DOI: 10.1002/bdr2.1212

RESEARCH ARTICLE

WILEY Birth Defects Research THE TERATOLOGY SOCIETY EST. 1960

Rethinking developmental toxicity testing: Evolution or revolution?

Anthony R. Scialli[1] | George Daston[2] | Connie Chen[3] | Prägati S. Coder[4] | Susan Y. Euling[5] | Jennifer Foreman[6] | Alan M. Hoberman[7] | Julia Hui[8] | Thomas Knudsen[9] | Susan L. Makris[10] | LaRonda Morford[11] | Aldert H. Piersma[12] | Dinesh Stanislaus[13] | Kary E. Thompson[14]

[1] Reproductive Toxicology Center and Scialli Consulting LLC, Washington, DC

[2] Proctor & Gamble, Mason, Ohio

[3] ILSI Health and Environmental Sciences Institute, Washington, DC

[4] Charles River Laboratories, Ashland, Ohio

[5] Office of Children's Health Protection, U.S. Environmental Protection Agency, Washington, DC

[6] ExxonMobil Biomedical Sciences, Inc, Annandale, New Jersey

[7] Charles River Laboratories, Horsham, Pennsylvania

[8] Celgene Corporation, Summit, New Jersey

[9] National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

[10] National Center for Environmental Assessment, U.S. Environmental Protection Agency, Washington, DC

[11] Lilly Research Laboratories, Indianapolis, Indiana

[12] Center for Health Protection, National Institute for Public Health and the Environment RIVM, Bilthoven and Institute for Risk Assessment Sciences, Utrecht University, Utrecht, Netherlands

[13] GlaxoSmithKline, King of Prussia, Pennsylvania

[14] Drug Safety Evaluation, Bristol-Myers Squibb, New Brunswick, New Jersey

Correspondence
Anthony R. Scialli, Scialli Consulting LLC, 2737 Devonshire Pl NW #120, Washington DC 20008-3459.
Email: ascialli@scialliconsulting.com

Funding information
Funding for the Workshop was provided by the ILSI Health and Environmental Sciences Institute. Author employers are as listed

Background: Current developmental toxicity testing adheres largely to protocols suggested in 1966 involving the administration of test compound to pregnant laboratory animals. After more than 50 years of embryo-fetal development testing, are we ready to consider a different approach to human developmental toxicity testing?

Methods: A workshop was held under the auspices of the Developmental and Reproductive Toxicology Technical Committee of the ILSI Health and Environmental Sciences Institute to consider how we might design developmental toxicity testing if we started over with 21st century knowledge and techniques (revolution). We first consider what changes to the current protocols might be recommended to make them more predictive for human risk (evolution).

Results: The evolutionary approach includes modifications of existing protocols and can include humanized models, disease models, more accurate assessment and testing of metabolites, and informed approaches to dose selection. The revolution could start with hypothesis-driven testing where we take what we know about a compound or close analog and answer specific questions using targeted experimental techniques rather than a one-protocol-fits-all approach. Central to the idea of hypothesis-driven testing is the concept that testing can be done at the level of mode of action. It might be feasible to identify a small number of key events at a molecular or cellular level that predict an adverse outcome and for which testing could be performed in vitro or in silico or, rarely, using limited in vivo models. Techniques for evaluating these key events exist today or are in development.

Discussion: Opportunities exist for refining and then replacing current developmental toxicity testing protocols using techniques that have already been developed or are within reach.

KEYWORDS

adverse outcome pathways, developmental toxicity testing, embryo-fetal toxicity testing, hypothesis-based testing, Segment II, virtual embryo

1 | INTRODUCTION

In 1966, Edwin I. Goldenthal, Chief of the Drug Review Branch at the US Food and Drug Administration, wrote a letter to pharmaceutical companies about methods of evaluating drugs for adverse reproductive effects. The letter started, “During the past several years following the thalidomide episode, we have been recommending a study designed to determine the potential of drugs for producing adverse effects on the reproductive process.” Attached to this letter were guidelines that outlined the three-segment protocol with which we are now familiar. Goldenthal ended his letter with the statement, “It must be realized that even these improved guidelines reflect merely the “state of the art” at the present time, and undoubtedly further modifications will be needed in the future as additional knowledge in this area is developed.” Goldenthal understood that testing was intended to predict human hazard and was not an end in itself.

In the more than 50 years since this letter was written, additional knowledge has been developed, but the three-segment study designs have remained, although there have been some modifications in them introduced over the years (International Conference on Harmonisation, 2005), and they have been used for evaluation of nonpharmaceutical as well as pharmaceutical chemicals. At present, testing is conducted in one or two laboratory species using intact animals given the test article by gavage, inhalation, dermally, or by subcutaneous or intravenous injection, depending on the anticipated route of human exposure and/or information about the kinetics of the compound. An in vivo protocol commonly used to test for prenatal developmental toxicity (OECD, 2001) assesses malformations, structural variations, resorptions, and fetal growth in litters of pregnant rats or rabbits exposed to a test compound during the period of organogenesis. A vehicle-control and at least three dose levels of the compounds are used, the highest of which produces some minimal adult toxicity and the lowest of which is a low-order multiple of the anticipated human exposure level. Evaluation of pregnancy outcome often involves removal of fetuses about one day prior to delivery and evaluation of external, soft tissue, and skeletal alterations. In other protocols, males and females are dosed prior to mating and conceptuses evaluated after implantation, or pregnant animals are dosed and young are delivered and raised by their mothers or by foster mothers with testing of offspring viability and functional characteristics.

We wondered how we might design protocols for developmental testing of drugs and other chemicals if we were to start over, using 21st century methodology to approach questions about the possible effects of xenobiotics on human development. Would we end up with the same study design involving the dosing of pregnant laboratory animals and evaluation of embryos, fetuses, and pups, or might we adopt a different approach?

Under the auspices of the Developmental and Reproductive Toxicology Technical Committee of the International Life Sciences Institute (ILSI) Health and Environmental Sciences Institute (HESI), we held a workshop in April, 2017 at which this question was explored. Conclusions of the workshop were presented at the 57th annual meeting of the Teratology Society in June, 2017. We here summarize the workshop presentations and discussions. Although the discussions were directed primarily at developmental toxicity testing, application of novel ideas to reproductive toxicity testing can also be considered. We divided the ideas generated in the workshop into those representing an evolution of existing protocols to improve the predictive value of current whole-animal test methods and those representing a revolutionary approach starting from an animal-free testing design. The lines between evolution and revolution can be blurred, and we noted considerable overlap in some areas of discussion.

