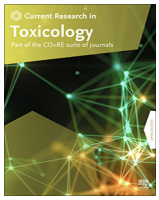
Current Research in Toxicology 3 (2022) 100074

Contents lists available at ScienceDirect

Current Research in Toxicology

journal homepage: www.elsevier.com/locate/crtox



Pluripotent stem cell assays: Modalities and applications for predictive developmental toxicity

Aldert H. Piersma [a,*], Nancy C. Baker [b], George P. Daston [c], Burkhard Flick [d], Michio Fujiwara [e], Thomas B. Knudsen [f], Horst Spielmann [g], Noriyuki Suzuki [h], Katya Tsaioun [i], Hajime Kojima [j]

[a] Center for Health Protection, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
[b] Leidos, Research Triangle Park, NC, USA
[c] Global Product Stewardship, The Procter & Gamble Company, Cincinnati, OH, USA
[d] Experimental Toxicology and Ecology, BASF SE, Ludwigshafen am Rhein, Germany
[e] Drug Safety Research Labs, Astellas Pharma Inc., Tsukuba-shi, Japan
[f] Center for Computational Toxicology and Exposure, U.S. Environmental Protection Agency, Research Triangle Park, USA
[g] Institute for Pharmacy, Faculty of Biology, Chemistry, and Pharmacy, Freie Universität, Berlin, Germany
[h] Cell Science Group Environmental Health Science Laboratory, Sumitomo Chemical Co., Ltd., Osaka, Japan
[i] Evidence-Based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[j] National Institute of Health Sciences, Kawasaki, Japan

ARTICLE INFO

Keywords:
Developmental toxicity
Predictive toxicology
Embryonic stem cell test
Teratogenesis

ABSTRACT

This manuscript provides a review focused on embryonic stem cell-based models and their place within the landscape of alternative developmental toxicity assays. Against the background of the principles of developmental toxicology, the wide diversity of alternative methods using pluripotent stem cells developed in this area over the past half century is reviewed. In order to provide an overview of available models, a systematic scoping review was conducted following a published protocol with inclusion criteria, which were applied to select the assays. Critical aspects including biological domain, readout endpoint, availability of standardized protocols, chemical domain, reproducibility and predictive power of each assay are described in detail, in order to review the applicability and limitations of the platform in general and progress moving forward to implementation. The horizon of innovative routes of promoting regulatory implementation of alternative methods is scanned, and recommendations for further work are given.

General introduction

Regulatory test guidelines for developmental toxicity

Developmental toxicology has long been considered an area of specific interest in the safety assessment of chemicals, pharmaceuticals, crop protection products and biocides, especially since the thalidomide episode in the early 1960s. This experience prompted the need for premarketing safety testing for developmental toxicity, which was initially focused on using rodent and nonrodent small mammals such as mice, rats and rabbits, to predict human safety. Worldwide regulatory animal study protocol guidelines were established at the Organisation for Economic Co-operation and Development (OECD) since the early 1980s and have been updated since as deemed necessary (https://www.oecd.org/chemicalsafety/testing/oecd-guidelines-testing-chemicals-related-documents.htm, accessed November 2021). OECD test guidelines (TG) for the developmental/reproductive toxicity of chemicals include the Prenatal Developmental Toxicity Study (TG 414), the Two-Generation Reproduction Toxicity Study (TG 416), the Reproduction/Developmental Toxicity Screening Test (TG 421), the Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test (TG 422), the Developmental Neurotoxicity Study (TG 426) and the Extended One-Generation Reproductive Toxicity Study (TG 443). Test guidelines on *in vitro* endocrine screening such as TGs 455, 456, 458 and 493, and the *in vivo* uterotrophic assays (TG 440) and Hershberger assay (TG 441) are also associated with assessing developmental/reproductive toxicity.

* Corresponding author.
E-mail address: aldert.piersma@rivm.nl (A.H. Piersma).

https://doi.org/10.1016/j.crtox.2022.100074
Received 7 February 2022; Revised 21 April 2022; Accepted 9 May 2022

The embryonic stem cell test (EST) as an alternative test for predictive toxicology

In vitro profiling of drugs and chemicals utilizing biomarker responses of pluripotent stem cell (PSC) lines has been an active area of investigation and one of the most promising alternatives to pregnant animal testing. As collective experience increased with the use of PSC lines, derived either from mouse embryos (mESCs) or as established human cell lines, the strengths and limitations as to predicting human safety have become apparent. Concomitantly, the reduction, refinement and replacement of experimental animals for toxicity testing (3Rs) is a major driver of the need for new approach methods (NAMs) to reliably identify developmental hazards and characterize their risk to healthy human pregnancy outcomes (Knudsen et al., 2021). The development of alternative assay platforms have led to a wide variety of platforms based on simple model organisms classified in both non-vertebrate and vertebrate species, tissue and organ cultures, and cultures with primary cells and human cell lines (Brown, 1987; Piersma, 2006).

The initial embryonic stem cell test (EST) was based on the inhibition by test compounds of cardiac muscle cell differentiation as observed by scoring contracting foci under the light microscope (Scholz et al., 1999). This readily observed endpoint is a consequence of heterotypic interactions between different germ layers during the cellular progression from pluripotency to differentiation (cardiopoiesis). Furthermore, a variety of differentiation routes and other readout systems have been assessed that broaden the applicability of EST to compound testing in PSC lines from diverse species, including biological domain, readout endpoint, availability of standardized protocols, chemical applicability domain, reproducibility and predictive power (Hartung et al., 2004). Here, we review the spectrum of EST assays currently available for *in vitro* testing, with specific focus on those assays for which validation studies have been published.

Relevance to principles of developmental toxicology

Apical endpoints

The purpose of developmental toxicity testing is to evaluate the developmental hazard potential for an agent (drug, chemical, or physical stressor) to adversely affect pregnancy outcome following maternal (or parental) exposure. Clinically, this means changes in embryonic development that lead to structural malformations, growth retardation, death of the developing organism, and/or functional deficits in offspring (Friedman, 2010). Modifications or additional tests may be needed that are sensitive to functional deficits, particularly developmental neurotoxicity (Scialli et al., 2018). Apical endpoints from regulatory studies may be analogous or homologous to those observed in human populations, so their relevance and applicability for predicting outcome in humans is high. However, a battery of *in vitro* assays may be needed for sufficient biological coverage of key developmental pathways and processes. Assays employing ESCs show many of the upstream pathways that intersect with more apical adverse endpoints of regulatory concern and could therefore be predictive of developmental toxicity potential observed in animal studies or human populations. Since differentiation is a key process in the formation of every structure in the embryo, it is reasonable to assume that differentiation of stem cells is relevant for human development. The unique features of ESC lines that make them valuable additions to an *in vitro* battery are (i) pluripotency (the capacity to give rise to most cells of the embryo-fetus), (ii) self-renewal (ESC lines can be maintained for extended periods in culture, and (iii) autopoiesis (self-organizing capacity to form rudimentary tissues and organoids) (Martello and Smith, 2014). Validation studies that compare *in vitro* with *in vivo* results for data-rich drugs and chemicals, such as retinoids, chemotherapeutic agents, and pesticides provide support that the readouts being measured in EST assays are relevant for regulatory toxicology (Genschow et al., 2004; Chapin et al., 2007).

Exposure-based modeling

A central tenet of developmental toxicology (and of all toxicology) is that a threshold concentration exists for an exposure above which adverse effects are produced, and the prevalence and degree of effects increases with increasing concentration. Therefore, in validation studies it is important not just to determine whether the test agent produces an effect, but at what concentration. Most validation studies present their data as positive or negative; a recent publication (Daston et al., 2014) provides a list of exposures (compound plus dosimetric) that can be used for validation. In many cases, the same compound can be both a positive or a negative depending on the concentration at which it is tested.

Bioavailability

The *in vivo* teratogenic potential of a compound is influenced by ADME (absorption, distribution, metabolism, excretion) kinetics in the mother-conceptus. In some cases, it is the metabolite of a chemical that is the proximate toxicant, in which case the chemical must either be bioactivated by the embryo-placenta or distributed to the embryo after maternal metabolism. *In vitro*, the test compound is directly added to the culture media. A nominal concentration tested in the absence of ADME requires *in vitro* to *in vivo* extrapolation in order to appropriately evaluate the developmental hazard. This common weakness of *in vitro* toxicity assays may be partially circumvented by characterizing the endogenous xenobiotic metabolic capacity of the cultured cells or supplementation by an exogenous metabolizing system. There are also some physical characteristics of chemicals that can make testing *in vitro* a challenge or may interfere with the readouts (e.g., volatility, fluorescence, solubility and other pharmacokinetic characteristics (Judson et al., 2013).

Species susceptibility

One of the principal uncertainties of developmental toxicity studies in animals is the extent to which the model is relevant in its susceptibility and reproducibility to the agent being tested. The basis for species differences are pharmacokinetic and/or pharmacodynamic (Toutain et al., 2010). While kinetic variables include explicit differences in the rate and extent of ADME, the species-dependent variables in dynamic response such as drug/chemical binding to receptors or enzymes and subsequent stress-response pathways in the culture models are implicitly assumed to be the same as for *in vivo* models (Zeng et al., 2020; Wang et al., 2021). This may provide a predictive advantage for human over rodent models, and furthermore for the use of induced pluripotent stem cells (iPSCs) from different human donors to assess the potential for individual variability.

Stage vulnerability

Embryonic development is characterized by rapid cell proliferation, formation and patterning of body axes, and differentiation of cells into specialized forms that perform specific functions in tissues and organs that usually have a highly organized architecture (Xing et al., 2015; Warkus and Marikawa, 2017). Each organ/structure has its own timetable for formation, differentiation, and susceptibility to drug and chemical disruption (Scialli et al., 2018). Stem cells undergo a pattern of differentiation that is dependent on factors added to the culture media. Although cardiomyocyte differentiation is a readily observed endpoint in the mouse EST (mEST) other differentiated cell types may form spontaneously or by specific growth conditions (zur

Nieden et al., 2001; Suzuki et al., 2011; van Dartel and Piersma, 2011; Le Coz et al., 2015; Kameoka et al., 2014). The assay systems that rely on embryoid body formation may allow for differentiation into multiple cell types within the same assay system (Marikawa et al., 2020). The time period over which differentiation occurs in these assays is different from the time course for embryonic development; however, the test compound is present over the entire time course of the *in vitro* exposure, so it is the opinion of the authors that it covers episodic vulnerability across a range of pathways relevant to human exposure.

Initiating mechanisms

One of the important principles of toxicology is that there are specific mechanisms of action by which chemicals interact with the biological system to produce adverse effects. A considerable amount of research has gone into identifying the targets and initiating mechanisms for developmental toxicants (van Gelder et al., 2010). These can be divided into broad categories of reactive chemicals, chemicals with a defined molecular target, changes in embryonal or placental physiology that are developmentally adverse, or changes in maternal homeostasis that indirectly affect the embryo (Wu et al., 2013). Reactive chemicals may adduct nucleic acids or proteins or induce oxidative stress and are generally non-specific in the type of cells they affect. While oxidative stress may target cells in a broader manner than if a chemical were to target a specific receptor or enzyme, somatic cells of different lineages, stages of differentiation, or species of origin may react differently to oxidative stress depending on their antioxidant capacity and stress-response pathways (Dai et al., 2020; Gao et al., 2021). Chemicals that interact with specific biomolecules are probably more specific, as their toxicity is dependent on the target being expressed in a particular cell type at a specific time during development. Notably, the mEST assays have a high concordance with animal models for teratogens that affect structures besides the heart, which suggests that many teratogens may not have specific modes of action, that conserved signaling pathways may affect many end points, and/or that multiple modes of action are involved in cardiomyocyte differentiation.

Alternative methods in developmental toxicology

From animal test to alternative method

Regulatory chemical hazard assessment for human prenatal developmental toxicity is usually based on the OECD Test Guideline (TG) 414. This guideline describes the protocol for prenatal post-implantation exposure in pregnant rat dams or rabbit does, with adverse outcomes observed in fetuses at term. This protocol requires a larger number of animals due to the presence of parental and offspring generations within a single study. Non-mammalian alternatives to pregnant animal studies are amenable to mechanistic processes that offer insights into chemical modes of action for human-relevant pathways.

Alternative methods vary in complexity from whole embryo culture, organ and cell culture to molecular assays, each with their pros and cons as to throughput, predictive value and mechanistic insight for *in vitro* evaluation of developmental hazard potential (Piersma, 2006). The implementation of alternative assays in prenatal developmental toxicology is hampered by the complexity of prenatal development (Piersma et al., 2014). The development of the implanted conceptus through the various embryonic and fetal stages to term is accompanied by a host of complex mechanisms programmed throughout evolution, which show different vulnerabilities in time and location in the conceptus. Covering all these possible targets for toxic disruption in *in vitro* systems is a significant challenge. Clearly, individual reductionist *in vitro* assays cannot be expected to sufficiently cover the entire prenatal developmental landscape (Piersma et al., 2013).

Validation of individual alternative methods

Validation of an alternative testing method as one-to-one replacements of animal studies has a technical and a relevance component (Hartung et al., 2004). The former consists of assessing the intra- and inter-laboratory variability and reproducibility, and the latter includes the description of the biological and chemical domains of the assay, and its predictive power. This situation has two limitations. First, it assumes that the alternative method, which is by definition reductionist, will accurately predict all developmental toxicity in the intact organism. Second, it assumes that the animal study is the gold standard, and the target for chemical hazard assessment is the human. Therefore, the interpretation of validation findings is not straightforward. Predictive power highly depends on the gold standard against which the *in vitro* assay is validated. For prenatal developmental toxicity this generally includes harmonized protocols most commonly using pregnant rats or rabbits for drugs and chemicals. Many existing assays that have undergone a validation exercise were tested with a limited set of data-rich compounds. This often resulted in an overall predictive capacity of around 80% or higher, based on a quantitative positive/negative scoring (Brown, 1987; Genschow et al., 2002). The significance of this predictivity is not always clear, given the gold standard comparison and the limited coverage *in vitro* of chemical space and of the biological space of prenatal development. Moreover, it is difficult to determine predictivity for data-poor and weaker toxicants that may be less potent with regards to the incidence and magnitude of adverse developmental outcomes. This situation has hampered regulatory acceptance of individual alternative assays for prenatal developmental toxicity (Kugler et al., 2017). Classically, validation has employed positive and negative reference chemicals to calibrate balanced accuracy of the *in vitro* assay for assessing model performance for sensitivity (e.g., detecting toxicants) and specificity (getting it right); however, as noted earlier the induction of effects is clearly dependent on the exposure level of the tested chemical. Therefore, quantitative prediction based on *in vitro* concentration–response assessment followed by quantitative *in vitro* to *in vivo* extrapolation (QIVIVE), also referred to as reverse dosimetry (Thomas et al., 2019), may be expected to give better informed estimation of assay performance based on the inclusion of activities such as maternal ADME and trans-placental kinetics (Louisse, Beekmann, and Rietjens, 2017). Proposals as to study designs for quantitative predictivity assessment have been published (Fragki et al., 2017; Punt et al., 2018; McNally et al., 2018).

Validated methods in developmental toxicology

Three prenatal developmental toxicity assays were validated in a European Centre for the Validation of Alternative Methods (ECVAM) validation study in four participating laboratories (Genschow et al., 2002). It included the rat whole embryo culture, the rat limb bud micromass, and the mouse cardiac EST. The *in vitro tests* were subjected to a limited set of around 20 chemicals, including positive developmental toxicants, negatives and an intermediate group of weak toxicity observed across animal models of human developmental toxicity. Predictive capacity (overall accuracy including positive predictive value (sensitivity) and negative predictive value (specificity)) was close to 80% for these chemicals and inter-laboratory comparison was generally acceptable. None of the assays reached regulatory acceptance, probably for a combination of different reasons such as default uncertainty factors. Although the EST appeared less successful as to predictivity in follow-up investigations with additional chemicals (Chapin et al., 2007; Paquette et al., 2008; Marx-Stoelting et al., 2009), it is still used in industry for prescreening and prioritization

purposes, as well as for research into the molecular mechanisms of chemical interference with embryonic cell differentiation (Robinson and Piersma, 2013). Furthermore, a host of additional stem cell tests have been developed, e.g. with genetic modifications such as for molecular effect markers, and for different differentiation routes such as the neural and osteogenic lineages, and with different automated readout systems (Schmidt et al., 2017; Madrid et al., 2018).

Non-mammalian small model organisms (SMOs)

Developmental toxicity studies have been conducted using birds (chickens) (Stark and Ross, 2019), amphibians (Xenopus) (Berg, 2019), fish (zebrafish) (Beekhuijzen et al., 2015), insects (Drosophila) (Affleck and Walker, 2019), and other lower animals such as Hydra (Fu et al., 1990). Advantages with SMO species as alternatives to mammalian studies is their bioavailability to direct effects of chemicals avoiding the maternal system as well as the vast information and knowledge about their embryology. The latter is in the context of phylogenetic conservation of fundamental developmental pathways and processes during the gestational period corresponding to peak sensitivity to teratogenic intervention. Added value over cell culture assays is provided by the higher complexity of complete embryo models. The PSC assays may be covering only limited windows during developmentally susceptible lifestage, e.g. only the pluripotent status (Zurlinden et al., 2020), or the window up to a certain differentiation stage, e.g. the differentiation into cardiomyocytes (Genschow et al., 2004). SMO-based assays, like those using vertebrate embryos, provide further development comparable with later windows of prenatal developmental. Therefore, the SMO-based assays are used as an integral part of most *in vitro* test batteries to predict prenatal developmental toxicity. They can be seen as complementary assays to PSC-based assays regarding the biological domain.

New approach methods (NAMs)

In 2007, the National Research Council published Toxicity Testing in the 21st Century: A Vision and a Strategy (NRC, 2007). This report addressed the potential for automated high-throughput screening (HTS) and high-content screening (HCS) assays and technologies to identify chemically induced biological activity in human cells and to develop predictive models of *in vivo* biological response that would ignite a shift in thinking from traditional animal endpoint-based testing to human pathway-based risk assessment paradigm. Concurrent with the NRC 2007 report, the US EPA launched the ToxCast research program that utilized statistical methods and machine learning algorithms for profiling biological pathways and building bioactivity signatures predictive of toxicity (Judson et al., 2010; Kavlock et al., 2012; Judson et al., 2014; Richard et al., 2016; Thomas et al., 2019). A richness of HTS/HCS data has since fueled the building and testing of integrative models for "encoding the toxicological blueprint of active substances that interact with living systems" (Sturla et al., 2014; Knudsen et al., 2015; Juberg et al., 2017). The ToxCast program is part of the federal Tox21 consortium to develop a cost-effective approach for efficiently prioritizing the toxicity testing of thousands of chemicals and the application of this information to assessing human toxicology (Collins et al., 2008). Operationalizing NAMs for predictive developmental toxicity was a major theme covered in the recent FutureTox-4 conference (Knudsen et al., 2021).

