely, but without denying the public access to the benefits of these chemicals. The risk-based NAS paradigm, which calls for stronger links between exposure modeling, kinetic modeling, and signaling pathway data, represents a platform to foster continual improvements in risk assessment. As the technology to estimate human exposures and internal dosimetry improves, so will the accuracy and relevance of human risk assessments.

The specific details of how this new paradigm might be implemented remains to be elucidated, but were discussed in subsequent presentations in this workshop. Suffice it to say that the transition to new testing paradigms needs to be done methodically and systematically as we cannot afford to erode public trust in our safety assessment system. It will be critical to anchor and validate new approaches against existing whole animal data, as well as human data where they exist. As we embark on new directions in developmental toxicity testing, not only must we address the weaknesses of current models, but we must also leverage their strengths.

REFERENCES

Andersen ME, Krewski D. 2009. Toxicity testing in the 21st century: bringing the vision to life. Toxicol Sci 107:324–330.

Bailey J, Knight A, Balcombe J. 2005. The future of teratology research is in vitro. Biogenic Amines 19:97–145.

Beyer BK, Chernoff N, Danielsson BR, Davis-Bruno K, Harrouk W, Hood RD, Janer G, Wandel Liminga U, Kim JH, Rocca M, Rogers J, Scialli AR. ILSI/HESI Maternal Toxicity Workshop Summary: Maternal Toxicity and Its Impact on Study Design and Data Interpretation. Birth Defects Res. (Part B) Devel Reprod Toxicol 92: 36–51.

Cappon GD, Fleeman TL, Chapin RE, Hurtt ME. 2005. Effects of feed restriction during organogenesis on embryo-fetal development in rabbit. Birth Defects Res B Dev Reprod Toxicol 74:424–430.

Carney EW. 1996. Maternal physiological disruption. In: Daston GaRK, editor. Handbook of Experimental Pharmacology: Drug Toxicity in Embryonic Development. Springer. p 573–594.

Carney EW, Kimmel CA. 2007. Interpretation of skeletal variations for human risk assessment: delayed ossification and wavy ribs. Birth Defects Res B Dev Reprod Toxicol 80:473–496.

Carney EW, Scialli AR, Watson RE, DeSesso JM. 2004. Mechanisms regulating toxicant disposition to the embryo during early pregnancy: an interspecies comparison. Birth Defects Res C Embryo Today 72:345–360.

Carney EW, Tornesi B, Keller C, et al. 2007. Refinement of a morphological scoring system for postimplantation rabbit conceptuses. Birth Defects Res B Dev Reprod Toxicol 80:213–222.

Carney EW, Tornesi B, Markham DA, et al. 2008. Species-specificity of ethylene glycol-induced developmental toxicity: toxicokinetic and whole embryo culture studies in the rabbit. Birth Defects Res B Dev Reprod Toxicol 83:573–581.

Chahoud I, Ligensa A, Dietzel L, Faqi AS. 1999. Correlation between maternal toxicity and embryo/fetal effects. Reprod Toxicol 13: 375–381.

Chernoff N, Rogers JM, Kavlock RJ. 1989. An overview of maternal toxicity and prenatal development: considerations for developmental toxicity hazard assessments. Toxicology 59:111–125.

Clark RL, Robertson RT, Peter CP, et al. 1986. Association between adverse maternal and embryo-fetal effects in norfloxacin-treated and food-deprived rabbits. Fundam Appl Toxicol 7:272–286.

Collins TF, Welsh JJ, Black TN, et al. 1987. Potential reversibility of skeletal effects in rats exposed in utero to caffeine. Food Chem Toxicol 25:647–662.

Conolly RB, Beck BD, Goodman JI. 1999. Stimulating research to improve the scientific basis of risk assessment. Toxicol Sci 49:1–4.

Daston GP, Seed J. 2007. Skeletal malformations and variations in developmental toxicity studies: interpretation issues for human risk assessment. Birth Defects Res B Dev Reprod Toxicol. 80:421–424.

ECSCF. 2001. Guidance on Submissions for Food Additive Evaluations by the Scientific Committee on Food (dated 11 July 2001).

Edwards JA. 1968. The external development of the rabbit and rat embryo. Adv Teratol 3:239–263.

FDA. 2000. Redbook 2000 Chapter IV.C.9.a.: Guidelines for Reproduction Studies.

Georgiades P, Ferguson-Smith AC, Burton GJ. 2002. Comparative developmental anatomy of the murine and human definitive placentae. Placenta 23:3–19.

Grindon C, Combes R, Cronin MT, et al. 2006. Integrated testing strategies for use in the EU REACH system. Altern Lab Anim 34:407–427.

Guerri C, Bazinet A, Riley EP. 2009. Foetal alcohol spectrum disorders and alterations in brain and behaviour. Alcohol Alcohol 44:108–114.

Hansen JM, Harris C. 2004. A novel hypothesis for thalidomide-induced limb teratogenesis: redox misregulation of the NF-kappaB pathway. Antioxid Redox Signal 6:1–14.

Holsapple MP, Wallace KB. 2008. Dose response considerations in risk assessment—an overview of recent ILSI activities. Toxicol Lett 180: 85–92.

Hood RD. 1989. A perspective on the significance of maternally mediated developmental toxicity. Regul Toxicol Pharmacol 10:144–148.

Hotchkiss AK, Rider CV, Blystone CR, et al. 2008. Fifteen years after ''Wingspread''—environmental endocrine disrupters and human and wildlife health: where we are today and where we need to go. Toxicol Sci 105:235–259.