2 | EVOLUTION

The revolution we discuss below is underway but not ready to replace our original thinking. Whole-animal testing will be important in developmental toxicity assessment for some time to come. Are there ways we can improve the predictivity of whole-animal testing for human health risk assessment?

2.1 | Humanized models

Insofar as we are using rodents and rabbits and sometimes other species to support human risk assessment, can we make our laboratory animal models more like human beings? Humanized models include animals in which genes have been edited, knocked out, or knocked in to make the genetic substrate of the model more human-like while avoiding ethical concerns of human fetal tissue research. RNAs that modify translation of protein also can make an animal model more human. Humanized models include the testing of surrogate molecules that more closely produce in the test species the effects anticipated from the use of the molecule of interest in human patients. For example, an antibody intended to bind and inactivate a human cellular receptor might have no activity on the analogous receptor in a mouse. A mouse in which that receptor has been knocked out might be a suitable model of the activity of the pharmaceutical candidate or, alternatively, an antibody against the mouse receptor might be useful to explore potential developmental effects of a different antibody against the human receptor (Enright et al., 2011).

Challenges of using a humanized model include:

- Time and cost of developing the model;
- Use of a compound in testing that is not the compound intended for marketing;
- Possible increase in use of animals due to the need to complete standard embryo-fetal developmental studies in addition to studies using the humanized model;
- Lack of historical control data in an animal model that has been altered;
- Lack of information on off-target toxicity;
- Alterations in the viability or health of genetically altered animals;
- Lack of regulatory guidance when testing nonpharmaceutical chemicals.

The time and cost of developing a model continues to go down. CrispR/Cas technology has shortened the time and brought the cost to within reach of most academic laboratories, but the characterization of altered animals, which has never been straightforward, has only become more complicated. Apparent inactivation of a gene in a knock-out or otherwise altered animal model might lead to compensatory mechanisms coming online. There may be differences between the complete inhibition of a protein associated with a knock-out and the partial inhibition associated with anticipated human therapy. Differences in gene activity between heterozygotes and homozygotes will require determination of which model is more appropriate for the expected therapeutic scenario.

2.2 | Disease models

Most pharmaceutical products are used in the treatment of human disease. The testing of these compounds in healthy animals may produce developmental toxicity that is irrelevant to the assessment of risk for pregnant women with a disease under treatment. A drug intended to restore normal physiology in a disease may produce altered physiology in healthy pregnant animals. An example is an antidiabetic drug used to make blood glucose concentrations normal in a diabetic woman that might cause hypoglycemia, with consequent developmental toxicity, when given to a healthy pregnant animal (e.g., Hofmann, Horstmann, & Stammberger, 2002). Other pharmaceutical products that might restore normalcy in diseased women but produce toxicity in normal animals include nutrients, hormones, vasoconstrictors, neurotransmitters, and immune modulators.

Healthy animals may not predict developmental effects associated with disease. For example, obesity may enhance susceptibility to developmental toxicity, and the use of normal weight animals to model obese women may fail to identify this enhanced susceptibility. Healthy animals may not express the target for a pharmaceutical product. Examples of missing targets include microorganisms in infection and abnormal proteins in Alzheimer or Parkinson diseases (Barrow et al., 2017).

The use of animal models of disease in testing for reproductive or developmental toxicity has not been addressed in the literature in a systematic manner. It may be useful to consider animal models used for discovery studies as subjects for safety studies, but there are potential problems with the use of diseased animals for these studies:

- Inducing the disease may introduce confounding factors. For example, use of hyperglycemic animals for studies of antidiabetic drugs requires the use of genetically hyperglycemic animals that have not been well characterized or the use of chemicals that destroy pancreatic islet cells (e.g., streptozotocin).
- Increased use of animals may arise from the need to test the drug in healthy models as well as disease models.
- Lack of a historical control database for diseased animals.

- Diseased animals may not reproduce normally or at all.

There is unpublished experience in industry with models of particular human diseases for which there has been abundant product development. It would be useful to gather these experiences in a published forum for consideration by scientists in industry, academia, and government. The development of historical control databases for diseased animals (e.g., diabetic rats) would be useful.

2.3 | Metabolites

The ability to accurately characterize metabolites of test compounds in humans and experimental animals is an important part of safety testing. However, pharmacokinetic data for pregnant women is almost always missing for pharmaceutical products, and human toxicokinetic data are usually missing for nonpharmaceutical chemicals. Moreover, testing of metabolites in pregnant experimental animals is usually conducted at high dose levels, which may not reflect human kinetics. This problem has been addressed to some extent by read-across assessments (discussed more fully in section 3.1, below) using data for structurally similar chemicals or the development of mouse models with humanized livers, either using whole liver or replacement of Phase I and II enzymes. Not all compounds are metabolized by the liver, and use of nonliver cell lines may improve the identification of additional metabolites.

2.4 | Dose-response assessment

Dose setting, including the levels, frequency, and interval of exposure, have a huge impact on dose response assessment. Current developmental toxicity study designs are conducted without attention to critical windows of development; therefore, the impact of exposure is considered to be uniform across the entirety of development. There are cases when suboptimal dose timing can miss a critical window and give a false negative result; for example, use of a single bolus dose of a chemical with rapid metabolism could preclude exposure over the critical window. However, a number of data types, including clinical studies (pharmaceuticals) or human exposure studies (chemicals), in silico information, and pharmacokinetic/pharmacodynamic (or toxicokinetic/toxicodynamic) data, can be used to inform the optimal time or critical windows of exposure, which in turn can be incorporated into dose setting.

Another aspect of dose setting is dose level selection, which is typically based on adult systemic toxicity studies or on a limit dose. However, dose range-finding data in the appropriate population (e.g., pregnant dams or offspring) is critical to establishing dose levels for developmental toxicity studies. Further, dose range finding should approximate the 10% effect level for toxicity and should not include doses that lead to excessive maternal toxicity (reviewed in Beyer et al., 2011). In some regulatory systems, an upper level for testing of 1000 mg/kg/day is used for nonpharmaceutical chemicals, given the low likelihood that a human dose would approach this level. Data-driven dose setting involving better understanding of toxicokinetic properties provides a biological basis for dose setting that often will give a lowered upper limit dose for a given compound without a loss of confidence in the safety of the testing strategy (Saghir et al., 2012).