In the European Union FP7 scientific research funding programme, various multinational research projects addressed the use of embryonic stem cell-based test systems amongst other alternative assays for developmental toxicity testing. The ReProTect project aimed at generating a battery of assays covering most of the different segments of the reproductive cycle (Hareng et al., 2005). The combination of complementary assays was deemed essential in order to enhance reliability of predictions. The predictive capacity of the EST in different chemical and pharmaceutical areas was discussed at an implementation workshop (Marx-Stoelting et al., 2009). The added value of mechanistic knowledge of chemicals tested in relation to understanding the biological domain of the assay was considered important in view of the interpretation of assay results. The EU FP7 ESNATS project generated a number of neurodevelopmental toxicity assay protocols based on human ESC lines (Krug et al., 2013). Valproic acid and methylmercury chloride were successfully used as data-rich positive control compounds to study the responsiveness of the assays. One of the main conclusions of the project related to the importance of culture conditions and assay protocol characteristics such as duration of culture and timing and duration of exposure, in addition to readout parameters (Rovida et al., 2014).

In 2018, sixteen US federal agencies contributed to the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) Strategic Roadmap for Establishing New Approaches to Evaluate the Safety of Chemicals and Medical Products in the United States (ICCVAM, 2018). This roadmap described a framework to enable development of, establish confidence in, and ensure use of new approaches to toxicity testing that improve human health relevance and reduce or eliminate the need for testing in animals. These regulations highlight the need for NAMs to evaluate the potential toxicity of chemicals in screening and prioritization contexts.

In silico approaches to integrate in vitro testing

Given the reductionist nature of *in vitro* alternative assays as opposed to the complexity of prenatal development, and the lack of their regulatory implementation, increasing efforts are invested in combining assays with complementary biological domains in order to enhance the coverage of developmental mechanisms. This can be done pragmatically by combining available assays and assessing combined predictivity over individual assays (e.g. ToxCast) (Piersma et al., 2013; Richard et al., 2016). Combinations of assays can also be designed from a more mechanistic perspective, defining the developmental landscape that needs to be covered and subsequently filling that landscape with the necessary models (e.g. virtual embryo, ontologies) (Kavlock and Dix, 2010; Hutson et al., 2017; Staal et al., 2017; Hessel et al., 2018; Scialli et al., 2018). The latter requires sufficient knowledge of developmental biology, ideally leading to an *in silico* systems biology model covering all necessary developmental elements. This would enable reliable quantitative predictions from the integration of quantitative *in vitro* data from all assays involved. The complexity of development will require advanced machine learning techniques, which in turn require sufficiently detailed input of the biology of the system (Piersma et al., 2019). Large data sets that are now being generated on developmental gene expression responses to chemical exposures in various *in vitro* systems may aid in fulfilling that requirement. However, the timeline of these developments towards regulatory implementation is yet uncertain. Such approaches would eventually overcome the limitations of individual *in vitro* assays and would optimize their use in hazard assessment.

In the context of developmental toxicity, a NAM-based approach that combines *in vitro* data with *in silico* models may be encapsulated in the adverse outcome pathway (AOP) concept (Wittwehr et al., 2017). Extending data from PSC-based assays to pregnancy and development requires AOPs with sufficient biological coverage to move from one key event (KE) to the next, such that measurement of a KE downstream is predictive of KE (or adverse outcome) upstream. In order to achieve quantitative risk assessment, the AOP concept needs to be mirrored by the kinetic component that comes before triggering of the initiating event. It is the kinetic route from external exposure of the intact organism, via absorption, distribution, metabolism and excretion (ADME), that determines which molecular initiating events in what target tissues are triggered and with what dose-time profile (Alqahtani, 2017). This is essential information for hazard assessment

that goes beyond this review of toxicodynamic alternative *in vitro* assays for developmental toxicology, but needs to be incorporated in integrated systems biology approaches (Przybylak et al., 2018).

Stem cell biology and its application in dart testing

Pluripotent stem cell origins

Embryonic stem cells (ESCs) are derived from the inner cell mass of the blastocyst. They are both self-renewing (capable of unlimited, undifferentiated proliferation *in vitro*) and pluripotent (may differentiate to essentially all non-trophectodermal cell lineages in the embryo proper). To fully harness this resource, it is necessary to understand their history and biology.

Stem cell progenitors were discovered in mice first as hematopoietic precursors (Becker et al., 1963) and subsequently as precursors to the embryo proper (Evans and Kaufman, 1981). The isolation of mouse ES cell lines enabled revolutionary generation of transgenic mice for phenotypic assessment that are now used by the millions worldwide to study and model a range of human diseases and toxicities. This technology exploited two key properties of ES cells: self-renewal, which permitted propagation *in vitro*; and pluripotency, which following injection into the inner cell mass spread the genetic modifications to a range of somatic cells in the embryo, including on occasion the germ cell line. The term 'ES cell' was introduced to distinguish embryo-derived pluripotent cells from teratocarcinoma-derived pluripotent embryonal carcinoma cells that could give rise to cystic 'embryoid bodies' *in vitro* (Martin and Evans, 1975) and that secreted growth factor(s) that stimulated proliferation and/or inhibited differentiation of normal pluripotent mESCs (Martin, 1981).

A major advance in stem cell biology came in 1998 when Thomson and colleagues reported isolating ES cells (ESCs) from human blastocysts. The cells retained pluripotency after undifferentiated proliferation *in vitro* for 4–5 months and retained the potential to form derivatives of all three embryonic germ layers, including gut epithelium (endoderm); cartilage, bone, smooth muscle, and striated muscle (mesoderm); and neural epithelium, embryonic ganglia, and stratified squamous epithelium (ectoderm) (Thomson et al., 1998). They also retained normal karyotype, exhibited high levels of telomerase activity and expressed cell surface markers unique to primate ESCs. Here, for the first time, was a renewable source of human cell types for testing new therapeutics and regenerative medicine. However, their procurement from viable human embryos opened bioethical debate "*about the beginnings of life and the ends of science*", as stated by US President George W. Bush. Bush imposed a moratorium on federally funded research on new hESC lines created after August 9, 2001 (Murugan, 2009). This limited research to a small number (21) of existing ESC lines until 2009, when the Obama administration eased the funding moratorium making it possible for federally funded scientists to use excess embryos "*created using in vitro fertilization for reproductive purposes and were no longer needed for this purpose*" to obtain ESCs for research purposes (Murugan, 2009). Controversy over hESC research cooled in 2006 when Yamanaka and colleagues reported the striking finding that dermal fibroblasts from an adult mouse could be reprogrammed to a pluripotent stem cell state under ESC culture conditions simply by altering the expression of four genes (*Oct3/4, Sox2, c-Myc, Klf4*) (Takahashi and Yamanaka, 2006). This breakthrough led to human induced pluripotent stem cell (iPSC) technology. Today, iPSCs are promising tools for application in personalized medicine that may capture individual variability in a cell type with pluripotent developmental potential (Ferreira and Mostajo-Radji 2013). Patient-derived iPSC lines have enabled not only toxicity testing but likely will be an important tool in the drug discovery process (Funakoshi and Yoshida, 2021; Harvey et al., 2021). In addition, generation of iPSCs from individuals who exhibit specific side effect profiles or idiosyncratic reaction to drugs may prove useful in screening for relatively rare but serious toxic effects. The potential of patient-specific iPSCs to be able to identify patients that would respond adversely or favorably to a drug could provide powerful new tools for personalized medicine.

Comparing sensitivity of human versus mouse ESC platforms

To date, most developmental toxicity testing utilizing PSCs has been performed using mESCs, and the implementation of hESCs in developmental toxicity testing has been slower to evolve. Ethical and legal issues of performing toxicity testing with hESCs aside, hESCs have been more difficult to culture than mESCs; however, advances in culturing techniques have eliminated this issue (Desai et al., 2015). The most apparent advantage of using hESCs in addition to or instead of mESCs is to limit the possibility of false negatives that may arise due to species-specific differences. Differences exist at least for some popular reference compounds, and it remains to be determined if hESC-based assays outperform mESC-based assays (this point is addressed below for specific case examples). Furthermore, species-specific differences are often due to *in vivo* differences in metabolism or toxicokinetics, which may not apply to *in vitro* assays. Finally, it is important to note that mESCs are shown to be more naïve than hESCs (Dong et al., 2019), meaning the molecular features of mESCs more closely resemble those of pluripotent cells in the early embryo. This may infer that mESCs are a more appropriate model for chemicals that specifically target the ICM to epiblast transition and leads to speculation that hESCs may have greater sensitivity to chemicals that target later processes, such as gastrulation and beyond.

Advantages and drawbacks of hESC versus iPSC platforms

Induced pluripotent stem cells (iPSCs) are derived from somatic cells that are reprogrammed back to an embryonic-like state. The differentiation routes of iPSCs *in vitro* can be modified by compound exposures, enabling the study of basic mechanisms of differentiation and embryotoxicity. As an example, retinoic acid exposure was shown to induce mesoderm at the expense of endoderm differentiation (Saili et al., 2020). Thalidomide disrupted mesoderm differentiation, and valproic acid induced a shift from neuroectoderm to neural crest differentiation (Matyskiela et al., 2018; Meisig et al., 2020). An indepth description of the expanding literature in this area is beyond the scope of this review focusing on test system development. Although iPSCs are pluripotent, their use in developmental toxicity testing has remained limited. This may, in part, be due to the fact that many iPSC lines tend to have lineage bias towards the lineage of origin, which may be due to an incomplete reset of DNA methylation back to an "embryonic state" (Liang and Zhang, 2013). However, many commonly used ESC lines have also been shown to have lineage biases (Bock et al., 2011; Tsankov et al., 2015). Additionally, donor age, sex, race, and exposure history may all influence the toxicant response of iPSCs. Therefore, the true power of iPSCs may lie in the field of personalized toxicology, and in their utility to incorporate genetic diversity into *in vitro* developmental toxicology studies [reviewed in (Jennings, 2015; Liu et al., 2017)].

Readout parameters of EST

Biological domain of EST

The capacity of mouse-derived mESCs (D3 line) to differentiate *in vitro* into a wide variety of cell types forms the basis of the mEST. The traditional mEST entails spontaneous embryoid body (EB) differentiation for 9–10 days after which beating cardiomyocytes and/or myosin heavy chain (MHC) expression are measured relative to cytotoxicity (Genschow et al., 2000; zur Nieden et al., 2001; Genschow et al., 2004; Peters et al., 2008; Seiler and Spielmann, 2011;

Chandler et al., 2011; Barrier et al., 2012). However, while the mEST easily identified strong embryotoxic compounds, it was found to be less accurate resolving weak teratogens from non-teratogens (Chapin et al., 2007). Given these limitations, a variety of modified versions of the mEST have been developed, although they have not been validated as thoroughly. For example, protocols have been optimized for EB formation in 96-well plates to increase assay throughput (Peters et al., 2008), while fluorescent-activated cell sorting has been used to assess cardiomyocyte differentiation, thus eliminating subjectivity of scoring contracting cardiomyocytes (Buesen et al., 2009). Dimopoulou et al., (Dimopoulou et al., 2018) incorporated placental BeWo b30 cells into the mEST and demonstrated that the incorporation of placental transport velocity data can increase assay predictivity. A great deal of effort has also been expended to incorporate transcriptomics into the mEST (>50 studies published (reviewed in: van Dartel and Piersma, 2011). Of note, a molecular EST, in which expression of 12 developmentally regulated genes are assessed, reduced assay time leading to beating cardiomyocytes at 10 down to 4 days based on gene expression and with a similar degree of accuracy (72%–83%) as the original mEST (Panzica-Kelly et al., 2013). It should be noted, however, that accuracy comparisons between assays have limited value as differences in chemicals tested, end points measured, and scoring criteria may affect outcome.

Transcriptomics

Improvements in endpoint scoring, test duration, definition of the predictivity and the applicability domain for developmental hazard detection were needed for implementation of the EST into regulatory testing strategies. One suggested improvement was to better underscore early changes in gene expression profiles across multiple lineages by transcriptomic profiling. A 'differentiation track' of gene expression that discriminated several mechanistically diverse developmental toxicants was resolved with Affymetrix chips to potentially improve the predictivity and expand the applicability domain for developmental hazard detection of the ECVAM-validated mEST (van Dartel et al., 2010). Another approach mined the cardiogenic effects of 309 ToxCast chemicals against ~500 diverse *in vitro* assays in the ToxCast dataset (Chandler et al., 2011). That analysis reported statistically significant associations in the mEST response for 26 of the chemicals tested. A correlation against the multiplex reporter assays in the ToxCast portfolio inferred increased bioactivity expressed for several critical developmental regulators, including BMPR2, PAX6 and OCT1, in association with decreased ESC differentiation. Changes to multiple genetic regulators in reactive oxygen species signaling pathways (NRF2, ABCG2, GSTA2, HIF1A) were also strongly correlated with decreased ESC differentiation as a potential mode of action that accounted for disruption of the cardiogenic readout.

Metabolomics

Metabolomics is the study of small molecules (metabolites) that are the end-product of various cellular processes, including energy metabolism. Quantitative analysis of various metabolites in the culture medium ('secretome') is being implemented in developmental toxicity testing and has allowed researchers to identify small molecules that can serve as putative biomarkers of myriad diseases (Cezar et al., 2007; Palmer et al., 2013; West et al., 2010). The utility of metabolomics in developmental toxicity testing was first established by demonstrating that valproate, a neurodevelopmental toxicant, can alter the secretome of hESCs, affecting processes such as tryptophan and glutamate metabolism (Cezar et al., 2007). In a follow-up study, hESCs were exposed to the ECVAM test set, and the secretome was analyzed using an untargeted metabolomics approach, which led to the identification of 8 metabolites (dimethylarginine, aspartic acid, arginine, glutamate, GABA, malate, succinate, isoleucine) that correlated with teratogenicity (West et al., 2010). A predictive model based on similar results from untargeted metabolomics could accurately classify 88% (7/8) of drugs and 73% (8/11) of environmental toxicants in two separate blinded test sets (Kleinstreuer et al., 2011). Subsequently, it was found that a 12% reduction in the ornithine to cystine ratio in the secretome of hESCs maintained in a pluripotent state during a 3-day exposure accurately classified developmental toxicity potential of blinded test sets of drugs or environmental toxicants when compared with cell viability across concentration–response profile (Palmer et al., 2013). The targeted biomarker (ratio of ornithine to cystine secreted or consumed from the media) could accurately classify 77% of compounds in a 13-compound test set, but the overall sensitivity of the assay was weak as only 57% (4/7) of reference teratogens were properly classified. This commercial assay (https://www.stemina.com) has been recently used to screen 1065 ToxCast phase I and II chemicals in single-concentration or concentration–response) (Zurlinden et al., 2020). The analysis of this ToxCast dataset showed that 19% of 1065 chemicals yielded a prediction of developmental toxicity based on the ornithine:cystine biomarker. Predictive performance of the assay reached 79%–82% accuracy with high specificity (>84%) but modest sensitivity (<67%) against well-qualified, data-rich developmental toxicants; however, sensitivity declined as the evidence requirements applied to the animal studies were relaxed, such as fetal effects in one species (rat or rabbit), species discordance, or concurrent maternal toxicity. Statistical analysis of the most potent chemical hits on specific biochemical targets in different ToxCast assays revealed positive and negative associations with the stem cell response, providing insights into the mechanistic underpinnings of the targeted endpoint and its biological domain (Zurlinden et al., 2020). Changes in the ornithine:-cystine biomarker have been used successfully to rank the developmental toxicity potential of a dozen retinoid compounds using human iPSCs rather than hESCs (Palmer et al., 2017). These results indicate that metabolic processes in pluripotent hESCs or iPSCs can be exploited for quantitative prediction (in terms of critical response concentration) of potential developmental hazards with high specificity; however, it will be important to understand how the targeted biomarker response of pluripotent stem cells (eg, critical drop in the ratio of ornithine to cystine secreted or consumed from the media) works to predict a developmental hazard.

Morphometry

The involvement of morphometric approaches can be useful in implementing the EST assay for predictive developmental toxicology. An annular pattern of mesoendoderm differentiation, epithelial-mesenchymal transition and cell migration of human iPSCs was shown to quantify and classify teratogenic potential of compounds exposed *in vitro* (Xing et al., 2015). Because embryogenesis requires biomechanical forces and biochemical microenvironments, the involvement of morphometric approaches can be useful in implementing the EST assay for self-organizing potential of hESCs becomes useful in a morphogenetic sense. Standardized and quantitative systems have been described that mass produce uniformly sized spheroids that synchronously differentiate into EBs (Flamier et al., 2017; Tronser et al., 2018). EBs are 3D aggregates of PSCs that spontaneously differentiate into all 3 embryonic germ layers (ectoderm, mesoderm, and endoderm), in a process that recapitulates many of the molecular events that occur throughout early embryogenesis (Weitzer, 2006). Furthermore, EBs can recapitulate the three-dimensional growth and axial elongation of early embryos during gastrulation and early organogenesis (Warkus and Marikawa, 2017). In these systems, stem cells are influenced by local cell–cell and cell-extracellular matrix interactions. Microscale technologies such as micropatterned and microfluidic systems, along with embryoid body-on-a-chip modalities, are now emerging as models for studying human embryogenesis and high-throughput testing platforms (Knudsen et al., 2017; Rico-Varela et al., 2018).

Several groups have proposed that EB growth dynamics can be used to predict the embryotoxic potential of chemicals (Flamier et al., 2017; Kang et al., 2017; Warkus and Marikawa, 2017). Kang

et al., demonstrate a strong correlation between EB area and cardiomyocyte beating, indicating that a reduction in EB size closely reflects cardiomyocyte differentiation (Kang et al., 2017). This led to the development of a novel mEST variation in which cardiomyocyte beating was replaced with an EB size measurement, termed the EB stem cell test (EBT). The accuracies of the mEST and EBT were found to be similar (86.9% and 90.5%, respectively) when compared across a 21-compound test set. It is, however, well known that EB size affects lineage specification even in the absence of any toxic exposure. For example, EBs inoculated from 200 cells prefer to form ectodermal derivatives. Cardiomyocytes differentiate best in larger EBs of 750 cells and chondrocytes optimally develop in EBs made from 800 cells. These lineage effects are likely based on differential amounts of soluble signals that drive differentiation outcome and how densely cells are arranged with neighboring cells, which would also trigger differential cell-internal signaling cascades (see Moon et al., 2014; Nath et al., 2017). Therefore, it is not surprising that a cytotoxic chemical that decreases the number of cells, would decrease the number of beating cells. However, it would be difficult to discern whether that chemical is cytotoxic only or whether it truly interferes with a developmental (differentiation) process.