ICH. 1997. S6: Preclinical Safety Evaluation of Biotechnology-Derived Pharmaceuticals (ICH Harmonised Tripartite Guideline; dated 16 July 1997).

ICH. 2005. S5(R2): detection of toxicity to reproduction for medicinal products & toxicity to male fertility In: ICH Harmonised Tripartite Guideline; Parent Guideline dated 24 June 1993 AdN, incorporated in November 2005, editor.

Khera KS. 1984. Maternal toxicity—a possible factor in fetal malformations in mice. Teratology 29:411–416.

Khera KS. 1985. Maternal toxicity: a possible etiological factor in embryo-fetal deaths and fetal malformations of rodent-rabbit species. Teratology 31:129–153.

Leazer TM, Daston GP, Keen CL, Rogers JM. 2003. The embryolethality of lipopolysaccharide in CD-1 and metallothionein I-II null mice: lack of a role for induced zinc deficiency or metallothionein induction. Toxicol Sci 73:442–447.

Marr MC, Price CJ, Myers CB, Morrissey RE. 1992. Developmental stages of the CD (Sprague-Dawley) rat skeleton after maternal exposure to ethylene glycol. Teratology 46:169–181.

Martin PL, Breslin W, Rocca M, et al. 2009. Considerations in assessing the developmental and reproductive toxicity potential of biopharmaceuticals. Birth Defects Res B Dev Reprod Toxicol 86:176–203.
Matsuzawa T, Nakata M, Goto I, Tsushima M. 1981. Dietary deprivation induces fetal loss and abortion in rabbits. Toxicology 22:255–259.
Minot CS, Taylor E. 1905. Normal Plates of the Development of the Rabbit (Lepus cuniculus). Jena: Verlag von Gustav Fishcher. 101p.
Narotsky MG, Best DS, Guidici DL, Cooper RL. 2001. Strain comparisons of atrazine-induced pregnancy loss in the rat. Reprod Toxicol 15: 61–69.
Nishimura M, Iizuka M, Iwaki S, Kast A. 1982. Repairability of drug-induced ''wavy ribs'' in rat offspring. Arzneimittelforschung 32: 1518–1522.
NAS. 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy. Washington, DC: The National Academies Press.
Palmer AK. 1968. Spontaneous malformations of the New Zealand White rabbit: the background to safety evaluation tests. Lab Anim 2: 195–206.
Petrere JA, Rohn WR, Grantham 2nd LE, Anderson JA. 1993. Food restriction during organogenesis in rabbits: effects on reproduction and the offspring. Fundam Appl Toxicol 21:517–522.
Pitt JA, Carney EW. 1999. Development of a morphologically-based scoring system for postimplantation New Zealand White rabbit embryos. Teratology 59:88–101.
Pollard I. 2007. Neuropharmacology of drugs and alcohol in mother and fetus. Semin Fetal Neonatal Med 12:106–113.
Price CJ, Strong PL, Marr MC, et al. 1996. Developmental toxicity NOAEL and postnatal recovery in rats fed boric acid during gestation. Fundam Appl Toxicol 32:179–193.
Rich KJ, Boobis AR. 1997. Expression and inducibility of P450 enzymes during liver ontogeny. Microsc Res Tech 39:424–435.
Rogers JM, Chernoff N, Keen CL, Daston GP. 2005. Evaluation and interpretation of maternal toxicity in Segment II studies: issues, some answers, and data needs. Toxicol Appl Pharmacol 207:367–374.
Rovida C, Hartung T. 2009. Re-evaluation of animal numbers and costs for in vivo tests to accomplish REACH legislation requirements for chemicals - a report by the transatlantic think tank for toxicology (t(4)). Altex 26:187–208.
Saghir SA, Marty MS, Clark AJ, et al. 2009. A dietary dose range-finding and toxicokinetic (TK) study of 2,4-dichlorophenoxyacetic acid (2,4-D) in adult CRL:CD(SD) rats and their offspring: I. Toxicokinetics. Toxicologist 114(Supplement 1):Abstract 1161.
Schardein JL. 2000. Chemically Induced Birth Defects. New York: Marcel Dekker. p. 40.
Scialli AR. 2008. The challenge of reproductive and developmental toxicology under REACH. Regul Toxicol Pharmacol 51:244–250.
Smith 3rd JM, Reul Jr GJ, Wukasch DC, Cooley DA. 1979. Retro-esophageal subclavian arteries: surgical management of symptomatic children. Cardiovasc Dis 6:331–334.
Stewart J. 2009. Developmental toxicity testing of monoclonal antibodies: an enhanced pre- and postnatal study design option. Reprod Toxicol 28:220–225.
Taubeneck MW, Daston GP, Rogers JM, Keen CL. 1994. Altered maternal zinc metabolism following exposure to diverse developmental toxicants. Reprod Toxicol 8:25–40.
Van der Jagt KMS, Torslov J, de Bruijn J. 2004. Alternative approaches can reduce the use of test animals under REACH 2004. European Commission Report EUR 21405 EN.
Waterman AJ. 1943. Studies of normal development of the New Zealand White strain of rabbit. Am J Anat 72:473–515.
Wilson GN, Howe M, Stover JM. 1985. Delayed developmental sequences in rodent diabetic embryopathy. Pediatr Res 19:1337–1340.
Winkelmann CT, Wise LD. 2009. High-throughput micro-computed tomography imaging as a method to evaluate rat and rabbit fetal skeletal abnormalities for developmental toxicity studies. J Pharmacol Toxicol Methods 59:156–165.