Current guidelines for dose spacing typically recommend three doses plus an untreated control. Dose-response modeling based on a small number of doses that may not include the point of departure (POD) will not provide adequate information to either determine the shape of the dose-response curve or identify the POD (Bercu, Morinello, Sehner, Shipp, & Weideman, 2016). The use of benchmark dose modeling is preferable to the NOAEL/LOAEL approach, but does not completely alleviate the dose spacing issue if dose levels are not near the POD. The approach to dose selection when testing mixtures presents further challenges; in some instances, a real-world mixture may have set ratios (e.g., crude oil mixture formulations), while, in other cases, a mixture could exist in an infinite variety of dose combinations (e.g., phthalate ester mixtures) in the environment.

Considering enhancements as part of the evolution of the existing testing paradigm, dose setting should be designed based on the available data and hypothesis-testing rather than simply relying on the use of standard practices. In the ideal, dose setting should be based on the internal dose and not the administered dose and incorporate critical window information for the chemical or a mechanistically similar chemical. Dose-response methods should be used that allow for nonlinear curves (e.g., nonmonotonic dose response curves; Chevillotte et al., 2017) because these have been observed for some chemicals with developmental toxicity and for syndromes of mechanistically related endpoints (e.g., the phthalate syndrome; US EPA, 2013).

In an animal-free testing approach, discussed in the next section, a dose-response issue with the use of human cell assays is that while animal to human extrapolation will be removed, in vitro to in vivo data extrapolation will be introduced. Other challenges will be building pathway-based information that considers the impacts of nonchemical stressors, multiple modes of action for a single chemical and cumulative risk assessment for multiple, mechanistically related compounds, and for complex mixtures. Dose-related recommendations for the coming animal-free testing approaches include development of:

- a knowledgebase to predict dose response in whole traditional animal assays from genomics and other pathway-level data, currently being performed in the pharmaceutical sector;

- methods for determining internal dose information in animal-free assessments; and
- in vitro to in vivo dose extrapolation methodologies.

3 | REVOLUTION

3.1 | Hypothesis-driven testing

The current testing environment uses the same models, often pregnant rabbits, rats, and/or mice exposed to a test compound during all or most of gestation with evaluation of fetuses just prior to anticipated delivery, regardless of the compound being evaluated. Yet, prior to developmental toxicity testing, there may be substantial information available about the compound or closely related analogs that could allow us to change the study design to optimize the chance of predicting human developmental toxicity potential. Hypothesis-driven testing is the use of existing information about a chemical to generate hypotheses that could be tested using customized models and protocols. Modifications could be as straightforward as adjusting the dosing regimen so that the internal dosimetry of the chemical is more similar to human pharmacokinetics or choosing a model that is pharmacodynamically more similar to humans. Alternatively, modified testing could entail an entirely different approach that does not involve Segment 2 (embryofetal toxicity testing)-like protocols.

In the drug discovery and development process, much is known about the activity of a compound. The developer usually knows the molecular target of the compound and possible secondary targets, either identified by high-throughput receptor binding/enzyme activity panels or inferred from in vivo safety pharmacology protocols or other toxicology protocols conducted prior to developmental testing. These data are generally not available for nonpharmaceutical chemicals, which (except for pesticides) are not designed to have specific biological activity, but there is still information available to shed light on the possible toxicity of a chemical including its relatedness to previously tested chemicals based on two or three-dimensional structure, physical chemical properties, or with relatively easily generated data on gene expression in a panel of cell types or high-throughput screening. This kind of information can be used to formulate and evaluate hypotheses about the toxicity of a new, related chemical.

In most cases, the hypotheses will involve testing for the mode(s) of action (the critical biological responses, usually occurring at a molecular or subcellular level, that underlie an adverse effect) of the chemical, which will of necessity involve different test methods than the Segment 2 protocol. The Segment 2 protocol, along with every other in vivo testing protocol developed in the mid-20th century, has adverse outcomes as the read-out. These adverse outcomes are the end result of a series of pathogenic events that begin with external exposure and compound-specific distribution in the organism, followed by the interaction of an exogenous chemical with an endogenous molecular target. While we are still a long way from understanding all of the steps in these pathogenic (or adverse outcome) pathways, a considerable amount is known about mode of action at the molecular and cellular level, and there are methods available to test modes of action.

There are several approaches for the prediction of possible molecular initiating events based on structure-activity relationships (Wu et al., 2013), high throughput screening (Judson et al., 2016; Kavlock et al., 2012), or toxicogenomics (De Abrew et al., 2016). Information about key events that result from the molecular initiating event may be less certain, and questions about possible key events will be the subjects of testable hypotheses. These hypotheses might be tested in whole-animal models, but human cells or tissues might offer less expensive and potentially more relevant models for the questions at hand. Once the key events and their quantitative (dose-response) relationships have been characterized, prediction of the adverse outcome may be straightforward based on experience with other compounds operating through the same key events.

The use of analogs for predicting toxicity can be facilitated by large databases that are searchable by chemical structure. There are databases with toxicology-relevant information on hundreds of thousands of chemicals (e.g., EPA ACToR database), including more than 23,000 entries for developmental and reproductive toxicity data. Although consideration of analogs has been used as part of read-across (discussed below) for many years, read-across has been used for the purpose of filling data gaps, and not as a prospective exercise in hypothesis generation and test optimization.

A key component in the development of hypothesis-driven testing is the understanding that there are a finite number of modes of action involved in developmental toxicity. We do not yet know all possible modes of action or the number of pathways that could be involved, but we believe that these pathways are knowable. There are, for example, a limited number of ways in which retinoic acid signaling can be disrupted, and developmental toxicity that is dependent on this signaling could be predicted with targeted testing as discussed below (Tonk, Pennings, & Piersma, 2015). Moreover, toxicity pathways are likely interconnected as part of the physiological network in the body, which should allow for selection and monitoring of a finite number of key events sufficient for establishing the overall toxicological profile of a chemical. We will return to these networks in the next section.