EB growth can be monitored using high-content imaging (HCI) devices. Recently, a pre-validation study of the EBT was conducted in Korea (Lee et al., 2019). This pre-validation study evaluated the predictive accuracy of the EBT using 26 coded test substances by two steps: intra-laboratory and inter-laboratory reproducibility tests. Since some substances have different embryotoxic potentials at different pregnancy periods, in this study, a new prediction model consisting of non-toxic and toxic classes was used, instead of the existing prediction model assessing embryotoxicants in three or four classes. The results of the intra- and inter-laboratory tests had an accuracy above 80% when substances were classified using the predictive model. In this report, EBT can accurately classify various embryotoxicants in a short time with less effort and greater validation to reflect 'growth retardation and embryo mortality' (Lee et al., 2019). Therefore, the EBT may be most sensitive to chemicals that exhibit cytotoxicity toward the cells (which may lead to a negative developmental outcome) raising question as to the sensitivity of EBT to detect chemicals that truly elicit a developmental inhibition in the absence of generalized alterations on EB mass.

Spontaneous differentiation of EBs

To test for developmental toxicity, Shinde et al. (2015) exposed differentiating EBs to teratogens throughout a 14-day differentiation window, and then assessed EB differentiation using transcriptomics and immunocytochemistry. Although the overall accuracy of this approach has not yet been established, several proof-of-principle experiments with cytosine arabinoside (Jagtap et al., 2011), thalidomide (Meganathan et al., 2012), valproic acid (Krug et al., 2013), and methyl mercury (Shinde et al., 2015) have demonstrated the validity of this approach. For example, thalidomide perturbed the expression of genes associated with limb and heart development (Meganathan et al., 2012), which coincides with clinical observations of thalidomide toxicity in humans.

Despite these promising characteristics, this assay is relatively low-throughput and does not utilize consistently sized or individually cultured EBs. This may prove problematic for high throughput developmental toxicant screening, as EB size can influence PSC differentiation, likely due to reduced diffusion of nutrients and oxygen into the core of larger EBs, resulting in increased cell death and altered differentiation patterns (Moon et al., 2014; Nath et al., 2017). Furthermore, pooled EBs rapidly (within several hours) fuse, resulting in increased EB size, cell death, and altered differentiation (Dang et al., 2002), which may also contribute to assay variability. However, advances in EB formation protocols may allow researchers to overcome these limitations (reviewed in (Pettinato, Wen, and Zhang, 2015; Cornwall-Scoones et al., 2021).

Lineage specification and differentiation

Additional differentiation models have also been incorporated into the mEST in an attempt to expand the assay's applicability domain. For example, (Adler et al., 2008) developed a human EST to limit potential false negatives that may arise due to species-specific differences. Furthermore, many embryotoxic compounds can adversely affect skeletal development (an endpoint regularly monitored in *in vivo* developmental toxicity studies), which has led researchers to incorporate osteoblast differentiation protocols into the mEST (Chen et al., 2015; zur Nieden et al., 2010a, 2010b). Developmental osteotoxicity was assessed by morphometric analysis of calcified matrix, measurement of calcium levels, and activity of alkaline phosphatase (an enzyme involved in matrix calcification), and expression of osteocalcin (exclusive to mineralized tissues and a biomarker of developmental osteotoxicity) (zur Nieden et al., 2003). This shows the value of Ca2+ deposition in the EST as a reliable endpoint for routine industrial assessment of developmental osteotoxicity.

Because the original cardiac differentiation EST failed to detect methyl mercury (MeHg), numerous research groups have worked to incorporate neuron differentiation protocols into the mEST (Baek et al., 2012; Stummann et al., 2009; Theunissen et al., 2010). This has resulted in the proper classification of MeHg as a developmental neurotoxicant. Baek et al aimed to improve the EST for detecting developmental neurotoxicants using a neuronal endpoint (Tuj-1 ID50) and flow cytometry analysis of Tuj-1-positive cells to detect the effects of MeHg, valproic acid, and sodium arsenate in an adherent monoculture differentiation method (Baek et al., 2012). Using Tuj-1 ID50 (concentration inhibiting differentiation by 50%) instead of cardiac ID50 in the EST, all of the tested chemical positives were classified as embryotoxic, while the negative controls were correctly classified as nonembryotoxic. To support the validity of Tuj-1 ID50, they compared the results from two experimenters that independently tested MeHg using modified EST. An additional neuronal endpoint (MAP2 ID50), obtained by analyzing the relative quantity of MAP2 mRNA, was used to classify the same chemicals. There were no significant differences in the three endpoint values of the two experimenters or in the classification results, except for isoniazid. These results indicate that Tuj-1 ID50 can be used as a surrogate endpoint of the traditional EST to screen developmental neurotoxicants correctly and can also be applied to other chemicals.

With regards to CNS morphogenesis, Piersma et al., developed a murine neural embryonic stem cell test (ESTn) that can mimic parts of early differentiation of embryonic brain. Their aim was to investigate whether this test would rank the potencies of three valproic acid analogues and reveal mode of action by investigating their individual effects on four cell types: stem cells, neurons, astrocytes and neural crest cells. Using biomarkers for immunocytochemical (GFAP) and qPCR (*Fut4, Cdh1*) readouts at different time points during differentiation, they observed that a combined evaluation of some endpoints was useful for ranking of valproic acid analogues consistent with the *in vivo* developmental toxicity potency of these compounds (de Leeuw et al.. 2019).

High- content screening (HCS)

High content screening (HCS), in combination with automated image analysis software, have allowed researchers to rapidly screen large suites of compounds for biological activity (Buesen et al., 2009; Knudsen et al., 2013). As an example of this technology, Kameoka et al. (2014) directed hESCs to differentiate down the mesendoderm lineage with activin and testing 55 pharmaceutical compounds throughout a 3-day differentiation window. Cell viability and differentiation were then assessed by staining for DAPI (nuclear)

and SOX17 (definitive endoderm). Teratogenic risk was based upon a compounds ability to reduce nuclear translocation of the SOX17 transcription factor. Here, 94% of pharmaceutical compounds (67/71), and 87% of environmental toxicants (13/15) with known *in vivo* teratological outcomes were properly classified using this approach. Whether these outcomes were limited to endodermal derivatives is not clear; however, given the rapid (72 h) and automated nature of this approach, the platform represented a promising advancement in the field of stem cell toxicology.

Genetically modified EST systems

Transgenic reporter strains (morpho-regulatory pathways)

Early embryogenesis is primarily under the control of six key signaling pathways—the Wnt/β-catenin, transforming growth factor beta (TGF-β), Notch, Hedgehog, receptor tyrosine kinase/Ras, and cytokine receptor signaling pathways. The crucial nature of these pathways is demonstrated by the fact that genetic manipulation results in embryonic lethality or developmental defects in mammalian models (Loebel et al., 2003). This has led to the development of several transgenic ESC reporter lines that can be used to monitor pathway activity following toxicant exposure. For example, the ReProGlo Assay utilizes mESCs transfected with the SuperTopFlash luciferase reporter, which can be used to monitor Wnt signaling pathway activity following exposure (Uibel et al., 2010). In the initial validation study, the ReProGlo assay properly classified 76% of compounds in a 17-compound test set. Although the low-cost and rapid (24-h) nature of the ReProGlo assay made it an attractive tool for developmental toxicant screening, follow-up studies reported high false negative rates, suggesting the applicability domain of the assay needs to be better defined with regards to sensitivity (Uibel and Schwarz, 2015). Generation of additional transgenic lines capable of assessing the activity of the other signaling pathways (i.e., Notch, Hedgehog, TGF-β, receptor tyrosine kinase/Ras, and cytokine receptor signaling) may help overcome this limitation when used in combination.

Another strategy used to generate transgenic ES cell lines is to isolate ESCs from transgenic mouse models, which has led to the generation of two reporter mESC lines that can be used to monitor Wnt and TGF-βb signaling (Kugler et al., 2017; Kugler et al., 2016). Both models have been tested with a small set of developmental toxicants (valproic acid, retinoic acid, and 6-aminonicotinamide), and reporter activity correlates well with subsequent cardiomyocyte differentiation assays, demonstrating the validity of these approaches for the small set of chemicals tested. However, the overall accuracy of these models is yet to be tested.

Systems engineered for specific reporter genes (Hand1-Luc EST).

Transgenic reporter mESCs that can be used to monitor cardiomyocyte differentiation have also been generated (Le Coz et al., 2015; Nagahori et al., 2016; Suzuki et al., 2011). In these models, luciferase reporter gene expression is driven by either the *Hand1* or *Cmya1* promoters, both of which are considered indispensable for proper heart development. This provides an early marker that replaces the need for scoring cardiomyocytes with the rapid and quantitative endpoint of luciferase fluorescence. Furthermore, both the *Hand1*-Luc and *Cmya1*-Luc assays have been shown to properly predict 83% and 92% of developmental toxicants, respectively. In an extensive validation study involving four independent laboratories, the predictivity of the *Hand1*-Luc EST was evaluated with 71 chemicals. First, the positive predicted value was 80.8%. However, in parallel, accuracy and sensitivity are low (60.6% and 47.7% respectively), emphasizing that the *Hand1*-Luc EST has limitations, and that no conclusion can be drawn if a negative result is triggered in the test.

EST systematic scoping literature review

Systematic scoping review of embryonic stem cell tests

In order to add transparency and reproducibility to the review, systematic scoping review of the literature was conducted. Several formats and methodologies for literature review exist such as expert narrative review, systematic scoping review, systematic map, rapid systematic review, or full systematic review. Systematic reviews have been implemented in clinical research by Cochrane Collaboration over the last 40 years and brought comprehensiveness, objectivity, reproducibility and transparency to medicine and are the foundation of evidence-based medical practice (Higgins et al., 2011). Full systematic review is considered the highest quality evidence with the least risk of bias but requires a focused question. Systematic scoping review framework (Moher et al., 2015) was the methodology considered most appropriate for the scope of this project. This methodology was used because of the broad objective defined for this review that does not fit the more narrowly defined pillar of a Population (including animal species), Exposure, Comparator, and Outcomes (PECO) framework. On the basis of results from this broad scoping review of literature, the EST methodologies are discussed in the context of application for non-animal chemical and pharmaceutical safety assessment of prenatal developmental toxicity. This discussion includes relevance for human safety assessment, chemical applicability domains, and a path for international regulatory acceptance. Additionally, the list of included studies and chemicals that were tested in the assays may serve as a resource for OECD and other parties interested in the subject of developmental toxicity testing.

Protocol

The protocol for this scoping review was put together by an international expert working group assembled by the Japanese Center for the Validation of Alternative Methods (JaCVAM), published on July 10, 2019, and can be accessed: https://zenodo.org/record/2528920. The eligibility criteria were defined before the conduct of the review and are part of the protocol. Inclusion criteria were: (1) publications that provide detailed methodologies and primary data using *in vitro* embryotoxicity tests, (2) exposure to at least one chemical substance, (3) no publication date restriction, and (4) any publication status. Exclusion criteria were: (1) whole organism tests (mammals, fish, chicken, nematode), (2) publications containing no primary data (narrative reviews, opinion letters, conference abstracts and similar), and (3) publications in non-English language. We define "*in vitro* research" as a study in which living parts of animals (including humans) up to and including the level of organization of tissues, but not whole organs, are exposed to chemicals in order to detect adverse effects. Our definition is intended to exclude: (1) intact animals, (2) whole embryo and *ex vivo* studies (in which whole organs are removed and studied outside the animal's body), and (3) tissue slices. Our definition includes studies in cell cultures (including stem cell and organo-typic cultures). Studies into the safety of drugs and other xenobiotic chemicals as well as biological effects of exposure to chemical substances (environmental toxicology) are both included. We limited the search to PubMed database using a search strategy listed herein. We accepted additional information supplied by the working group members.

Literature search strategy

We referred to the (Kohl et al., 2018) review of existing tools to identify appropriate screening, data extraction and management software. the PubMed literature search was done with the Abstract Sifter tool (Baker, Knudsen, and Williams, 2017); filtering for duplicates with Sciome SWIFT-Review and SWIFT Active Screener; screening

and selection with SWIFT Active Screener; extraction, coding, bibliographic information and data storage in Microsoft Excel. The PubMed SR search strategy uses the systematic review filter from (Shojania and Bero, 2001): *(Embryonic stem cells OR IPSC OR Induced Pluripotent Stem Cells OR embryonal stem cells OR "embryonic stem cell"[tw]) AND ((toxicity OR congenital abnormalities OR prenatal exposure delayed effects OR dysmorphogenesis OR dysmorphogenetic OR abnormalities, drug-induced) AND (fetus OR embryo OR embryonic OR embryonic development OR larva OR eggs OR prenatal OR pregnancy OR Gene Expression Regulation, Developmental) AND ((chemical OR drug OR compound OR toxicity tests OR High-Throughput Screening Assays) OR embryotoxicity OR embryotoxicants OR teratogenicity OR teratogens OR developmental toxicity OR developmental toxicants OR teratogenic agents OR "developmental neurotoxicity" OR "developmental cardiotoxicity"))*. The queries used and the resulting filtered and screened publication lists can be found (and the queries rerun) in the Abstract Sifter included as Supplemental File S1.

Selection of studies and data extraction

Screening was conducted by several working group members, EBTC staff and trained volunteer reviewers. Titles and abstracts were screened by two reviewers. All titles and abstracts were screened using Sciome SWIFT-Active Screener software with Artificial Intelligence (AI)-assisted methodology (Howard et al., 2020). In this method, both reviewers had to agree on inclusion or exclusion. All conflicts were discussed and resolved in weekly meetings. The AI-assisted software prioritized abstracts for screening based on relevance. The working group stopped the screening after reaching 95% predicted recall. Full texts were screened by two reviewers, with discussion to resolve conflict. Reviewers followed the protocol with pre-specified inclusion criteria (Stephens et al., 2018). The data extracted included biological domain, assay type, species, cell line, readout, and the chemical names and CASRN numbers, if present.

The query text as of 2/19/2019 returned 1,533 PubMed entries. The title and abstract (Level 1) screening was conducted in Sciome's SWIFT Active Screener cloud-based software. The two reviewers had to agree on inclusion or exclusion of each paper. All conflicts were discussed and resolved in weekly teleconferences and the project manager (KT), in consultation with expert group, resolved the conflicts which reviewers were unable to resolve themselves. The AI-assisted software prioritized abstracts for screening based on relevance determined based on the reviewers' decisions. The working group stopped the screening after reaching 96% predicted recall. Overall, 403 papers progressed to full-text stage (level 2 screen). Level 2 screening was conducted again by 2 reviewers per paper and resulted in 13 papers included. The search was updated on July 23, 2019, and March 31, 2021, and returned 89 and 439 papers, respectively, which were reviewed using the same criteria, which added 7 papers to the final list included in results synthesis. The PRISMA diagram showing the flow of the studies is shown in Fig. 1. The 20 included studies are listed in Table 1.

Mapping the chemical space tested in EST

The chemical names from each study were combined into one list and uploaded to the batch search form of the US EPA Chemicals Dashboard (https://comptox.epa.gov/dashboard/dsstoxdb/batch_search) (Williams et al., 2017). The DSSToxIDs and CASRN numbers of each of the chemicals was downloaded from the Dashboard. Chemical names with no match because of name misspelling were corrected and resubmitted. The chemical names and identifiers were combined with the study PubMed IDs and study authors and placed as a pivot table onto the CuratedLists sheet of the Abstract Sifter tool to get an overview of chemical coverage at this detailed level (Supplemental File S1).

In order to get a list of chemicals for graphical summaries, we reduced the granularity by mapping the DSSToxIDs to Medical Subject Heading (MeSH) identifiers. The National Library of Medicine has defined chemical MeSH names to index PubMed articles (https://meshb.nlm.nih.gov). One MeSH name can encompass more than one salt form, meaning that Penicillin G and Penicillin G potassium are collapsed into one term – Penicillin G. The majority of the chemicals were matched to a MeSH term (1053 out of 1256); the chemicals without a MeSH match included chemicals with little literature presence (e.g., Chlorethoxyfos). A pivot table was created with the MeSH chemical names and study detail and placed onto the CuratedMeSHList sheet (Supplemental File S1).

The mapping of the chemical identifier to MeSH terms also let us take advantage of the MeSH dictionary entries for each chemical, and specifically, of the mappings from chemicals to the MeSH tree categories of Pharmacologic Action (D27.505) and Specialty Uses of Chemicals (D27.720). Pharmacologic action categories include, for instance, mechanisms of action (e.g., Neurotransmitter Agents) and therapeutic uses (e.g., Tranquilizing Agents); specialty use terms include Flame Retardants and Disinfectants. We assembled these categories for each chemical in the list. The resulting categories and counts were inserted into a spreadsheet Supplemental File S2. The results are presented in four sheets. The first sheet has all chemicals and all MeSH categories. The second sheet has only chemicals in more than one study. The third and fourth sheet have chemicals in more than one study with just pharmacological action and just specialty use, respectively.

The review included papers that fell into the following two main categories: (1) Stem cell-derived models with either formal validation data on reference chemicals often used for alternative assays or specificity and sensitivity calculated for multiple chemicals (20 publications); and (2) Stem cell-derived models with no formal validation data or specificity / sensitivity data (typically with exposure to 1–8 compounds and more mechanistic in nature) (231 publications). Only papers in category 1 were deemed by the expert group relevant for the objectives of the study, with papers in category 2 serving as evidence of use of the models described in detail in the category 1 publications or providing mechanistic information (e.g., gene expression), important for advancing the field.

Results synthesis

Among the included 20 publications there were 8 mouse ESC (ES-D3, Cmya1-ES, DBA/1lacJ, ES-E14TG2a, ESCs from the NMRI strain, Hand1-ES, KOB1-ES, R1) and 1 human (WA09, also named H9) ESC-derived models. The ES-D3 hanging drop model (Spielmann et al., 1997) had the largest number of studies with statistical evaluation and formal validation information available. Fifteen of the studies employed the cardiac muscle cell differentiation route. Alternative routes included endoderm and mesoderm differentiation, or no differentiation at all. The observed endpoints and differentiation pathways were not always transparent, as readout could also be cell viability or the release of mediators in the culture medium. Morphological assessment of contracting muscle cell foci and staining for cardiac muscle cell specific biomarkers (gene or protein expression) were the readout parameters most clearly used to ascertain the outcome of cardiomyocyte differentiation. Details of test protocols differed among laboratories. Apart from cell line and readout parameters protocol differences included aspects such as culture medium, monolayer or aggregate culture, and exposure timing and duration. The studies summarized in Table 1 included 9 to 1065 test chemicals.

The total (non-redundant) number of chemicals tested was 1,274 with only 14 chemicals not returning a DSSToxID and link to the CompTox Chemicals Dashboard. Tretinoin (all-trans retinoic acid) and 5-fluorouracil were the most widely tested chemicals with results in 18 assays, followed by Penicillin G tested in 16 assays,

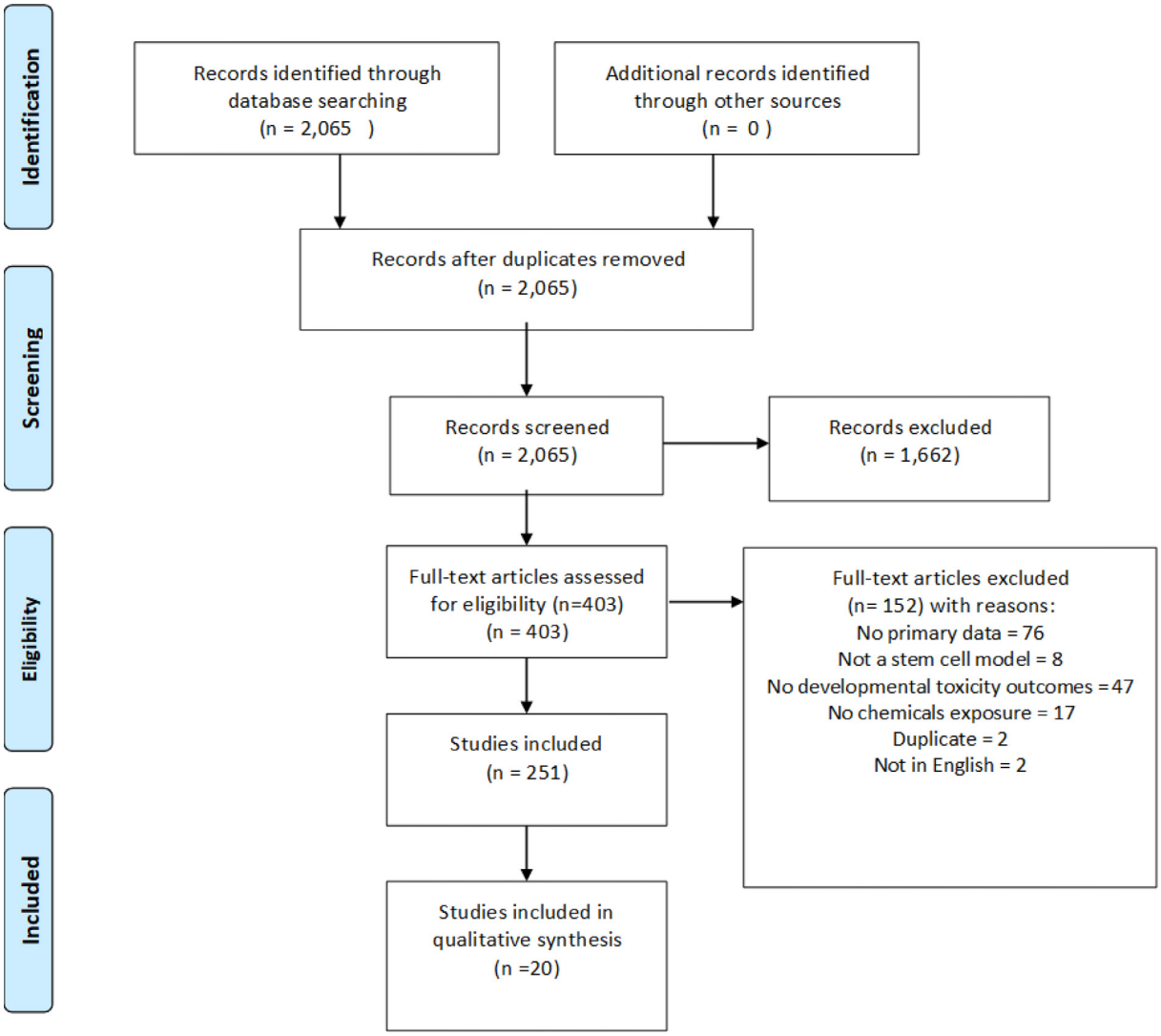


Fig. 1. PRISMA 2009 Flow Chart.

and Methotrexate and Valproic Acid in 14 studies. Summarizing the tested chemicals by MeSH names resulted in 1066 chemical entries. For 1046 of them a pharmaceutical action or specialty use was given. When mapped to the MeSH pharmacological actions and use categories, the pharmacological actions with the most chemicals were designed to influence the nervous system, enzymes and infective agents. The specialty use that described the most chemicals was pesticides including, for example herbicides, insecticides and fungicides.

Chemical space tested

Our selection process yielded 192 studies published between 1991 and 2021 (Fig. 2). In the 1990s, when *in vitro* embryotoxicity assay development was based on pluripotent stem cells, only a few studies were published. The advent of the ECVAM validation study of the murine embryonic stem cell test stimulated an increase in studies based on this methodology and consequently the number of publications increased (Genschow et al., 2002). In the last decade, the number of studies increased further presenting several approaches to establish new protocols using pluripotent stem cells aiming to broaden the applicability domain and/or improve the predictivity of prenatal development toxicity.

Based on the 20 studies, 1274 chemicals were investigated in the ESC or iPSC assays to predict prenatal developmental toxicity. The relatively high number of chemicals was mainly driven by one of the latest publications investigating over 1000 chemicals (Zurlinden et al., 2020). The number of chemicals investigated in at least two independent studies was 153. Within these chemicals 127 pharmaceutical actions and 20 specialty use (industrial chemicals) could be identified. An overview is given in Figs. 3 and 4, respectively.

During the development of *in vitro* embryotoxicity assays in general, several lists of reference compounds of prenatal development toxicity were established. Two of these lists are especially relevant in assessing test improvements and protocol modifications. These lists of special interest are the ECVAM list of 20 reference compounds used during the ECVAM validation trial of the murine EST (Brown, 2002), and second, the list of 29 reference compounds recommended to be used to support the qualification of alternative methods for animal testing according to the ICH S5 (R3) guideline on reproductive toxicology (ICH S5 (R3), 2020). Both lists have a significant overlap of reference compounds listed. Taking a look at the chemicals investigated in the finally assessed 20 studies in this review, the chemicals of both lists represent the those investigated in the highest number of published studies (Fig. 5).

Table 1
Summary table of included publications with embryonic stem cell models.

PMID	Brief reference	Assay type	Species	Cell line	Developmental cell type	Readout	Biological domain tested	Number tested chemicals
1807538	Laschinski, 1991	ESC	Mouse	ES D3	CM	Cell viability	Differentiation	28
20692990	Newall, 1996	ESC	Mouse	ES D3	endoderm-like cells	Morphology, Cytotoxicity	Differentiation	25
15588166	Genschow, 2004	ESC	Mouse	ES D3 & BALB/c 3Ts	CM	cytotoxicity	Differentiation	20
21964422	Suzuki, 2011	ESC	Mouse	Hand1-ES & BALB/c 3 T3	CM	Cell viability, gene expression	Differentiation	24
18361453	Paquette, 2008	ESC	Mouse	DBA/1lacJ	CM	Cell viability	Differentiation	48
20493898	West, 2010	ESC	Human	WA09	CM	relative cell viability	Differentiation	26
21925528	Kleinstreuer, 2011	ESC	Human	WA09	CM	Metabolome	Differentiation	11
24154490	Kameoka, 2014	ESC	Human	H9 & LSJ-1	Mesendoderm	SOX17 expression	Differentiation	86
24123775	Palmer, 2013	ESC	Human	H9 (WA09)	ESC	ornithine/cystine ratio	Undifferentiated	46
23042729	Panzica-Kelly, 2013	ESC	Mouse	ES D3	CM	gene expression	Differentiation	12
27445234	Cheng, 2016	ESC	Mouse	R1, SP3	CM	karyotype analysis	Differentiation	18
27444379	Nagahori, 2016	ESC	Mouse	KOB1-ES	CM	Cell Viability	Differentiation	71
30339957	Lee, 2018	ESC	Mouse	ES-E14TG2a & BALB/c 3 T3	CM	Cell viability	Differentiation	26
31636845	Kawamura, 2019	ESC	Mouse	ES D3	CM	beating CM	Differentiation	20
30934112	Zang, 2019	ESC	Mouse	D3	CM	fluorescent EGFP marker	Differentiation	9
32238694	Aikawa, 2020	IPSC	Human	iPSC / fibroblasts	CM	beating CM	Differentiation	14
31711903	Marikawa, 2020	ESC	Human	H9	EB	paraxial mesoderm, neuroectoderm markers	Differentiation	20
32073639	Zurlinden, 2020	ESC	Human	H9 (WA09)	ESC	ornithine/cystine ratio	Undifferentiated	1065
32633240	Lee, 2020	ESC	Mouse	ES-E14TG2a & BALB/c 3 T3	EB	cell viability and size of EBs	Growth	35
32205227	van Oostrom, 2020	ESC	Mouse	ES-D3	CM	beating of CM	Differentiation	24

Legend to table 1: ESC embryonic stem cells; IPSC induced pluripotent stem cells; CM cardiomyocytes; EB embryoid bodies. References to Table 1: (Laschinski, Vogel, and Spielmann, 1991; Newall and Beedles, 1996; Genschow et al., 2004; Paquette et al., 2008; West et al., 2010; Kleinstreuer et al., 2011; Suzuki et al., 2011; Palmer et al., 2013; Panzica-Kelly et al., 2013; Kameoka et al., 2014; Cheng et al., 2016; Nagahori et al., 2016; Lee et al., 2019; Kawamura et al., 2019; Zang et al., 2019; Aikawa, 2020; Marikawa et al., 2020; Zurlinden et al., 2020; Lee et al., 2020; van Oostrom, Slob, and van der Ven, 2020).

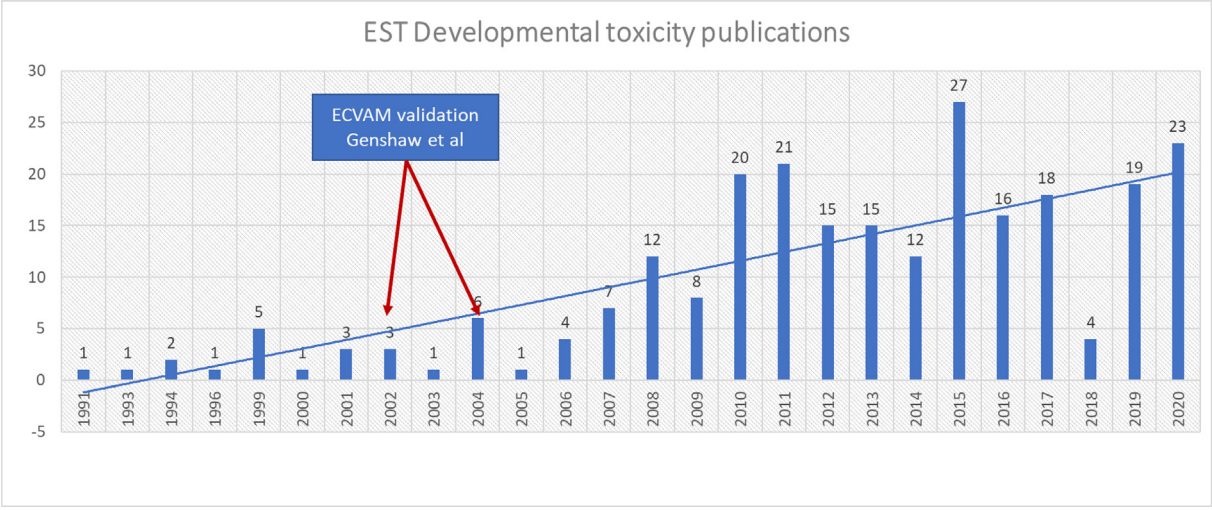


Fig. 2. Number of studies published in the last decades and included in this review about the use of the pluripotent stem cell to predict prenatal developmental toxicity.

Case studies of three selected chemicals

Based on the high number of chemicals included in the 20 studies assessed in this review, a closer look to all chemicals in detail is out of the scope. For an in-depth examination, we selected three chemicals, 5-fluorouracil, thalidomide and caffeine, and took a closer look at the effective concentrations in the different protocols in the study as well as their predicted teratogenic potential *in vivo* based on the *in vitro* results.

5-Fluorouracil (5-FU) was identified as one of the two (along with tretinoin) most often investigated chemicals in the studies in Table 1, appearing in 18 studies (Fig. 6), most likely because it was selected as a positive control for the ECVAM validation trail. It is considered to be an embryotoxic and teratogenic compound, its active metabolites

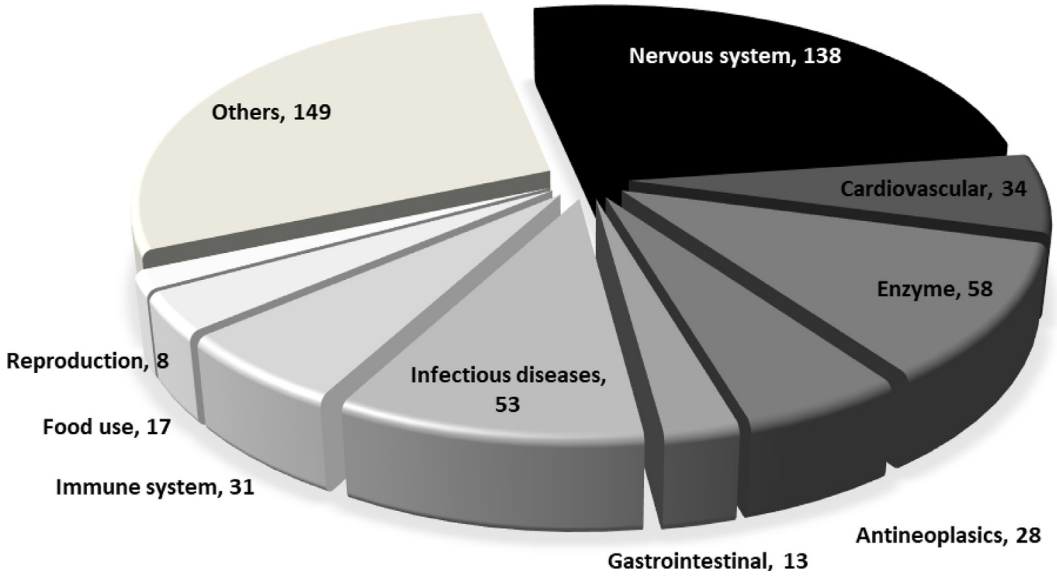


Fig. 3. Prevalent MeSH pharmaceutical action classes of chemicals studied in pluripotent stem cell assays.

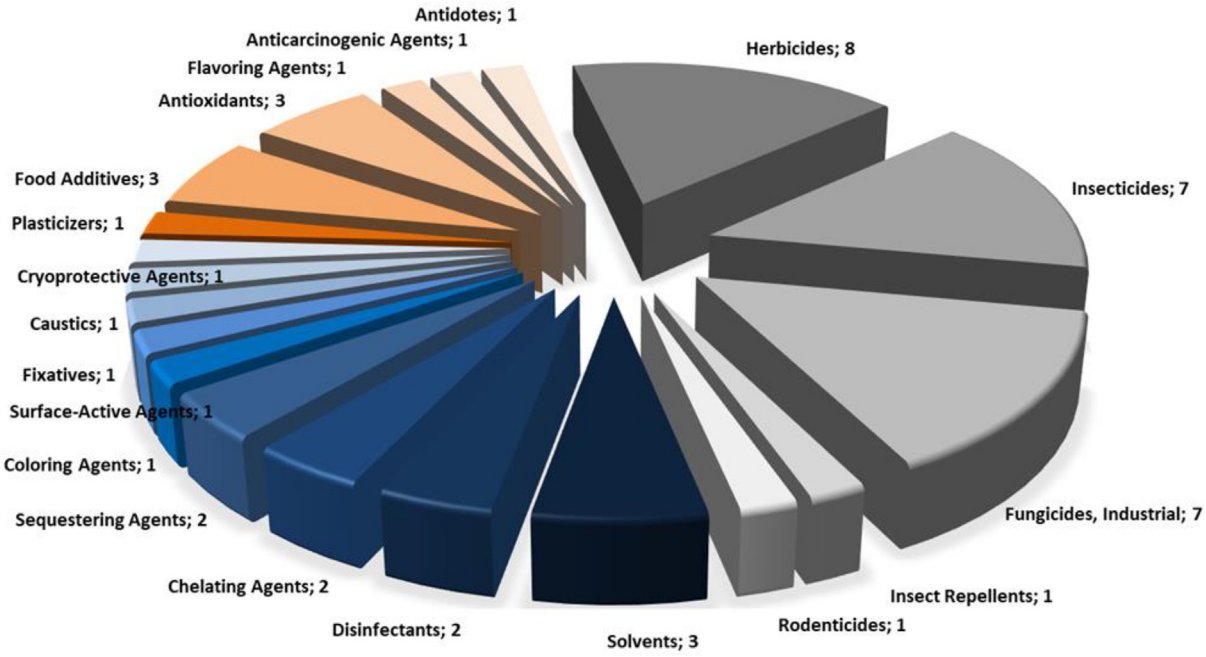


Fig. 4. Prevalent MeSH specialty uses of chemicals studied in pluripotent stem cell assays.

inhibiting thymidylate synthase, and with that affecting DNA synthesis and cell proliferation, resulting for instance in reduced fetal weight gain in rats (Lau et al., 2001; Setzer et al., 2001). In the studies included in this review, 5-FU was demonstrated to be causing test compound-related effects *in vitro* at relatively low concentrations (0.01 to 1.00 μg/ml). As an exception, West et al., determined the effective concentration to be 27 μg/ml but did not test concentration dependency and, thus, might not have identified the lowest effect level (West et al., 2010). In tests with cells other than PSCs the effective concentrations were higher (e.g., (Suzuki et al., 2011)) or equal to the effective concentration in PSC. The comparison of effective concentrations on proliferation and differentiation of PSC was equivocal in the 12 studies addressing both endpoints. In 6 studies proliferation was less sensitive than differentiation and in 6 studies vice versa. Being aware that nominal concentrations *in vitro* and plasma concentrations *in vivo* should not be compared directly without proper kinetic extrapolation, in almost all studies the effective concentrations *in vitro* were lower than the therapeutic plasma concentration level in humans being 1 μg/ml (Casale et al., 2004). In cases where the prediction model included not two (positive / negative) but three prediction classes (strong, weak, and non-embryotoxic), strong and weak embryotoxic classifications were summarized as positive outcome for this review. Thereby, all studies predicted correctly the embryotoxic/teratogenic potential of 5-Fluorouracil *in vivo* based on the *in vitro* results.