Common modes of action can be inferred in some cases from gene expression. For example, estrogen receptor agonism can be predicted from gene expression and other studies in vitro (Browne, Judson, Casey, Kleinstreuer, & Thomas, 2015; Daston & Naciff, 2010). Decisions can be made about a compound with suspected estrogen-mediated toxicity without whole-animal testing, at least with regard to modes of action for which estrogen receptor agonism is a key event and with regard to potency/efficacy considerations. Confidence in testing of steps in a mode of action will require that the full range of possible modes of action has been considered; data supporting this approach are increasingly available.

Questions remain about the relationship between exposure level and response, particularly when the underlying information about a chemical is based on toxicity data using compounds other than the chemical of interest. Decisions about acceptable levels of exposure will need to incorporate an understanding about similarities and differences in toxicokinetics and disposition in target organs between the compounds, and uncertainty factors will continue to be necessary in decision-making about acceptable exposure levels. The selection of uncertainty factors may be influenced by the source of data, for example, in vivo, in vitro, in silico, or a combination of sources.

Hypothesis-based testing in the pharmaceutical industry considers the pharmacology, toxicology, and clinical use of the drug in designing a testing strategy. Considerations include both the effect of intended pharmacology and off-target effects on embryo-fetal and post-natal development. The intended clinical use informs exposure questions such as whether therapy requires constant engagement with the target, for example by a dermal drug that needs to be present 24 hr/day or an antibiotic that needs to maintain a minimum systemic concentration.

Figure 1 shows a case example illustrating the pharmacokinetic profile of a potential therapeutic agent being developed for a dermatological indication (Stanislaus D, unpublished). Because systemic exposure is necessary to understand the hazard potential for a developing embryo, the compound was given subcutaneously to rabbits twice/day and four times/day. The same dose spread over 24 hr/day produced developmental toxicity that was not evident when exposure was present for only half the day. Due to the unexpected developmental toxicity, further investigation of the target was conducted and it was found that microdeletions of the chromosomal segments that contained this target produced craniofacial malformations in humans. Adapting a standard developmental toxicity study to test a hypothesis about the kinetic requirements for toxicity allowed better understanding of the hazard potential, leading to better decision making.

Nonpharmaceutical chemicals differ from pharmaceutical products in that biological activity in humans is unintended and physical performance or activity against insects, fungi, or rodents is the primary focus for most of these compounds. Physical hazards (flammability, explosive limits), low vapor concentrations, and solubility can be challenges in toxicity testing. Poorly characterized mixtures or complex reaction products may show variability, and it may not be easy to show that a material tested for toxicity in the lab is identical to the marketed product (Daston et al., 2015).

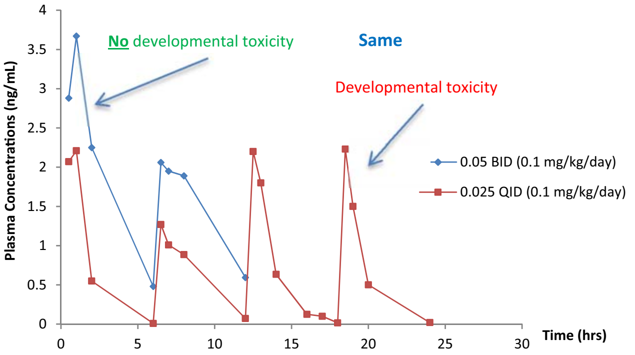

FIGURE 1 Potential therapeutic agent given twice/day (6 hr apart) or four times/day (6 hr apart) to rabbits. The same total daily dose was used in both exposure scenarios, but developmental toxicity occurred only with four times/day dosing. From Slanislaus D, unpublished.

The European Chemicals Agency (ECHA) supports read-across, a procedure in which gaps in toxicology knowledge about a compound may be filled using data from related compounds based on biotransformation to common compounds or similar properties of the related compounds (ECHA, 2017). Read-across may reduce the need to test a compound in whole animals, provided there are adequate data on a sufficiently similar compound or series of compounds. Read-across can be similar to the use of Quantitative Structure-Activity Relationships (QSAR) in inferring activity of a compound from structural similarity to another compound or series of compounds for which data are available.

The need for additional testing to characterize endpoint specific hazards also can be assessed based on estimated human exposures being below a threshold of toxicological concern (TTC; van Ravenzwaay, 2011, 2012, 2017; Kroes et al., 2004). The TTC method empirically derived a distribution of effect levels for maternal and developmental toxicity for a large number of tested chemicals. Application of a safety factor to the 5th percentile effect level gives an exposure level below which adverse effects are unlikely for any compound, tested or not. A related TTC scheme used structural chemical alerts to set TTCs for reproductive and developmental toxicity (Laufersweiler et al., 2012). Compounds with low alerts had a TTC value of 131 μg/kg bw/day, and those with high alerts had a TTC value of 3.1 μg/kg bw/day. Using this method, estimated human exposures below these levels would not require testing in whole animals.

In some regulatory systems, an upper level for testing of 1000 mg/kg/day is used for nonpharmaceutical chemicals, given the low likelihood that a human dose would approach this level. Hypothesis-driven testing involving kinetics could explore the question of whether this limit dose could be lowered for a given compound without a loss of confidence in the safety of the testing strategy (Saghir et al., 2012).

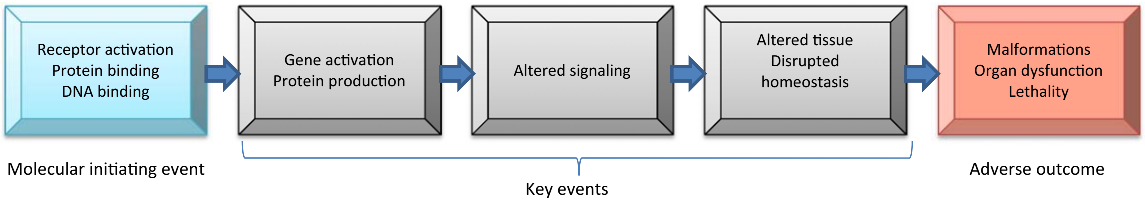

FIGURE 2 Adverse outcome pathway describing key events leading from a molecular initiating event to an adverse outcome

Complex substances or mixtures can be grouped using known toxicological characteristics of tested chemicals within the mixtures or results from testing similar chemicals/mixtures. A collection of compounds or mixtures representing worst-case scenarios could be developed to stand for related materials that are predicted to have lower toxicity potential. For example, a mixture of aliphatic compounds with varying chain lengths could be represented by data for a single compound within the mixture that is believed to have the greatest toxicity potential (McKee, Nicolich, Roy, White, & Daughtrey, 2014).