Thalidomide is one of the most well-known teratogens and shows clear species differences in the likelihood to cause a teratogenic adverse outcome, which are not observed in rats but manifested in rabbits and humans. Seven out of 9 studies investigating thalidomide

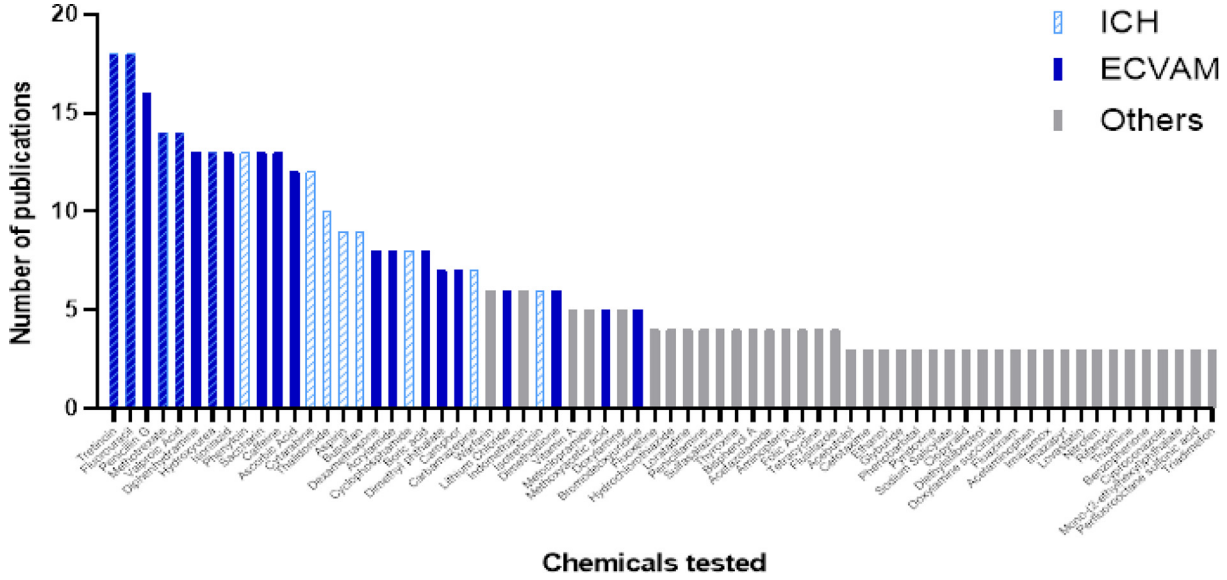


Fig. 5. Number of studies investigating the test compounds in relation to their listing as reference compounds. The blue color of the bars indicate that the chemical was listed as a reference compounds for developmental toxicity (PDT) by (Brown, 2002). The striped bars indicate chemicals listed as a reference compound for PDT in the ICH S5-R3 test guideline. Grey bars indicate other reference chemicals absent in these two lists. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

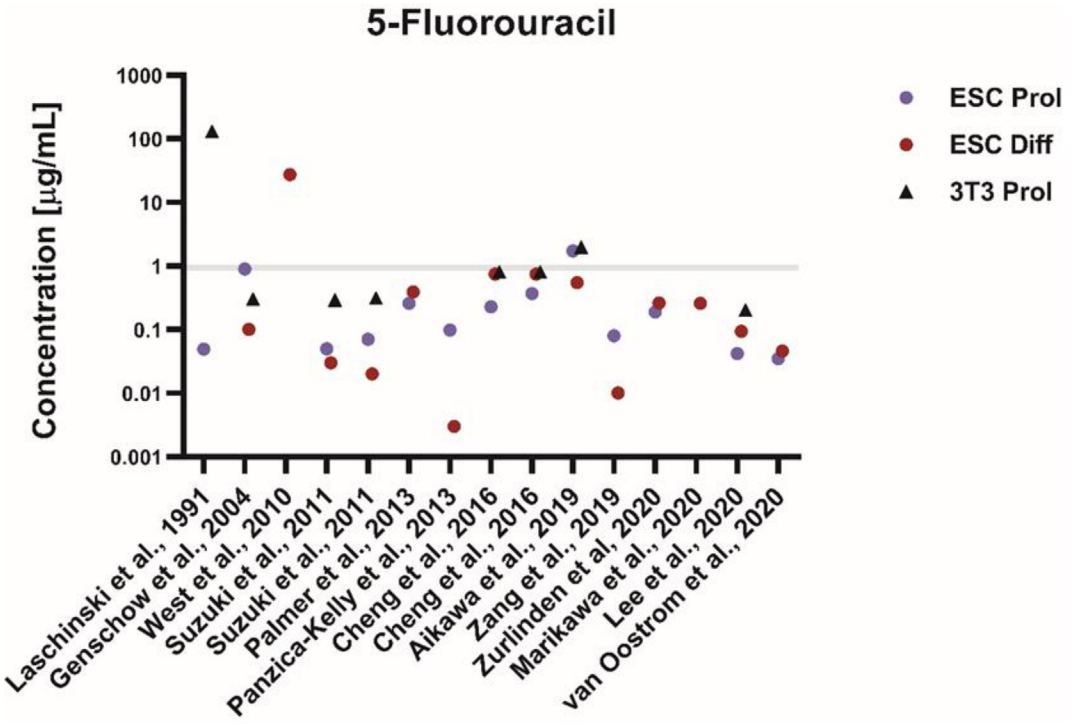


Fig. 6. Effective nominal concentrations of 5-Fluoruracil in 15 studies. The triangle indicates the lowest effective nominal concentration of 5-FU given by the authors to alter the proliferation of 3T3 BALB/c fibroblasts or comparable cells. The dots indicate the lowest effective concentration altering the growth of embryonic stem cells or comparable cells, blue dots for the proliferation and red dots for the differentiation of those cells. The grey line represents the therapeutic plasma concentration of 5-FU in humans. The publications Suzuki et al., 2011 and Cheng et al., 2016 are listed twice because they contain two data sets based on two different embryonic stem cell lines each. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in vitro were using human cells, with remaining two using murine ESCs (Fig. 7). Murine ESCs in the included publications used additional endpoints in combination with the cell viability: the size of embryoid bodies (Lee et al., 2020) and mRNA expression analyses (Panzica-Kelly et al., 2013). However, both murine ESC studies were not able to predict the teratogenic potential correctly, testing up to the maximum soluble concentration. In other cells than PSC, the effective concentrations were higher (e.g. (Aikawa, 2020)) or equal to the effective concentration in PSC. In all 5 studies addressing thalidomide effects on proliferation and differentiation of human PSC, proliferation

was less sensitive (about 10–100 μg/ml) than differentiation (about 0.1–1.0 μg/ml) in all cases. Therefore, the nominal effective concentration *in vitro* for the alteration of differentiation was below and for the alteration of proliferation was above the therapeutic plasma concentration in humans (Eriksson et al., 2001). All studies using human pluripotent stem cells predicted correctly the embryotoxic/teratogenic potential of thalidomide *in vivo* based on the *in vitro* results.

Caffeine shows an inconsistent classification of prenatal development toxic potential in literature. In 7 out of 14 studies (Fig. 8) the authors state that caffeine is non-embryotoxic or non-teratogenic as well as in the other 7 studies as embryotoxic and teratogenic. With the exception of one study (Laschinski et al., 1991) considering the *in vitro* outcome as false negative), all authors interpreted the *in vitro* outcomes as true results matching their claimed *in vivo* classification. The effective concentrations *in vitro* are relatively high, generally in the range of 100 to 1000 μg/ml. One exception is an observed effective concentration at 4.4 μg/ml (West et al., 2010). In other cell types the effective concentrations were equal to the effective concentration in PSC, with one exception (Lee et al., 2019) in which fibroblasts were less sensitive. Without the consideration of plasma concentrations in humans, the *in vitro* results would be as equivocal as the *in vivo* results are interpreted. The nominal minimal effective concentration *in vitro* for the alteration of the growth of all cell types was significantly above the plasma concentration after regular coffee in humans (Cappelletti et al., 2018). Since in all cases except the West et al. 2010 study these values are two magnitudes higher, it seems unlikely that caffeine has an embryotoxic or teratogenic potential in humans under an average coffee consumption. This summary highlights the importance of *in vivo* effective plasma concentrations in the target species, to put *in vitro* results in the context of *in vivo* reality. These concentrations can be either obtained from public sources for known chemicals or drugs or calculated using Physiologically Based Kinetic (PBK) models.

Case study conclusions

Among the 20 included studies we discovered a wide range of differences in the protocols with regard to the pluripotent cells used (murine, human, non-rodent, transgenic, or different strains), readouts (mRNA expression, metabolites, morphology, or function), culture durations (ranging from 1 to 10 days), prediction models (based on different decision trees, ratios, or cut-off values). Apart from such methodological differences, we found overall good reproducibility of effective nominal concentrations *in vitro* as well as the correct prediction of *in vivo* potential to cause prenatal developmental toxicity for 5-Fluorouracil and Thalidomide. Therefore, the comparison of the nominal concentration *in vitro* to the therapeutic concentration *in vivo* was supportive for the interpretation, but not required to make a prediction of *in vivo* teratogenicity. In the case of caffeine this comparison clarified that even the equivocal findings *in vitro* were only observed at concentrations, which are very unlikely to be reached by coffee consumption in humans. The species-specific effects of thalidomide in prenatal developmental toxicity were also identified in these assays in the two studies using murine ESC resulting in false negative outcomes. Overall, the three examples showed that the principle of testing for embryotoxicity using PSC seems to be predictive, robust and reproducible, and that inclusion of PBK modelling data is necessary to improve the translation of these *in vitro* results to realistic *in vivo* exposures.

Systematic scoping review of EST models

This literature analysis of ESC derived test systems for developmental toxicity has confirmed the growing popularity of such models as animal-free assays for developmental toxicity testing of chemicals. Twenty validation studies that employed > 9 chemicals tested per

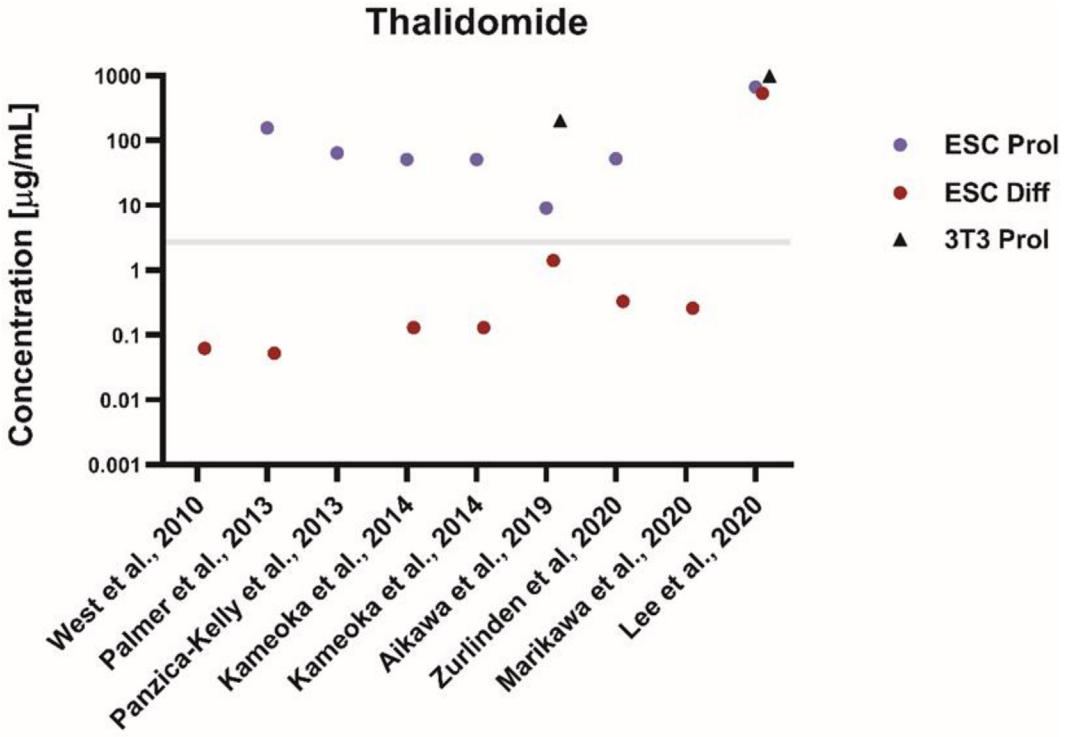


Fig. 7. Effective nominal concentrations of Thalidomide in 9 studies. The triangle indicates the minimal effective nominal concentration of thalidomide to alter the proliferation of 3 T3 BALB/c fibroblasts or comparable cells. The dots indicate the minimal effective concentration altering the growth of embryonic stem cells or comparable cells, blue dots for the proliferation and red dots for the differentiation of those cells. The grey line represents the therapeutic plasma concentration of thalidomide in humans. The publication Kameoka et al., 2014 is listed twice because it contains two data sets based on two different embryonic stem cell lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

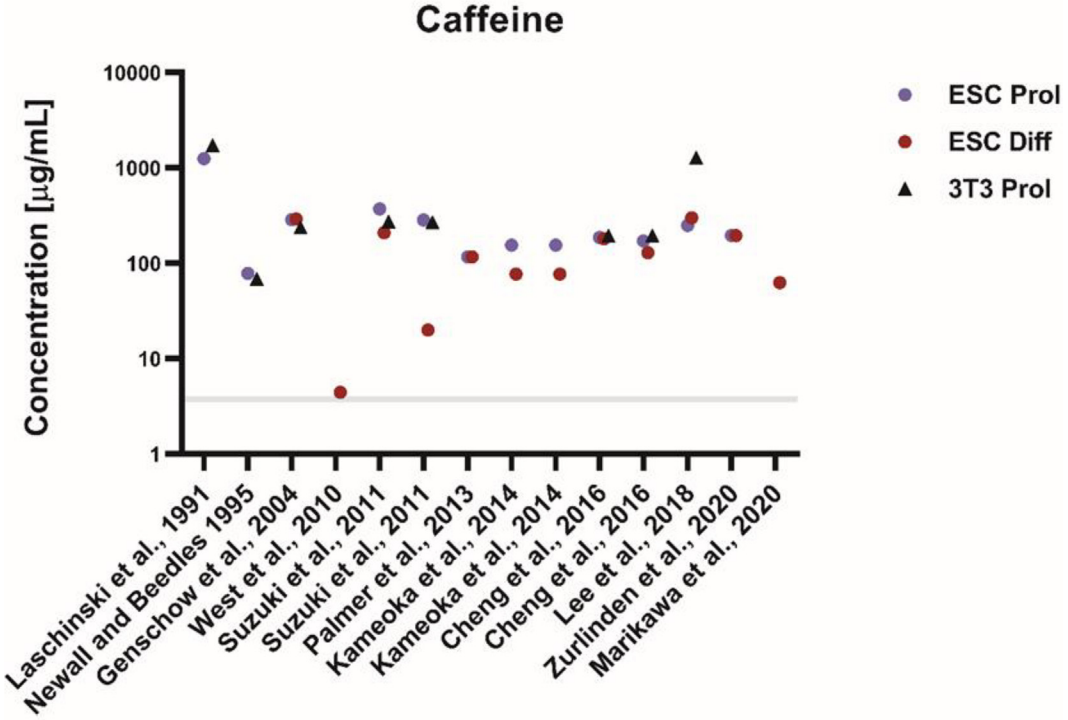


Fig. 8. Effective nominal concentrations of Caffeine in 14 studies. The triangle indicates the minimal effective nominal concentration of caffeine to alter the proliferation of 3T3 BALB/c fibroblasts or comparable cells. The dots indicate the minimal effective concentration altering the growth of embryonic stem cells or comparable cells, blue dots for the proliferation and red dots for the differentiation of those cells. The grey line represents the therapeutic plasma concentration of caffeine in humans. The publications Suzuki et al., 2011, Kameoka et al., 2014, and Cheng et al., 2016 are listed twice because they contain two data sets based on two different embryonic stem cell lines each. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

study have been published (Table 1), with over 200 supporting publications on methodology, mechanistic insights and specific chemical entities (Supplementary File S1). The limitations of this study are using only one publicly available database (PubMed), limiting the language to English, lack of formal critical appraisal of studies and no quantitative synthesis of the data. All these limitations were made consciously and detailed in the pre-published protocol. Nevertheless, this overview shows the abundance of dedicated studies into the use of embryonic stem cells in developmental toxicity assays and highlights the usefulness of embryonic stem cell lines in studies of perturbation of embryonic cell differentiation.

Perspectives

Test system accuracies

The identification of the limited number of 20 validation studies with at least 10 compounds tested (Table 1) out of 251 included studies (Fig. 2) shows that the abundance of research in this area has been dedicated to test development rather than to validation. Whilst this is partly due to the fact that test development naturally precedes validation, it also illustrates the enhanced interest in studying molecular mechanisms of action in these assays. It may also suggest awareness that the nominal accuracy of individual assays for a given group of tested compounds may not be the main driver for its usefulness in a testing strategy. Rather, mechanistic information about individual compound effectiveness within the biological domain of the assay may advise their interpretation within the broader scope of a test battery.

A variety of stem cell model-based assays for developmental toxicity have been developed over three decades, with different cell lines, protocols and readout parameters employed (Fig. 9). In spite of increased sophistication of these assays with time, their overall accuracy has remained around 80%. There are several reasons for this apparent stagnation, related to the intrinsic characteristics of the models used and the approaches to validation employed. They include the limitation of the biological domain, chemical selection and applicability domain, prediction models employed, and the existing data against which validation was performed. These aspects are addressed in more detail below.