3.2 | Toxicological pathway networks

Although a standard approach to considering adverse outcome pathways assumes a linear, unidirectional scheme such as shown in Figure 2, biological systems are more likely to be complex, and adverse outcome pathways are interconnected in multiple ways and directions (Browne, Noyes, Casey, & Dix, 2017; Wittewehr et al., 2017; Figure 3). If we envision all possible interconnected pathways involved in development, we would get a very busy diagram. It is likely, although that along the pathways leading to an outcome, certain steps would be particularly important, like rate-limiting steps in a complex series of biochemical reactions. Figure 4 gives an example of how such a network might be displayed with six key events, identified as stars, representing important steps mediating toxicity. The development of advanced computational methods and experimental models is required for analyzing and integrating data generated for this purpose, although it might not be necessary to know all the steps in the complex pathways leading to development. Instead, it might be sufficient to be able to test for the ability of a chemical exposure to disrupt any of the six key steps. In this way, a battery of six tests in an alternative model, perhaps in vitro or in silico, would predict potential developmental effects of a chemical, at least with respect to the outcomes shown. The predictivity of testing a limited number of key steps would need to be empirically verified.

The role of retinoic acid in embryo development gives rise to an example of a network of adverse outcome pathways (Tonk et al., 2015). Retinoic acid is involved in cortical neurogenesis, mediated by *Wnt3a* and in progenitor cell proliferation, mediated by *Fgf8*. Neural tube patterning and axial

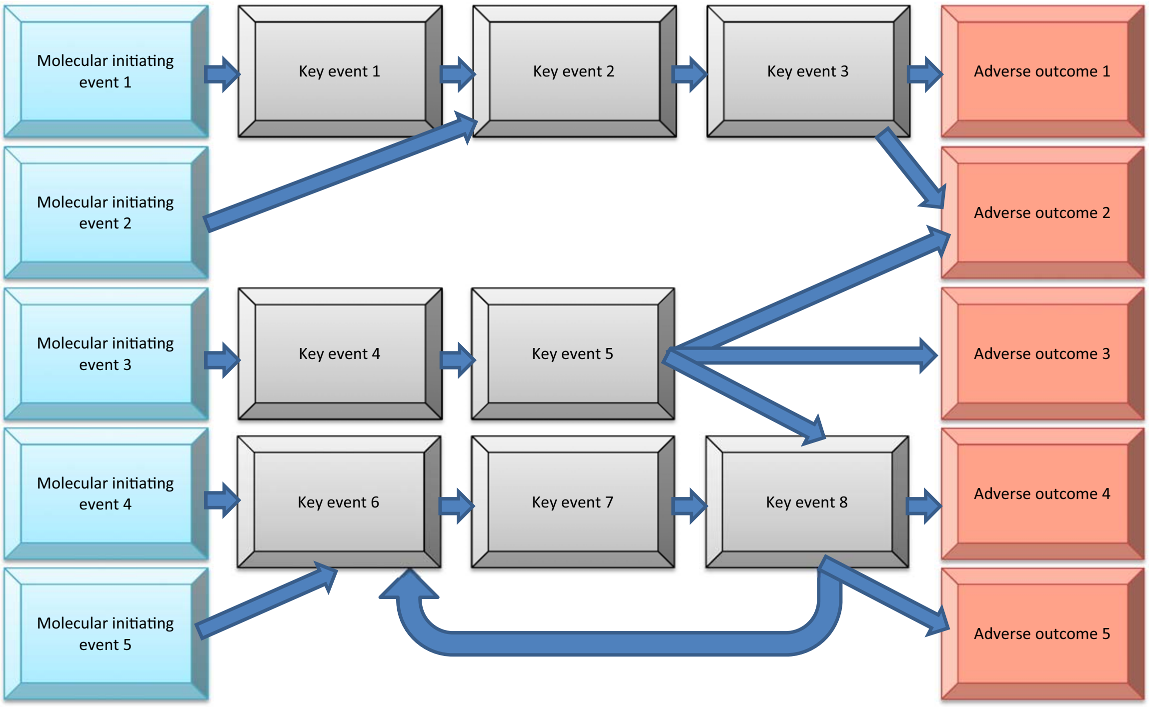

FIGURE 3 Interconnected adverse outcome pathways

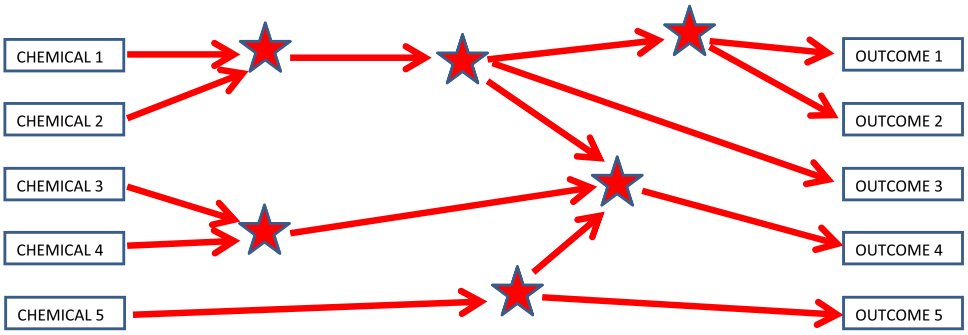

FIGURE 4 Toxicological pathway network showing six key events (stars) leading to outcomes

pattern are mediated by additional families of genes. Excess retinoic acid can be associated with heart defects, cleft palate, anencephaly, caudal regression, craniofacial, and limb defects, mediated by down-regulation of *Dhrs3* and *Cyp26a1*, *26b1*, and *26c1*. Insufficient retinoic acid has been associated with craniofacial, cardiac, and limb malformations associated with down-regulation of *Rdh10* and *Raldh2*. These relationships can be mapped as in Figure 4, giving a multistep, interconnected network of processes that can be evaluated in experimental systems that do not necessarily involve intact mammalian organisms.

We have not arrived at a place where we can perform testing of all developmental pathways at this level of detail, but we are getting an idea of what that place will look like. Table 1 is a list of the steps that will be necessary before we are ready to reliably model human developmental toxicity without the use or with only limited use of whole-animal models.