Biological domain

Clearly, the vulnerability of ESC differentiation has proven to be an important aspect worthy of assessment in developmental toxicology. As to biological applicability domain, the test is restricted to the sensitivity of ESCs and its differentiation potential into cardiomyocytes (leaving other routes of differentiation out of consideration for now). This limits the predictive capacity for developmental toxicity in general. However, it could be argued that the developmental processes in EST, including embryonic cell proliferation, interaction (e.g., in aggregate culture) and cardiac cell differentiation incorporate a host of mechanisms that occur more widely in embryogenesis, enhancing the biological domain of EST. Moreover, it has been shown that the cardiac EST at its end stage contains up to 17% myosin heavy chain positive cardiac muscle cells in untreated controls (Seiler et al., 2004). In addition, other cell types, of ectodermal, mesodermal and endodermal origin have been observed in EST (Theunissen et al., 2013; Mennen, Pennings, and Piersma, 2019). Cell interactions between cell types may play an important role in cardiac differentiation, similar to the rise of mesoderm between endoderm and ectoderm in the embryo, followed by heart formation as the first functional

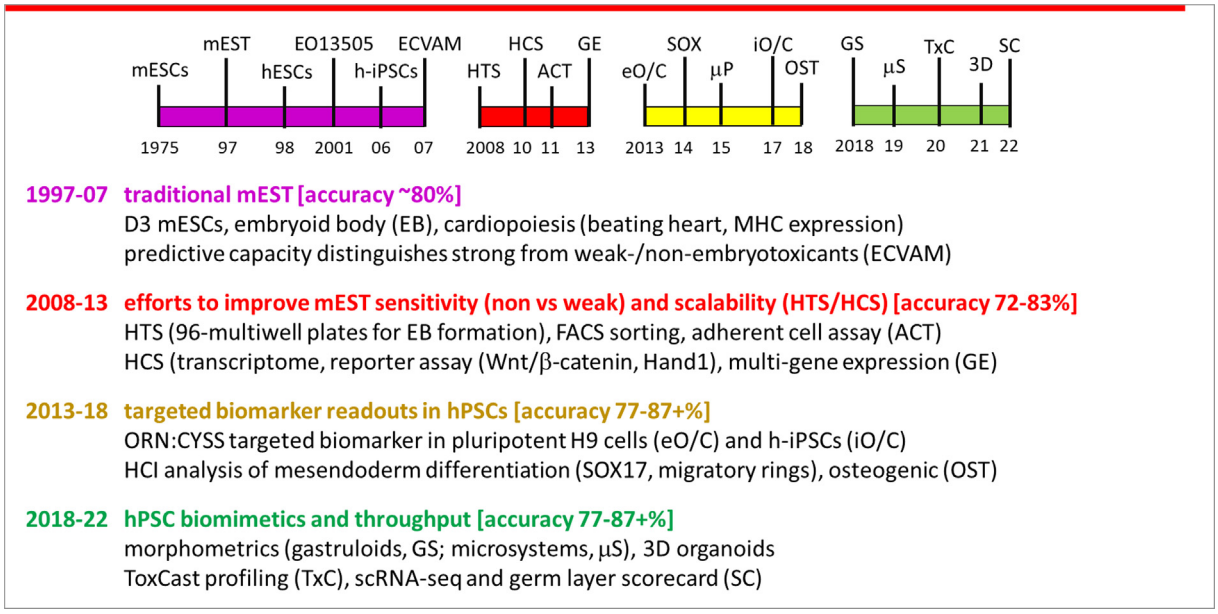


Fig. 9. Timeline of PSC-based modalities for developmental toxicity. Advances in the EST approach has led to increased throughput, reproducibility, diverse readouts, and microengineering; however, accuracies have generally remained in the 72–87% range across commonly tested, well-curated positive and negative reference developmental toxicants.

organ in the embryo. This simplified corollary of pattern formation in the embryo again points to the possibility of a wider biological domain in EST than just a single lineage of embryonic cell differentiation.

Nevertheless, it should be realized that the biological processes covered in embryonic stem cell tests are some among many processes in embryo development. Multiple alternative processes at the cellular level can be identified that may be affected by developmental toxicants that do not affect embryonic stem cell differentiation. As a consequence, a certain percentage of false negatives should be expected to occur in validation studies, again dependent on the chemical selection employed. Since the introduction of the cardiac cell differentiation EST (Seiler and Spielmann, 2011), alternative differentiation routes have also been explored, including endoderm, bone and neural differentiation (Chang and Zandstra, 2004; Madrid et al., 2018; Schmidt et al., 2017). The latter differentiation route has been especially widely applied in mechanistic studies aimed at understanding developmental neurotoxicity. A host of protocols have been published, based on mouse as well as human embryonic stem cell lines. No formal validation studies using > 10 chemicals have been identified in our search. Nevertheless, the neural EST is probably currently the most studied variant in the EST realm, which coincides with increased interest in testing possible neurodevelopmental effects of chemicals.

Chemical selection and validation

Most *in vitro* assay validation employs a list of chemicals that are categorized as either positive or negative; i.e., they either do or do not cause developmental toxicity *in vivo*. Validation studies have also been conducted by dividing toxicants into three categories (strong embryotoxicants, weak embryotoxicants, non-embryotoxicants) (Genschow et al., 2002). In this instance, a mEST performed well against this list. This is, of course, an oversimplification of toxicology, in that everything can be toxic depending on the dose. A more appropriate validation list took into account the concentration at which a chemical is expected to be toxic, and the concentration at which it is expected to have no effect. A validation list for developmental toxicity screens was assembled using pharmacokinetic data that identified peak concentration in maternal serum/plasma associated with developmental toxicity (Daston et al., 2014). To date, two labs have used this list for validating a stem cell assay (Warkus and Marikawa, 2017; Zurlinden et al., 2020), with good results.

The selection of chemicals used in a validation study is one critical factor affecting its outcome. The classical EST protocol, for which the landmark validation study by (Genschow et al., 2002) showed a relevance of around 80%, was also employed by (Chapin et al., 2007) with a different set of chemicals, which returned less favorable results. A workshop, reported in (Marx-Stoelting et al., 2009) addressed the issue and concluded that the prediction model used, based on the validation study with a limited set of compounds, had been shown to be insufficient for general application. They recommended that alternative endpoints (e.g. differentiation routes, genomics) and alternative methods of scoring compound effects (e.g. potency, as opposed to positive/negative) should be explored, in order to improve the predictive capacity of the system. Usually, chemicals are tested that are data-rich and can be clearly designated as positive or negative for developmental toxicity *in vivo,* mostly based on existing prenatal developmental toxicity studies in animals (Aschner et al., 2017; Daston et al., 2014). This ignores the grey area of chemicals with uncertain or limited effects on development, which may represent the majority of chemicals around. The most frequent effect observed in otherwise negative developmental toxicity studies is perhaps an effect on fetal weight, indicating growth retardation that may or may not be secondary to maternal toxicity. Such effects with complex etiologies may not be mimicked in specific *in vitro* assays, although in EST protocols it may perhaps be mimicked in some cases by effects on cell proliferation that could affect cell differentiation in the test. But such extrapolations between *in vitro* parameters and *in vivo* end points generally meet with considerable uncertainty. This stipulates the difficulties of chemical selection and of the extrapolation between *in vitro* and *in vivo* data.

The necessary extrapolation to the human situation is another complicating issue of using animal data as the gold standard. Although the majority of validation studies has been performed using mouse ES cell lines, increasing numbers of studies have employed human ES cell lines (Sachinidis et al., 2019). The latter takes away one issue of human relevance of *in vitro* testing. However, the number of proven human teratogens is too limited to enable carrying out validation studies with such chemicals only. Moreover, apart from the lack of human data for many chemicals, many proven animal teratogens provide

useful test chemicals for validation studies, even though the limited human relevance of the animal study as the gold standard needs always be kept in mind.

Chemical applicability domain

Given that many studies have been done with EST, including larger scale validation studies, there is quite some knowledge around about the chemical applicability domain of the assay (Riebeling et al., 2012). However, a formal assessment of the limitations in view of chemicals that can be tested has not been performed. As in many *in vitro* tests, chemicals with limited solubility, with high volatility or autofluorescence will pose practical issues. Otherwise, no specific drawbacks have occurred in EST that may not generally occur.

Prediction models

In most cases, the outcome used in validation studies is a positive versus negative scoring, which is then usually fed into mathematical prediction models (Genschow et al., 2004; Zurlinden et al., 2020; van der Burg et al., 2015). This scoring may depend simply on the induction of a certain predefined effect size occurring below a certain predefined limit concentration of the chemical tested. This approach is obviously a pragmatic simplification of reality. Therefore, the interpretation of the results of *in vitro* testing should ideally include comparison of the concentration–response characteristics in view of embryotoxic dose levels of individual compounds in pregnancy. Figs. 6–8 give those comparisons for the three case studies explored in this manuscript. First, the toxicity of a compound critically depends on the concentration at the target tissue. Thus, a compound can be positive or negative, dependent on the dose applied. Daston et al., have taken this notion forward, suggesting a series of chemicals for validation studies with two preset concentration levels, one proposed positive and one proposed negative concentration (Daston et al., 2014). This could accommodate the issue of tested concentration to some extent. This list can be considered as an example selection of data-rich chemicals, but the exact choice of chemicals for a given validation study may vary dependent on specific validation aims, e.g. related to certain chemical classes of interest. Second, assuming that the *in vitro* concentration in the assay can be considered comparable to the *in vivo* target organ concentration, it needs to be extrapolated to the external exposure in order to perform realistic hazard identification. Different chemicals with different kinetics may result in different target concentrations after the same external exposure. Such differences may also result in false negative or false positive readouts in the *in vitro* assay. Third, a very common issue with *in vitro* assays in general is the lack of metabolism that may activate or inactivate chemicals, which may also lead to false positives and false negatives. Usually, chemicals for which metabolism is crucial for correct toxicity scoring are not included in validation studies, which avoids this issue but also affects the significance of the validation study outcome.

Relevance to human health effects

The human health effects that we predict from developmental toxicity studies are the potential to produce structural malformations, growth retardation, in utero death, and functional deficits. Stem cell tests for developmental toxicity have been optimized for identifying the potential for structural malformations. Malformations are often on a continuum with growth retardation, functional deficit and death, so it is possible that predicting the potential for teratogenesis may also be predictive of these other manifestations, but only by inference. There are many mechanisms through which chemicals or drugs can affect development. In general, though, these mechanisms all involve perturbations in differentiation, signaling, metabolism and/or proliferation of cells at key timepoints during development. Because stem cells recapitulate the differentiation process that occurs in specific organs of the embryo, they are a relevant model for identifying changes to the differentiation program. That said, stem cell assays do not fully recapitulate embryonic development. Two important differences are 1) the time course of differentiation of stem cells *in vitro* may be different from the *in vivo* time course; and 2) depending on culture conditions, stem cells *in vitro* differentiate into one specific type of cells or cell groupings (e.g., cardiac myocytes, neuronal cells) and not into all cell types in the organism. Therefore, it is possible that not all possible molecular targets for teratogens are present in stem cell assays (or for the correct period of time). Whether this affects the predictive power of the assays is a question to be decided by validation studies. One criticism of using animal models to predict human toxicity is that the affinity of specific ligands for receptors or enzymes may exhibit species differences. This could also be a problem in stem cells derived from non-human species but is not an issue for human-derived stem cells (Sachinidis et al., 2019). The readout for stem cell assays is also different from the readout for *in vivo* developmental toxicity assays, where the outcome (structural malformations, decreased fetal weight, etc.) is directly relevant to the adverse human health effects we try to predict and prevent. The stem cell assay readout, whether anatomical, physiological (e.g., beating of cardiomyocytes) or biochemical (e.g., ornithine/cystine ratio), is a surrogate for those adverse effects, and it may not be obvious how (or whether) the readout is related to adverse outcomes *in vivo*. The elucidation of adverse outcome pathways may help to illustrate the relationship between these effects at the biochemical or cellular level with outcomes at the tissue/ organ level.

Draft ICH harmonized guideline S5(R3)

The ICH S5(R3) Guideline for Detection of Toxicity to Reproduction for Human Pharmaceuticals has been fully revised by the expert working group of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). This most recent revision, S5(R3), includes alternative test systems that have not yet qualified for regulatory acceptance for risk assessment of developmental toxicity in the clinical development and marketing approval of pharmaceuticals. The embryo stem cell test (EST), whole embryo culture (WEC), and zebrafish test (ZET) are well-known alternative test systems for evaluating the developmental toxicity of pharmaceuticals (Augustine-Rauch et al., 2016; Brannen et al., 2016), and many corroborative or validation studies have also been conducted for these test systems. (Ball et al., 2014; Cassar et al., 2019; Daston et al., 2014; Genschow et al., 2000; Genschow et al., 2002; Genschow et al., 2004; Gustafson et al., 2012; Piersma et al., 2004; Piersma, 2006; Padilla et al., 2012; Spielmann et al., 2006). The S5(R3) guideline requires the developer to demonstrate the relevance of using alternative assays for the evaluation of developmental toxicity based on data describing the performance of the alternative assays under consideration, including an explanation of the biological and chemical applicability domains is required to justify the application of an alternative assay. The extensive and continuing research in EST regarding these aspects, as elucidated in this review, will facilitate its regulatory application in pharmaceutical safety assessment.

OECD testing strategies

The broad interest and extensive research activities concerning the EST have given detailed insights into the biological domain covered as well as its chemical applicability domain. Moreover, the predictive capacity of the EST in its various forms has been documented extensively, as reported in this review. The biological coverage facilitates the use of the EST for testing selected key events in AOPs (Ankley

et al., 2010; Tollefsen et al., 2014). The AOP concept, providing a framework for documenting mechanistic toxicity pathways from initiating events to adverse outcomes, has been effectively adopted and promoted by OECD, resulting in an ever-expanding AOPwiki containing hundreds of AOP descriptions (https://aopwiki.org/aops). Awareness is growing that AOPs should eventually be updated to include quantitative descriptions, forming the basis for what could possibly evolve into animal-free quantitative risk assessment (Spinu et al., 2020). Moreover, AOPs are by nature mutually interrelated in a quantitative network within the domain of physiology. The qAOP network will ultimately cover the 'toxable' selection of physiological processes, those processes that need to be monitored in order to fully appreciate the toxic properties of chemicals (Piersma et al., 2019). In the wider context of developmental toxicity testing, recent history has shown that a single *in vitro* assay will not suffice for reliable prediction of developmental toxicity potential of a given chemical. For regulatory use, combinations of test systems with complementary biological domains will probably provide a better assessment than individual assays (Piersma et al., 2018a; Piersma et al., 2018b; Baker et al., 2020). One exemplary study with 12 chemicals illustrated this: a combination of assays, including the EST, was needed to achieve an 11/12 correct score (Piersma et al., 2013). Interestingly, the single false negative chemical had a mechanism of toxicity that was not covered in any of the tests in the battery, which shows the importance of complete coverage in the test battery of the biological domain underlying possible embryotoxicity. Thus, the EST and other assays should preferably be combined in Integrated Approaches to Testing and Assessment (IATA) or Defined Approaches (DA) in order to optimize *in vitro* prediction of *in vivo* developmental toxicity (Alépée et al., 2019; Tollefsen et al., 2014). A wide variety of assays considered non-animal in international regulations is available today, from assays at the molecular level to the level of intact vertebrate (zebrafish) embryos. The ToxCast program is particularly assay- and data-rich, including many hundreds of assays and currently over one thousand compounds tested in part or all of the battery tests (Judson et al., 2010). ToxCast now includes an array of public data on hESC predictive developmental hazard (Zurlinden et al., 2020). Such a battery combined with lower throughput assays at a higher level of complexity and machine-learning approaches to model the data show promising opportunities towards animal-free developmental toxicity testing. These batteries await integrating computational physiological models of embryogenesis, that enable the integration of individual assay data to the level of the intact individual, preferably fine-tuned to the human situation. Such computational models are being generated from the perspective of existing molecular developmental biology knowledge, and computational models of selected embryogenetic processes have been published (Knudsen et al., 2015; Knudsen et al., 2020). Further development of such models may help improve chemical hazard assessment.

Summary

In this manuscript we reviewed the history, development, application, and validation of EST in the context of predictive regulatory developmental toxicity testing. We reviewed the principles of developmental toxicology, the development of alternative methods, and the application of embryonic stem cells as a tool for designing animal-free alternative methods in this area of toxicity testing. We carried out a systematic scoping review of published literature on EST applications, and highlighted three case studies, pertaining to 5-fluorouracil, thalidomide and caffeine. Finally, we placed EST research within the broader perspective of validation and application in the regulatory context of pharmaceuticals (ICH) and chemicals (OECD). In conclusion, the expanding widespread attention to EST models marks their continuing central place as a tool in animal-free developmental toxicity testing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: T.B.K. is Editor in Chief of Current Research in Toxicology. G.P.D., K. T., and H.K. are founding members of the journal's editorial board. These individuals were not involved in the editorial process or peer review evaluation of the submission. The authors declare that they have no other known competing financial interests that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crtox.2022.100074.

References

Adler, S., Pellizzer, C., Hareng, L., Hartung, T., Bremer, S., 2008. First steps in establishing a developmental toxicity test method based on human embryonic stem cells. Toxicol. In Vitro 22, 200–211.

Affleck, J.G., Walker, V.K., 2019. Drosophila as a model for developmental toxicology: using and extending the drosophotoxicology model. Methods Mol. Biol. 1965, 139–153.

Aikawa, N., 2020. A novel screening test to predict the developmental toxicity of drugs using human induced pluripotent stem cells. J. Toxicol. Sci. 45, 187–199.

Alépée, N., Adriaens, E., Abo, T., Bagley, D., Desprez, B., Hibatallah, J., Mewes, K., Pfannenbecker, U., Sala, À., Van Rompay, A.R., Verstraelen, S., McNamee, P., 2019. Development of a defined approach for eye irritation or serious eye damage for neat liquids based on cosmetics Europe analysis of in vitro RhCE and BCOP test methods. Toxicol. In Vitro 59, 100–114.

Alqahtani, S., 2017. In silico ADME-Tox modeling: progress and prospects. Expert Opin. Drug Metab. Toxicol. 13, 1147–1158.

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Mount, D.R., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serrrano, J.A., Tietge, J. E., Villeneuve, D.L., 2010. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem. 29, 730–741.