TABLE 1 Steps needed to reliably model human developmental toxicity

- Map human developmental physiology from the molecular to the organism level
  - Aim at level of detail fit for the purpose of toxicity testing
- Integrate existing chemistry and toxicity knowledge
  - Identify the major modes of action of human developmental toxicity
  - Map the integrated adverse outcome pathways for the purpose of toxicity testing
  - Identify rate-limiting Key Events (the stars in Figure 4) and related biomarkers
  - Design biomarker-related test systems
- Build computational tools for toxicity prediction
  - Integrate quantitative test output into adverse outcome network model
  - Define thresholds of adversity at the integrated model level
- Embed toxicodynamic model within overall risk assessment
  - Consider use patterns and expected exposure scenarios
  - Model external to internal exposure – target organ concentration modeling
  - Consider timing, duration, and life cycle segment(s) of exposure
- Design flexible compound-dependent case-by-case fit-for-purpose testing strategy

3.3 | In vitro and in silico approaches for predictive toxicology

Predicting human developmental risk could in theory be based entirely on the testing of human cells or tissues or the in silico manipulation of models of human development. Such a strategy was generally envisioned by the National Research Council in the 2007 report, *Toxicity Testing in the 21st Century* (National Research Council, 2007), and in the intervening decade, we have come much closer than expected to making such testing a practical reality.

The development of predictive models uses a large amount of data such as has been developed by the ToxCast^{TM} and Tox21 efforts (US EPA, 2017a,b). ToxCast^{TM} has evaluated more than 1000 chemicals using multiple assays, and Tox21 includes data on many more chemicals using fewer assays. The results of high throughput assays of biochemical activities of hundreds of ToxCast^{TM} chemicals have been published, and the data are publically available (Knudsen et al., 2011; Judson et al., 2016; Richard et al., 2016; https://www.epa.gov/chemical-research/toxcast-dashboard). The evaluation of developmental pathways for a subset of these chemicals has been performed using ToxCast data and alternative models including zebrafish embryos and human embryonic stem cells, providing an estimate of concentration-related toxicity of these compounds in human developmental systems (Kleinstreuer et al., 2011; Palmer et al., 2013; Sipes et al., 2011).

Organotypic culture models under development will give rise to modeling that is more physiological than can be achieved with conventional monolayer cell culture. Organotypic models represent a three-dimensional framework of the embryo and can be used for human cell-based recapitulation

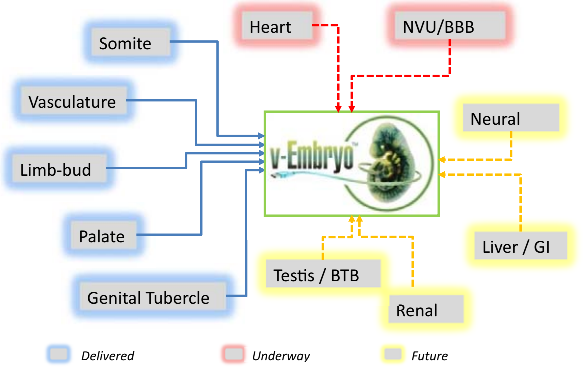

FIGURE 5 The predictive virtual embryo. Computer models that have been delivered include somites, vasculature, limb-bud, palate, and genital tubercle. Models that are underway include heart and neurovascular unit/blood-brain barrier (NVU/BBB). Future models include neural tube, liver/gastrointestinal (GI) tract, testis/blood–testis barrier (BTB), and renal development.

of key morphoregulatory pathways culminating in controlled tissue fusion events, heterotypic signaling, epithelial–mesenchymal transition, vascularization, and biomechanical forces (Knight, Sha, & Ashton, 2015; Knudsen, Klieforth, & Slikker, 2017). Such models may supplant mammalian animal studies focused on apical endpoints in favor of functional or mechanistic outcomes in engineered micro-tissues and microphysiological systems that inform processes around some of the stars in Figure 4.

Experimental models that reduce a complex biological system to simpler assays have the benefit of facilitating quantitative evaluation of cellular and molecular responses to chemical perturbation but at the drawback of eliminating the cellular interactions and spatial dynamics that make an embryo complex in the first place. When modeling developmental processes and the toxicity that disrupts them, we need to rebuild this complexity. Using computational methods, we can rebuild the complexity cell-by-cell and interaction-by-interaction. CompuCell3D, funded by NIH and EPA (http://www.compucell3d.org/), is an open-source modeling environment in which cells interact stochastically using rules governing individual cell behaviors. Using this platform, a computer model of somite development has been explored with and without the traditional clock-and-wavefront mechanism (Dias, de Almeida, Belmonte, Glazier, & Stern, 2014; Hester, Belmonte, Gens, Clendenon, & Glazier, 2011). Other embryological events that have been similarly modeled in dynamical computer simulations including urethral fusion during sexual diversification of the genital tubercle (Leung, Hutson, Seifert, Spencer, & Knudsen, 2016) and fusion of the secondary palatal processes (Hutson, Leung, Baker, Spencer, & Knudsen, 2017). Manipulation of these dynamic computer models can simulate exposures that interfere with development; for example, the effects of perturbing the androgen-dependent growth of the genital tubercle (Leung et al., 2016) or the *TGF*/*EGF* switch that controls fusion of palatal shelves (Hutson et al., 2017) can be used to predict critical effects of chemical exposures.

An ultimate goal of integrative computational modeling efforts is a predictive virtual embryo (Figure 5), a computer system that will represent our understanding of embryogenesis and permit the investigation of the singular or combinatorial effects of perturbing components of human development (Knudsen et al., 2017; https://www.epa.gov/chemical-research/virtual-tissue-models-predicting-how-chemicals-impact-development). Challenges will include identifying all relevant systems that need to be modeled (for example, utero-placental physiology) and identifying methods of determining the reversibility of modeled adverse effects.

Using the results of modeling in regulatory implementation will require consideration of diversity at the genetic, cellular, and organism level. Best-practice and global harmonization guidelines will be important, and acceptance by regulatory scientists will be important (Zaunbrecher et al., 2017). The transition from a system that relies on whole-animal testing to a system that is largely independent of whole-animal testing will need to be carefully managed. The expectation of perfection can mean that the revolution will fail, and we need to decide how we will know that the model is good enough to approximate reality.

4 | CONCLUDING REMARKS

Workshop participants were successful at identifying concerns about and limitations of the existing protocols used in reproductive and developmental toxicity testing. The consensus was that the evolutionary means to reduce the concerns and minimize the limitations, such as the use of humanized models or disease models, were temporary measures that would not be entirely satisfactory. A revolutionary transition to the use of human cells, tissue, or computer simulations of human development was generally seen as inevitable, but careful development and validation of new methods of risk assessment would require time and resources to increase confidence in the ability to replace whole mammal studies and improve human risk assessment. There was general optimism, however, that we would be successful in replacing the automatic use of whole animals with scientifically justified and thoughtful developmental toxicity testing strategies, perhaps in our own lifetimes, but that the revolution might be based on incremental, evolutionary change rather than on a cataclysmic retooling of the procedures we use.

DISCLAIMER

The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.

S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ORCID

Anthony R. Scialli http://orcid.org/0000-0003-1351-4572

REFERENCES

Barrow, P., Villabruna, L., Hoberman, A., Bohrmann, B., Richter, W. F., & Schubert, C. (2017). Reproductive and developmental toxicity studies with ganternerumab in PS2APP transgenic mice. *Reproductive Toxicology*, *73*, 362–371.

Bercu, J. P., Morinello, E. J., Sehner, C., Shipp, B. K., & Weideman, P. A. (2016). Point of departure (PoD) selection for the derivation of acceptable daily exposures (ADEs) for active pharmaceutical ingredients (APIs). *Regulatory Toxicology and Pharmacology*, *79* (Suppl 1), S48–S56.

Beyer, B. K., Chernoff, N., Danielsson, B. R., Davis-Bruno, K., Harrouk, W., Hood, R. D., . . . Scialli, A. R. (2011). ILSI/HESI maternal toxicity workshop summary: maternal toxicity and its impact on study design and data interpretation. *Birth Defects Research. Part B, Developmental and Reproductive Toxicology*, *92*(1), 36–51.

Browne, P., Judson, R. S., Casey, W. M., Kleinstreuer, N. C., & Thomas, R. S. (2015). Screening chemicals for estrogen receptor bioactivity using a computational model. *Environmental Science & Technology*, *49*(14), 8804–8814.

Browne, P., Noyes, P. D., Casey, W. M., & Dix, D. J. (2017). Application of adverse outcome pathways to U.S. EPA's endocrine disruptor screening program. *Environmental Health Perspectives*, *125*(9), 096001. https://doi.org/10.1289/EHP1304

Chevillotte, G., Bernard, A., Varret, C., Ballet, P., Bodin, L., & Roudot, A. C. (2017). Probabilistic assessment method of the non-monotonic dose-responses-Part I: Methodological approach. *Food and Chemical Toxicology*, *106*(Pt A), 376–385.

Daston, G. P., & Naciff, J. M. (2010). Predicting developmental toxicity through toxicogenomics. *Birth Defects Research. Part C, Embryo Today*, *90*(2), 110–117.

Daston, G., Knight, D. J., Schwarz, M., Gocht, T., Thomas, R. S., Mahony, C., & Whelan, M. (2015). SEURAT: Safety evaluation ultimately replacing animal testing—Recommendations for future research in the field of predictive toxicology. *Archives of Toxicology*, *89*(1), 15–23.

De Abrew, K. N., Kainkaryam, R. M., Shan, Y. K., Overmann, G. J., Settivari, R. S., Wang, X., . . . Daston, G. P. (2016). Grouping 34 chemicals based on mode of action using connectivity mapping. *Toxicological Sciences*, *151*(2), 447–461.

Dias, A. S., de Almeida, I., Belmonte, J. M., Glazier, J. A., & Stern, C. D. (2014). Somites without a clock. *Science*, *343*(6172), 791–795.

ECHA (2017). Read-Across Assessment Framework. European Chemicals Agency, Helsinki, Finland, Retrieved from https://echa.europa.eu/documents/10162/13628/raaf_en.pdf, last accessed 7 July 2017.

Enright, B. P., Davila, D. R., Tornesi, B. M., Blaich, G., Hoberman, A. M., & Gallenberg, L. A. (2011). Developmental and reproductive toxicology studies in IL-12p40 knockout mice. *Birth Defects Research (Part B)*, *92*(2), 102–110.

Hester, S. D., Belmonte, J. M., Gens, J. S., Clendenon, S. G., & Glazier, J. A. (2011). A multi-cell, multi-scale model of vertebrate segmentation and somite formation. *PLoS Computational Biology*, *7*(10), e1002155.

Hofmann, T., Horstmann, G., & Stammberger, I. (2002). Evaluation of the reproductive toxicity and embryotoxicity of insulin glargine (LANTUS) in rats and rabbits. *International Journal of Toxicology*, *21*(3), 181–189.

Hutson, M. S., Leung, M. C., Baker, N. C., Spencer, R. M., & Knudsen, T. B. (2017). Computational model of secondary palate fusion and disruption. *Chemical Research in Toxicology*, *30*(4), 965–979.

International Conference on Harmonisation (ICH) (2005). ICH Harmonised Tripartite Guideline. Detection of toxicity to reprotoduction for medicinal products & toxicity to male fertility S5(R2). Retrieved from http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Safety/S5/Step4/S5_R2__Guideline.pdf

Judson, R., Houck, K., Martin, M., Richard, A. M., Knudsen, T. B., Shah, I., . . . Thomas, R. S. (2016). Analysis of the effects of cell stress and cytotoxicity on in vitro assay activity in the toxcast dataset. *Toxicological Sciences*, *152*, 323–329.

Kavlock, R., Chandler, K., Dix, D., Houck, K., Hunter, S., Judson, R., . . . Sipes, N. (2012). Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. *Chemical Research in Toxicology*, *25*(7), 1287–1302.

Kleinstreuer, N. C., Smith, A. M., West, P. R., Conard, K. R., Fontaine, B. R., Weir-Hauptman, A. M., . . . Cezar, G. G. (2011). Identifying developmental toxicity pathways for a subset of ToxCast chemicals using human embryonic stem cells and metabolomics. *Toxicology and Applied Pharmacology*, *257*(1), 111–121.