Aschner, M., Ceccatelli, S., Daneshian, M., Fritsche, E., Hasiwa, N., Hartung, T., Hogberg, H.T., Leist, M., Li, A., Mundi, W.R., Padilla, S., Piersma, A.H., Bal-Price, A., Seiler, A., Westerink, R.H., Zimmer, B., Lein, P.J., 2017. Reference compounds for alternative test methods to indicate developmental neurotoxicity (DNT) potential of chemicals: example lists and criteria for their selection and use. ALTEX 34, 49–74.

Augustine-Rauch, K., Zhang, C.X., Panzica-Kelly, J.M., 2016. A developmental toxicology assay platform for screening teratogenic liability of pharmaceutical compounds. Birth Defects Res. B Dev. Reprod. Toxicol. 107, 4–20.

Baek, D.H., Kim, T.G., Lim, H.K., Kang, J.W., Seong, S.K., Choi, S.E., Lim, S.Y., Park, S.H., Nam, B.H., Kim, E.H., Kim, M.S., Park, K.L., 2012. Embryotoxicity assessment of developmental neurotoxicants using a neuronal endpoint in the embryonic stem cell test. J. Appl. Toxicol. 32, 617–626.

Baker, N.C., Sipes, N.S., Franzosa, J., Belair, D.G., Abbott, B.D., Judson, R.S., Knudsen, T. B., 2020. Characterizing cleft palate toxicants using ToxCast data, chemical structure, and the biomedical literature. Birth Defects Res. 112, 19–39.

Baker, N., Knudsen, T., Williams, A.J., 2017. Abstract Sifter: a comprehensive front-end system to PubMed. F1000Res 6.

Ball, J.S., Stedman, D.B., Hillegass, J.M., Zhang, C.X., Panzica-Kelly, J., Coburn, A., Enright, B.P., Tornesi, B., Amouzadeh, H.R., Hetheridge, M., Gustafson, A.L., Augustine-Rauch, K.A., 2014. Fishing for teratogens: a consortium effort for a harmonized zebrafish developmental toxicology assay. Toxicol. Sci. 139, 210–219.

Barrier, M., Chandler, K., Jeffay, S., Hoopes, M., Knudsen, T., Hunter, S., 2012. Mouse embryonic stem cell adherent cell differentiation and cytotoxicity assay. Methods Mol. Biol. 889, 181–195.

Becker, A.J., Mcculloch, E.A., Till, J.E., 1963. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. Nature 197, 452–454.

Beekhuijzen, M., de Koning, C., Flores-Guillén, M.-E., de Vries-Buitenweg, S., Tobor-Kaplon, M., van de Waart, B., Emmen, H., 2015. From cutting edge to guideline: A first step in harmonization of the zebrafish embryotoxicity test (ZET) by describing the most optimal test conditions and morphology scoring system. Reprod. Toxicol. 56, 64–76.

Berg, C., 2019. The Xenopus tropicalis model for studies of developmental and reproductive toxicity. Methods Mol. Biol. 1965, 173–186.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H., Gnirke, A., Eggan, K., Meissner, A., 2011. Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell 144 (3), 439–452.

Brannen, K.C., Chapin, R.E., Jacobs, A.C., Green, M.L., 2016. Alternative models of developmental and reproductive toxicity in pharmaceutical risk assessment and the 3Rs. ILAR J. 57, 144–156.

Brown, N.A., 2002. 'Selection of test chemicals for the ECVAM international validation study on in vitro embryotoxicity tests. European Centre for the Validation of Alternative Methods. Altern. Lab. Anim. 30, 177–198.
Brown, N.A., 1987. Teratogenicity testing in vitro: status of validation studies. Arch. Toxicol. Suppl. 11, 105–114.
Buesen, R., Genschow, E., Slawik, B., Visan, A., Spielmann, H., Luch, A., Seiler, A., 2009. Embryonic stem cell test remastered: comparison between the validated EST and the new molecular FACS-EST for assessing developmental toxicity in vitro. Toxicol. Sci. 108, 389–400.
Cappelletti, S., Piacentino, D., Fineschi, V., Frati, P., Cipolloni, L., Aromatario, M., 2018. Caffeine-related deaths: manner of deaths and categories at risk. Nutrients 10 (5), 611.
Casale, F., Canaparo, R., Serpe, L., Muntoni, E., Pepa, C.D., Costa, M., Mairone, L., Zara, G.P., Fornari, G., Eandi, M., 2004. Plasma concentrations of 5-fluorouracil and its metabolites in colon cancer patients. Pharmacol. Res. 50, 173–179.
Cassar, S., Beekhuijzen, M., Beyer, B., Chapin, R., Dorau, M., Hoberman, A., Krupp, E., Leconte, I., Stedman, D., Stethem, C., van den Oetelaar, D., Tornesi, B., 2019. A multi-institutional study benchmarking the zebrafish developmental assay for prediction of embryotoxic plasma concentrations from rat embryo-fetal development studies. Reprod. Toxicol. 86, 33–44.
Cezar, G.G., Quam, J.A., Smith, A.M., Rosa, G.J., Piekarczyk, M.S., Brown, J.F., Gage, F. H., Muotri, A.R., 2007. Identification of small molecules from human embryonic stem cells using metabolomics. Stem Cells Dev. 16, 869–882.
Chandler, K.J., Barrier, M., Jeffay, S., Nichols, H.P., Kleinstreuer, N.C., Singh, A.V., Reif, D.M., Sipes, N.S., Judson, R.S., Dix, D.J., Kavlock, R., Hunter 3rd, E.S., Knudsen, T. B., 2011. Evaluation of 309 environmental chemicals using a mouse embryonic stem cell adherent cell differentiation and cytotoxicity assay. PLoS ONE 6, e18540.
Chang, K.H., Zandstra, P.W., 2004. Quantitative screening of embryonic stem cell differentiation: endoderm formation as a model. Biotechnol. Bioeng. 88, 287–298.
Chapin, R., Stedman, D., Paquette, J., Streck, R., Kumpf, S., Deng, S., 2007. Struggles for equivalence: in vitro developmental toxicity model evolution in pharmaceuticals in 2006. Toxicol. In Vitro 21 (8), 1545–1551.
Chen, X., Hansen, D.K., Merry, G., DeJarnette, C., Nolen, G., Sloper, D., Fisher, J.E., Harrouk, W., Tassinari, M.S., Inselman, A.L., 2015. Developing osteoblasts as an endpoint for the mouse embryonic stem cell test. Reprod. Toxicol. 53, 131–140.
Cheng, W., Zhou, R., Liang, F., Wei, H., Feng, Y., Wang, Y., 2016. Application of mouse embryonic stem cell test to detect gender-specific effect of chemicals: a supplementary tool for embryotoxicity prediction. Chem. Res. Toxicol. 29 (9), 1519–1533.
Collins, F.S., Gray, G.M., Bucher, J.R., 2008. Toxicology. Transforming environmental health protection. Science 319, 906–907.
Cornwall-Scoones, J, Zernicka-Gotez, M, 2021. Unifying synthetic embryology. Dev Biol. 474, 1–4.
Dai, F, Zheng, Y.L., Zhou, B, 2020. Inhibiting NF-κB-Mediated Inflammation by Catechol-Type Diphenylbutadiene via an Intracellular Copper- and Iron-Dependent Pro-Oxidative Role. J Agric Food Chem. 68 (37), 10029–10035.
Dang, S.M., Kyba, M., Perlingeiro, R., Daley, G.Q., Zandstra, P.W., 2002. Efficiency of embryoid body formation and hematopoietic development from embryonic stem cells in different culture systems. Biotechnol. Bioeng. 78, 442–453.
Daston, G.P., Beyer, B.K., Carney, E.W., Chapin, R.E., Friedman, J.M., Piersma, A.H., Rogers, J.M., Scialli, A.R., 2014. Exposure-based validation list for developmental toxicity screening assays. Birth Defects Res. B Dev. Reprod. Toxicol. 101, 423–428.
de Leeuw, V.C., Hessel, E.V.S., Piersma, A.H., 2019. Look-alikes may not act alike: gene expression regulation and cell-type-specific responses of three valproic acid analogues in the neural embryonic stem cell test (ESTn). Toxicol. Lett. 303, 28–37.
Desai, N., Rambhia, P., Gishto, A., 2015. Human embryonic stem cell cultivation: historical perspective and evolution of xeno-free culture systems. Reprod. Biol. Endocrinol. 13, 9.
Dimopoulou, M., Verhoef, A., Gomes, C.A., van Dongen, C.W., Rietjens, I.M.C.M., Piersma, A.H., van Ravenzwaay, B., 2018. A comparison of the embryonic stem cell test and whole embryo culture assay combined with the BeWo placental passage model for predicting the embryotoxicity of azoles. Toxicol. Lett. 286, 10–21.
Dong, C, Fischer, L.A., Theunissen, T.W., 2019. Recent insights into the naïve state of human pluripotency and its applications. Exp Cell Res. 385 (1), 111645.
Eriksson, T., Björkman, S., Höglund, P., 2001. Clinical pharmacology of thalidomide. Eur. J. Clin. Pharmacol. 57, 365–376.
Evans, M.J., Kaufman, M.H., 1981. Establishment in culture of pluripotential cells from mouse embryos. Nature 292, 154–156.
Ferreira, L.M., Mostajo-Radji, M.A., 2013. How induced pluripotent stem cells are redefining personalized medicine. Gene 520, 1–6.
Flamier, A., Singh, S., Rasmussen, T.P., 2017. A standardized human embryoid body platform for the detection and analysis of teratogens. PLoS ONE 12, e0171101.
Fragki, S., Piersma, A.H., Rorije, E., Zeilmaker, M.J., 2017. In vitro to in vivo extrapolation of effective dosimetry in developmental toxicity testing: application of a generic PBK modelling approach. Toxicol. Appl. Pharmacol. 332, 109–120.
Friedman, J.M., 2010. The principles of teratology: are they still true? Birth Defects Res. A Clin. Mol. Teratol. 88, 766–768.
Fu, L.-J., Johnson, E.M., Newman, L.M., 1990. Prediction of the developmental toxicity hazard potential of halogenated drinking water disinfection by-products tested by the in vitro hydra assay. Regul. Toxicol. Pharm. 11 (3), 213–219.
Funakoshi, S, Yoshida, Y, 2021. Recent progress of iPSC technology in cardiac diseases. Arch Toxicol. 95 (12), 3633–3650.
Gao, T, Lin, M, Shao, B, Zhou, Q, Wang, Y, Chen, X, Zhao, D, Dai, X, Shen, C, Cheng, H, Yang, S, Li, H, Zheng, B, Zhong, X, Yu, J, Chen, L, Huang, X, 2020. BMI1 promotes steroidogenesis through maintaining redox homeostasis in mouse MLTC-1 and primary Leydig cells. Cell Cycle 19 (15), 1884–1898.
Genschow, E., Scholz, G., Brown, N., Piersma, A., Brady, M., Clemann, N., Huuskonen, H., Paillard, F., Bremer, S., Becker, K., Spielmann, H., 2000. Development of prediction models for three in vitro embryotoxicity tests in an ECVAM validation study. In Vitr Mol. Toxicol. 13, 51–66.
Genschow, E., Spielmann, H., Scholz, G., Pohl, I., Seiler, A., Clemann, N., Bremer, S., Becker, K., 2004. Validation of the embryonic stem cell test in the international ECVAM validation study on three in vitro embryotoxicity tests. Altern. Lab. Anim. 32, 209–244.
Genschow, E., Spielmann, H., Scholz, G., Seiler, A., Brown, N., Piersma, A., Brady, M., Clemann, N., Huuskonen, H., Paillard, F., Bremer, S., Becker, K., 2002. The ECVAM international validation study on in vitro embryotoxicity tests: results of the definitive phase and evaluation of prediction models. European Centre for the Validation of Alternative Methods. Altern. Lab. Anim. 30, 151–176.
Gustafson, A.L., Stedman, D.B., Ball, J., Hillegass, J.M., Flood, A., Zhang, C.X., Panzica-Kelly, J., Cao, J., Coburn, A., Enright, B.P., Tornesi, M.B., Hetheridge, M., Augustine-Rauch, K.A., 2012. Inter-laboratory assessment of a harmonized zebrafish developmental toxicology assay - progress report on phase I. Reprod. Toxicol. 33, 155–164.
Hareng, L., Pellizzer, C., Bremer, S., Schwarz, M., Hartung, T., 2005. The integrated project ReProTect: a novel approach in reproductive toxicity hazard assessment. Reprod. Toxicol. 20 (3), 441–452.
Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., Hoffmann, S., Roi, A.J., Prieto, P., Sabbioni, E., Scott, L., Worth, A., Zuang, V., 2004. A modular approach to the ECVAM principles on test validity. Altern. Lab. Anim. 32, 467–472.
Harvey, J.P., Sladen, P.E., Yu-Wai-Man, P, Cheetham, M.E., 2022. Induced Pluripotent Stem Cells for Inherited Optic Neuropathies-Disease Modeling and Therapeutic Development. J Neuroophthalmol. 42 (1), 35–44.
Hessel, E.V.S., Staal, Y.C.M., Piersma, A.H., 2018. Design and validation of an ontology-driven animal-free testing strategy for developmental neurotoxicity testing. Toxicol. Appl. Pharmacol. 354, 136–152.
Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (ed.)ˆ(eds.). 2011. *Cochrane Handbook for Systematic Reviews of Interventions version 5.1.0.*
Howard, B.E., Phillips, J., Tandon, A., Maharana, A., Elmore, R., Mav, D., Sedykh, A., Thayer, K., Merrick, B.A., Walker, V., Rooney, A., Shah, R.R., 2020. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. Environ. Int. 138, 105623.
Hutson, M.S., Leung, M.C.K., Baker, N.C., Spencer, R.M., Knudsen, T.B., 2017. Computational model of secondary palate fusion and disruption. Chem. Res. Toxicol. 30, 965–979.
Jagtap, S., Meganathan, K., Gaspar, J., Wagh, V., Winkler, J., Hescheler, J., Sachinidis, A., 2011. Cytosine arabinoside induces ectoderm and inhibits mesoderm expression in human embryonic stem cells during multilineage differentiation. Br. J. Pharmacol. 162, 1743–1756.
Jennings, P., 2015. The future of in vitro toxicology. Toxicol. In Vitro 29, 1217–1221.
Juberg, D.R., Knudsen, T.B., Sander, M., Beck, N.B., Faustman, E.M., Mendrick, D.L., Fowle 3rd, J.R., Hartung, T., Tice, R.R., Lemazurier, E., Becker, R.A., Fitzpatrick, S. C., Daston, G.P., Harrill, A., Hines, R.N., Keller, D.A., Lipscomb, J.C., Watson, D., Bahadori, T., Crofton, K.M., 2017. FutureTox III: bridges for translation. Toxicol. Sci. 155, 22–31.
Judson, R., Houck, K., Martin, M., Knudsen, T., Thomas, R.S., Sipes, N., Shah, I., Wambaugh, J., Crofton, K., 2014. In vitro and modelling approaches to risk assessment from the U.S. Environmental Protection Agency ToxCast programme. Basic Clin. Pharmacol. Toxicol. 115, 69–76.
Judson, R., Kavlock, R., Martin, M., Reif, D., Houck, K., Knudsen, T., Richard, A., Tice, R. R., Whelan, M., Xia, M., Huang, R., Austin, C., Daston, G., Hartung, T., Fowle 3rd, J. R., Wooge, W., Tong, W., Dix, D., 2013. Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. ALTEX 30, 51–56.
Judson, R.S., Houck, K.A., Kavlock, R.J., Knudsen, T.B., Martin, M.T., Mortensen, H.M., Reif, D.M., Rotroff, D.M., Shah, I., Richard, A.M., Dix, D.J., 2010. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. Environ. Health Perspect. 118, 485–492.
Kameoka, S., Babiarz, J., Kolaja, K., Chiao, E., 2014. A high-throughput screen for teratogens using human pluripotent stem cells. Toxicol. Sci. 137, 76–90.
Kang, H.Y., Choi, Y.K., Jo, N.R., Lee, J.H., Ahn, C., Ahn, I.Y., Kim, T.S., Kim, K.S., Choi, K.C., Lee, J.K., Lee, S.D., Jeung, E.B., 2017. Advanced developmental toxicity test method based on embryoid body's area. Reprod. Toxicol. 72, 74–85.
Kavlock, R., Chandler, K., Houck, K., Hunter, S., Judson, R., Kleinstreuer, N., Knudsen, T., Martin, M., Padilla, S., Reif, D., Richard, A., Rotroff, D., Sipes, N., Dix, D., 2012. Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. Chem. Res. Toxicol. 25 (7), 1287–1302.
Kavlock, R., Dix, D., 2010. Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. J. Toxicol. Environ. Health B Crit. Rev. 13, 197–217.
Kawamura, S., Horie, N., Okahashi, N., Higuchi, H., 2019. Implications for the predictivity of cell-based developmental toxicity assays developed two decades apart. Toxicol. Res. 35, 343–351.
Kleinstreuer, N.C., Smith, A.M., West, P.R., Conard, K.R., Fontaine, B.R., Weir-Hauptman, A.M., Palmer, J.A., Knudsen, T.B., Dix, D.J., Donley, E.L., Cezar, G.G., 2011. Identifying developmental toxicity pathways for a subset of ToxCast chemicals using human embryonic stem cells and metabolomics. Toxicol. Appl. Pharmacol. 257, 111–121.
Knudsen, T.B., Keller, D.A., Sander, M., Carney, E.W., Doerrer, N.G., Eaton, D.L., Fitzpatrick, S.C., Hastings, K.L., Mendrick, D.L., Tice, R.R., Watkins, P.B., Whelan,