Knight, G. T., Sha, J., & Ashton, R. S. (2015). Micropatterned, clickable culture substrates enable *in situ* spatiotemporal control of human PSC-derived neural tissue morphology. *Chemical Communications*, *51*(25), 5238–5241.

Knudsen, T. B., Houck, K. A., Sipes, N. S., Singh, A. V., Judson, R. S., Martin, M. T., . . . Kavlock, R. J. (2011). Activity profiles of 309 ToxCast^{TM} chemicals evaluated across 292 biochemical targets. *Toxicology*, *282*(1–2), 1–15.

Knudsen, T. B., Klieforth, B., & Slikker, W. Jr. (2017). Programming microphysiological systems for children's health protection. *Experimental Biology and Medicine*, *242*(16), 1586–1592.

Kroes, R., Renwick, A. G., Cheeseman, M., Kleiner, J., Mangelsdorf, I., Piersma, A., . . . Würtzen, G., European branch of the International Life Sciences Institute. (2004). Structure-based thresholds of toxicological concern (TTC): Guidance for application to substances present at low levels in the diet. *Food and Chemical Toxicology*, *42*(1), 65–83.

Laufersweiler, M. C., Gadagbui, B., Baskerville-Abraham, I. M., Maier, A., Willis, A., Scialli, A. R., . . . Daston, G. (2012). Correlation of chemical structure with reproductive and developmental toxicity as it relates to the use of the threshold of toxicological concern. *Regulatory Toxicology and Pharmacology*, *62*(1), 160–182.

Leung, M. C. K., Hutson, M. S., Seifert, A. W., Spencer, R. M., & Knudsen, T. B. (2016). Computational modeling and simulation of genital tubercle development. *Reproductive Toxicology*, *64*, 151–161.

McKee, R. H., Nicolich, M., Roy, T., White, R., & Daughtrey, W. C. (2014). Use of a statistical model to predict the potential for repeated dose and developmental toxicity of dermally administered crude oil and relation to reproductive toxicity. *International Journal of Toxicology*, *33*(1_suppl), 17S–27S.

National Research Council of the National Academies (2007). *Toxicity Testing in the 21st Century*. Washington DC: National Academies Press. Retrieved from https://www.nap.edu/download/11970

OECD. (2001). Guideline for the testing of chemicals. TG414. *Prenatal Developmental Toxcity Study*, Retrieved from http://www.oecd-ilibrary.org/environment/test-no-414-prenatal-development-toxicity-study_9789264070820-en.

Palmer, J. A., Smith, A. M., Egnash, L. A., Conard, K. R., West, P. R., Burrier, R. E., … Kirchner, F. R. (2013). Establishment and assessment of a new human embryonic stem cell-based biomarker assay for developmental toxicity screening. *Birth Defects Research. Part B, Developmental and Reproductive Toxicology*, *98*(4), 343–363.

Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., … Thomas, R. S. (2016). The ToxCast chemical landscape: Paving the road to 21st century toxicology. *Chemical Research in Toxicology*, *29*(8), 1225–1251.

Saghir, S. A., Bartels, M. J., Rick, D. L., McCoy, A. T., Rasoulpour, R. J., Ellis-Hutchings, R. G., … Bus, J. S. (2012). Assessment of diurnal systemic dose of agrochemicals in regulatory toxicity testing—An integrated approach without additional animal use. *Regulatory Toxicology and Pharmacology*, *63*(2), 321–332.

Sipes, N. S., Kleinstreuer, N. C., Judson, R. S., Reif, D. M., Martin, M. T., Singh, A. V., … Knudsen, T. B. (2011). Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data. *Toxicological Sciences*, *124*(1), 109–127.

Tonk, W. E. C., Pennings, J. L. A., & Piersma, A. H. (2015). An adverse outcome pathway framework for neural tube and axial defects mediated by modulation of retinoic acid homeostasis. *Reproductive Toxicology*, *55*, 104–113.

US EPA (2013). Nonmonotonic dose responses as they apply to estrogen, androgen, and thyroid pathways and EPA testing and assessment procedures. Retrieved from https://www.epa.gov/chemical-research/nonmonotonic-dose-responses-they-apply-estrogen-androgen-and-thyroid-pathways-and.

US EPA (2017a). Toxicity ForeCaster (ToxCast$^{\text{TM}}$) Data. Retrieved from https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data.

US EPA (2017b). Toxicology in the 21$^{\text{st}}$ Centurey. Retrieved from https://www.epa.gov/chemical-research/toxicology-testing-21st-century-tox21

van Ravenzwaay, B., Dammann, M., Buesen, R., & Schneider, S. (2011). The threshold of toxicological concern for prenatal developmental toxicity. *Regulatory Toxicology and Pharmacology*, *59*(1), 81–90.

van Ravenzwaay, B., Dammann, M., Buesen, R., Flick, B., & Schneider, S. (2012). The threshold of toxicological concern for prenatal developmental toxicity in rabbits and a comparison to TTC values in rats. *Regulatory Toxicology and Pharmacology*, *64*(1), 1–8.

van Ravenzwaay, B., Jiang, X., Luechtefeld, T., & Hartung, T. (2017). The threshold of toxicological concern for prenatal developmental toxicity in rats and rabbits. *Regulatory Toxicology and Pharmacology*, *88*, 157–172.

Wittwehr, C., Aladjov, H., Ankley, G., Byrne, H. J., de Knecht, J., Heinzle, E., … Whelan, M. (2017). How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology. *Toxicological Sciences*, *155*(2), 326–336.

Wu, S., Fisher, J., Naciff, J., Laufersweiler, M., Lester, C., Daston, G., & Blackburn, K. (2013). Framework for identifying chemicals with structural features associated with the potential to act as developmental or reproductive toxicants. *Chemical Research in Toxicology*, *26*(12), 1840–1861.

Zaunbrecher, V., Beryt, E., Parodi, D., Telesca, D., Doherty, J., Malloy, T., & Allard, P. (2017). Has toxicity testing moved into the 21st century? A survey and analysis of perceptions in the field of toxicology. *Environmental Health Perspectives*, *125125*(8), 087024.

How to cite this article: Scialli AR, Daston G, Chen C, et al. Rethinking developmental toxicity testing: Evolution or revolution? *Birth Defects Research*. 2018;00:1–11. https://doi.org/10.1002/bdr2.1212