M., . FutureTox II: in vitro data and in silico models for predictive toxicology. Toxicol. Sci. 143, 256–267.
Knudsen, T.B., Klieforth, B., Slikker Jr., W., 2017. Programming microphysiological systems for children's health protection. Exp. Biol. Med. (Maywood) 242, 1586–1592.
Knudsen, Thomas B, Suzanne Compton Fitzpatrick, K Nadira De Abrew, Linda S Birnbaum, Anne Chappelle, George P Daston, Dana C Dolinoy, Alison Elder, Susan Euling, Elaine M Faustman, Kristi Pullen Fedinick, Jill A Franzosa, Derik E Haggard, Laurie Haws, Nicole C Kleinstreuer, Germaine M Buck Louis, Donna L Mendrick, Ruthann Rudel, Katerine S Saili, Thaddeus T Schug, Robyn L Tanguay, Alexandra E Turley, Barbara A Wetmore, Kimberly W White, and Todd J Zurlinden. 2021. FutureTox IV workshop summary: predictive toxicology for healthy children, *Toxicol. Sci.*
Knudsen, T.B., Spencer, R.M., Pierro, J.D., Baker, N.C., 2020. Computational biology and in silico toxicodynamics. Curr. Opin. Toxicol. 23–24, 119–126.
Knudsen, Thomas, Matthew Martin, Kelly Chandler, Nicole Kleinstreuer, Richard Judson, and Nisha Sipes. 2013. 'Predictive Models and Computational Toxicology.' in Paul C. Barrow (ed.), *Teratogenicity Testing: Methods and Protocols* (Humana Press: Totowa, NJ).
Kohl, C., McIntosh, E.J., Unger, S., Haddaway, N.R., Kecke, S., Schiemann, J., Wilhelm, R., 2018. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. Environ. Evid. 7, 8.
Krug, A.K., Kolde, R., Gaspar, J.A., Rempel, E., Balmer, N.V., Meganathan, K., Vojnits, K., Baquié, M., Waldmann, T., Ensenat-Waser, R., Jagtap, S., Evans, R.M., Julien, S., Peterson, H., Zagoura, D., Kadereit, S., Gerhard, D., Sotiriadou, I., Heke, M., Natarajan, K., Henry, M., Winkler, J., Marchan, R., Stoppini, L., Bosgra, S., Westerhout, J., Verwei, M., Vilo, J., Kortenkamp, A., Hescheler, J., Hothorn, L., Bremer, S., van Thriel, C., Krause, K.-H., Hengstler, J.G., Rahnenführer, J., Leist, M., Sachinidis, A., 2013. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. Arch. Toxicol. 87 (1), 123–143.
Kugler, J., Kemler, R., Luch, A., Oelgeschläger, M., 2016. Editor's highlight: identification and characterization of teratogenic chemicals using embryonic stem cells isolated from a Wnt/β-catenin-reporter transgenic mouse line. Toxicol. Sci. 152 (2), 382–394.
Kugler, J., Huhse, B., Tralau, T., Luch, A., 2017. Embryonic stem cells and the next generation of developmental toxicity testing. Expert Opin. Drug Metab. Toxicol. 13, 833–841.
Laschinski, G., Vogel, R., Spielmann, H., 1991. Cytotoxicity test using blastocyst-derived euploid embryonal stem cells: a new approach to in vitro teratogenesis screening. Reprod. Toxicol. 5 (1), 57–64.
Lau, C., Mole, M.L., Copeland, M.F., Rogers, J.M., Kavlock, R.J., Shuey, D.L., Cameron, A.M., Ellis, D.H., Logsdon, T.R., Merriman, J., Setzer, R.W., 2001. Toward a biologically based dose-response model for developmental toxicity of 5-fluorouracil in the rat: acquisition of experimental data. Toxicol. Sci. 59 (1), 37–48.
Le Coz, F., Suzuki, N., Nagahori, H., Omori, T., Saito, K., 2015. Hand1-Luc embryonic stem cell test (Hand1-Luc EST): a novel rapid and highly reproducible in vitro test for embryotoxicity by measuring cytotoxicity and differentiation toxicity using engineered mouse ES cells. J. Toxicol. Sci. 40, 251–261.
Lee, J.H., Park, S.Y., Ahn, C., Kim, C.W., Kim, J.E., Jo, N.R., Kang, H.Y., Yoo, Y.M., Jung, E.M., Kim, E.M., Kim, K.S., Choi, K.C., Lee, S.D., Jeung, E.B., 2019. Pre-validation study of alternative developmental toxicity test using mouse embryonic stem cell-derived embryoid bodies. Food Chem. Toxicol. 123, 50–56.
Lee, J.H., Park, S.Y., Ahn, C., Yoo, Y.M., Kim, C.W., Kim, J.E., Jo, N.R., Kang, H.Y., Jung, E.M., Kim, K.S., Choi, K.C., Lee, S.D., Jeung, E.B., 2020. Second-phase validation study of an alternative developmental toxicity test using mouse embryonic stem cell-derived embryoid bodies. J. Physiol. Pharmacol. 71.
Liang, G., Zhang, Y., 2013. Genetic and epigenetic variations in iPSCs: potential causes and implications for application. Cell Stem Cell 13, 149–159.
Liu, S., Yin, N., Faiola, F., 2017. Prospects and frontiers of stem cell toxicology. Stem Cells Dev. 26, 1528–1539.
Loebel, D.A., Watson, C.M., De Young, R.A., Tam, P.P., 2003. Lineage choice and differentiation in mouse embryos and embryonic stem cells. Dev. Biol. 264, 1–14.
Louisse, J., Beekmann, K., Rietjens, I.M.C.M., 2017. Use of physiologically based kinetic modeling-based reverse dosimetry to predict in vivo toxicity from in vitro data. Chem. Res. Toxicol. 30, 114–125.
Madrid, J.V., Sera, S.R., Sparks, N.R.L., Zur Nieden, N.I., 2018. Human pluripotent stem cells to assess developmental toxicity in the osteogenic lineage. Methods Mol. Biol. 1797, 125–145.
Marikawa, Y., Chen, H.R., Menor, M., Deng, Y., Alarcon, V.B., 2020. Exposure-based assessment of chemical teratogenicity using morphogenetic aggregates of human embryonic stem cells. Reprod. Toxicol. 91, 74–91.
Martello, G., Smith, A., 2014. The nature of embryonic stem cells. Annu. Rev. Cell Dev. Biol. 30, 647–675.
Martin, G.R., 1981. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc. Natl. Acad. Sci. U S A 78, 7634–7638.
Martin, G.R., Evans, M.J., 1975. Differentiation of clonal lines of teratocarcinoma cells: formation of embryoid bodies in vitro. Proc. Natl. Acad. Sci. U S A 72, 1441–1445.
Marx-Stoelting, P., Adriaens, E., Ahr, H.J., Bremer, S., Garthoff, B., Gelbke, H.P., Piersma, A., Pellizzer, C., Reuter, U., Rogiers, V., Schenk, B., Schwengberg, S., Seiler, A., Spielmann, H., Steemans, M., Stedman, D.B., Vanparys, P., Vericat, J.A., Verwei, M., van der Water, F., Weimer, M., Schwarz, M., 2009. 'A review of the implementation of the embryonic stem cell test (EST). The report and recommendations of an ECVAM/ReProTect Workshop. Altern. Lab Anim 37, 313–328.
Matyskiela, M.E., Couto, S., Zheng, X., Lu, G., Hui, J., Stamp, K., Drew, C., Ren, Y., Wang, M., Carpenter, A., Lee, C.-W., Clayton, T., Fang, W., Lu, C.-C., Riley, M., Abdubek, P., Blease, K., Hartke, J., Kumar, G., Vessey, R., Rolfe, M., Hamann, L.G., Chamberlain, P.P., 2018. SALL4 mediates teratogenicity as a thalidomide-dependent cereblon substrate. Nat. Chem. Biol. 14 (10), 981–987.
McNally, K., Hogg, A., Loizou, G., 2018. A computational workflow for probabilistic quantitative in vitro to in vivo extrapolation. Front. Pharmacol. 9, 508.
Meganathan, K., Jagtap, S., Wagh, V., Winkler, J., Gaspar, J.A., Hildebrand, D., Trusch, M., Lehmann, K., Hescheler, J., Schlüter, H., Sachinidis, A., 2012. Identification of thalidomide-specific transcriptomics and proteomics signatures during differentiation of human embryonic stem cells. PLoS ONE 7, e44228.
Meisig, J., N. Dreser, M. Kapitza, M. Henry, T. Rotshteyn, J. Rahnenführer, et al. 2020. Kinetic modeling of stem cell transcriptome dynamics to identify regulatory modules of normal and disturbed neuroectodermal differentiation. Nucleic Acids Res. 48(22): 12577-12592.
Mennen, R.H.G., Pennings, J., Piersma, A.H.A., 2019. Neural crest related gene transcript regulation by valproic acid analogues in the cardiac embryonic stem cell test. Reprod. Toxicol. 90, 44–52.
Moher, D., Stewart, L., Shekelle, P., 2015. All in the Family: systematic reviews, rapid reviews, scoping reviews, realist reviews, and more. Syst. Rev. 4, 183.
Moon, S.H., Ju, J., Park, S.J., Bae, D., Chung, H.M., Lee, S.H., 2014. Optimizing human embryonic stem cells differentiation efficiency by screening size-tunable homogenous embryoid bodies. Biomaterials 35, 5987–5997.
Murugan, V, 2009. Embryonic Stem Cell Research: A Decade of Debate from Bush to Obama. Yale J Biol Med. 82 (3), 101–103.
Nagahori, H., Suzuki, N., Le Coz, F., Omori, T., Saito, K., 2016. Prediction of in vivo developmental toxicity by combination of Hand1-Luc embryonic stem cell test and metabolic stability test with clarification of metabolically inapplicable candidates. Toxicol. Lett. 259, 44–51.
Nath, S.C., Horie, M., Nagamori, E., Kino-Oka, M., 2017. Size- and time-dependent growth properties of human induced pluripotent stem cells in the culture of single aggregate. J. Biosci. Bioeng. 124, 469–475.
Newall, D.R., Beedles, K.E., 1996. The stem-cell test: an in vitro assay for teratogenic potential. Results of a blind trial with 25 compounds. Toxicol. In Vitro 10 (2), 229–240.
NRC. 2007. "Toxicity Testing in the 21st Century: A Vision and a Strategy." In. Washington, DC.
Padilla, S., Corum, D., Padnos, B., Hunter, D.L., Beam, A., Houck, K.A., Sipes, N., Kleinstreuer, N., Knudsen, T., Dix, D.J., Reif, D.M., 2012. Zebrafish developmental screening of the ToxCast™ Phase I chemical library. Reprod. Toxicol. 33, 174–187.
Palmer, J.A., Smith, A.M., Egnash, L.A., Colwell, M.R., Donley, E.L.R., Kirchner, F.R., Burrier, R.E., 2017. A human induced pluripotent stem cell-based in vitro assay predicts developmental toxicity through a retinoic acid receptor-mediated pathway for a series of related retinoid analogues. Reprod. Toxicol. 73, 350–361.
Palmer, J.A., Smith, A.M., Egnash, L.A., Conard, K.R., West, P.R., Burrier, R.E., Donley, E.L., Kirchner, F.R., 2013. Establishment and assessment of a new human embryonic stem cell-based biomarker assay for developmental toxicity screening. Birth Defects Res. B Dev. Reprod. Toxicol. 98, 343–363.
Panzica-Kelly, J.M., Brannen, K.C., Ma, Y., Zhang, C.X., Flint, O.P., Lehman-McKeeman, L.D., Augustine-Rauch, K.A., 2013. Establishment of a molecular embryonic stem cell developmental toxicity assay. Toxicol. Sci. 131, 447–457.
Paquette, J.A., Kumpf, S.W., Streck, R.D., Thomson, J.J., Chapin, R.E., Stedman, D.B., 2008. Assessment of the Embryonic Stem Cell Test and application and use in the pharmaceutical industry. Birth Defects Res. B Dev. Reprod. Toxicol. 83, 104–111.
Peters, A.K., Steemans, M., Hansen, E., Mesens, N., Verheyen, G.R., Vanparys, P., 2008. Evaluation of the embryotoxic potency of compounds in a newly revised high throughput embryonic stem cell test. Toxicol. Sci. 105, 342–350.
Pettinato, G., Wen, X., Zhang, N., 2015. Engineering strategies for the formation of embryoid bodies from human pluripotent stem cells. Stem Cells Dev. 24, 1595–1609.
Piersma, A.H., Bosgra, S., van Duursen, M.B., Hermsen, S.A., Jonker, L.R., Kroese, E.D., van der Linden, S.C., Man, H., Roelofs, M.J., Schulpen, S.H., Schwarz, M., Uibel, F., van Vugt-Lussenburg, B.M., Westerhout, J., Wolterbeek, A.P., van der Burg, B., 2013. Evaluation of an alternative in vitro test battery for detecting reproductive toxicants. Reprod. Toxicol. 38, 53–64.
Piersma, A.H., Burgdorf, T., Louekari, K., Desprez, B., Taalman, R., Landsiedel, R., Barroso, J., Rogiers, V., Eskes, C., Oelgeschläger, M., Whelan, M., Braeuning, A., Vinggaard, A.M., Kienhuis, A., van Benthem, J., Ezendam, J., 2018a. Workshop on acceleration of the validation and regulatory acceptance of alternative methods and implementation of testing strategies. Toxicol. In Vitro 50, 62–74.
Piersma, A.H., Genschow, E., Verhoef, A., Spanjersberg, M.Q.I., Brown, N.A., Brady, M., Burns, A., Clemann, N., Seiler, A., Spielmann, H., 2004. Validation of the postimplantation rat whole-embryo culture test in the international ECVAM validation study on three in vitro embryotoxicity tests. Altern. Lab. Anim. 32 (3), 275–307.
Piersma, A.H., van Benthem, J., Ezendam, J., Staal, Y.C.M., Kienhuis, A.S., 2019. The virtual human in chemical safety assessment. Curr. Opin. Toxicol. 15, 26–32.
Piersma, A.H., 2006. Alternative methods for developmental toxicity testing. Basic Clin. Pharmacol. Toxicol. 98, 427–431.
Piersma, A.H., Ezendam, J., Mirjam Luijten, J.J., Muller, A., Rorije, E., van der Ven, L.T. M., van Benthem, J., 2014. A critical appraisal of the process of regulatory implementation of novel in vivo and in vitro methods for chemical hazard and risk assessment. Crit. Rev. Toxicol. 24, 1–19.

Piersma, A.H., van Benthem, J., Ezendam, J., Kienhuis, A.S., 2018b. Validation redefined. Toxicol. In Vitro 46, 163–215.
Przybylak, K.R., Madden, J.C., Covey-Crump, E., Gibson, L., Barber, C., Patel, M., Cronin, M.T.D., 2018. Characterisation of data resources for in silico modelling: benchmark datasets for ADME properties. Expert Opin. Drug Metab. Toxicol. 14, 169–181.
Punt, A., Bouwmeester, H., Schiffelers, M.W.A., Peijnenburg, A., 2018. Expert opinions on the acceptance of alternative methods in food safety evaluations: Formulating recommendations to increase acceptance of non-animal methods for kinetics. Regul. Toxicol. Pharm. 92, 145–151.
Richard, A.M., Judson, R.S., Houck, K.A., Grulke, C.M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M.T., Wambaugh, J.F., Knudsen, T.B., Kancherla, J., Mansouri, K., Patlewicz, G., Williams, A.J., Little, S.B., Crofton, K.M., Thomas, R.S., 2016. ToxCast chemical landscape: paving the road to 21st century toxicology. Chem. Res. Toxicol. 29, 1225–1251.
Rico-Varela, J., Ho, D., Wan, L.Q., 2018. In Vitro Microscale Models for Embryogenesis. Adv Biosyst, p. 2.
Riebeling, C., Hayess, K., Peters, A.K., Steemans, M., Spielmann, H., Luch, A., Seiler, A.E. M., 2012. Assaying embryotoxicity in the test tube: current limitations of the embryonic stem cell test (EST) challenging its applicability domain. Crit. Rev. Toxicol. 42 (5), 443–464.
Robinson, J.F., Piersma, A.H., 2013. Toxicogenomic approaches in developmental toxicology testing. Methods Mol. Biol. 947, 451–473.
Rovida, C., Vivier, M., Garthoff, B., Hescheler, J., 2014. ESNATS conference - the use of human embryonic stem cells for novel toxicity testing approaches. Altern. Lab. Anim. 42, 97–113.
Saili, K.S., Antonijevic, T., Zurlinden, T.J., Shah, I., Deisenroth, C., Knudsen, T.B., 2020. Molecular characterization of a toxicological tipping point during human stem cell differentiation. Reprod. Toxicol. 91, 1–13.
Sachinidis, A., Albrecht, W., Nell, P., Cherianidou, A., Hewitt, N.J., Edlund, K., Hengstler, J.G., 2019. Road map for development of stem cell-based alternative test methods. Trends Mol. Med. 25, 470–481.
Schmidt, B.Z., Lehmann, M., Gutbier, S., Nembo, E., Noel, S., Smirnova, L., Forsby, A., Hescheler, J., Avci, H.X., Hartung, T., Leist, M., Kobolák, J., Dinnyés, A., 2017. In vitro acute and developmental neurotoxicity screening: an overview of cellular platforms and high-throughput technical possibilities. Arch. Toxicol. 91 (1), 1–33.
Scholz, G., Genschow, E., Pohl, I., Bremer, S., Paparella, M., Raabe, H., Southee, J., Spielmann, H., 1999. Prevalidation of the embryonic stem cell test (EST)-a new in vitro embryotoxicity test. Toxicol. In Vitro 13, 675–681.
Scialli, Anthony R., George Daston, Connie Chen, Prägati S. Coder, Susan Y. Euling, Jennifer Foreman, Alan M. Hoberman, Julia Hui, Thomas Knudsen, Susan L. Makris, LaRonda Morford, Aldert H. Piersma, Dinesh Stanislaus, and Kary E. Thompson. 2018. 'Rethinking developmental toxicity testing: Evolution or revolution?', 110: 840-50.
Seiler, A.E., Spielmann, H., 2011. The validated embryonic stem cell test to predict embryotoxicity in vitro. Nat. Protoc. 6, 961–978.
Seiler, A., Visan, A., Buesen, R., Genschow, E., Spielmann, H., 2004. Improvement of an in vitro stem cell assay for developmental toxicity: the use of molecular endpoints in the embryonic stem cell test. Reprod. Toxicol. 18 (2), 231–240.
Setzer, R.W., Lau, C., Mole, M.L., Copeland, M.F., Rogers, J.M., Kavlock, R.J., 2001. Toward a biologically based dose-response model for developmental toxicity of 5-fluorouracil in the rat: a mathematical construct. Toxicol. Sci. 59 (1), 49–58.
Shinde, V., Klima, S., Sureshkumar, P.S., Meganathan, K., Jagtap, S., Rempel, E., Rahnenführer, J., Hengstler, J.G., Waldmann, T., Hescheler, J., Leist, M., Sachinidis, A., 2015. Human Pluripotent stem cell based developmental toxicity assays for chemical safety screening and systems biology data generation. J. Vis. Exp. e52333.
Shojania, K.G., Bero, L.A., 2001. Taking advantage of the explosion of systematic reviews: an efficient MEDLINE search strategy. Eff. Clin. Pract. 4, 157–162.
Spielmann, H., Seiler, A., Bremer, S., Hareng, L., Hartung, T., Ahr, H., Faustman, E., Haas, U., Moffat, G.J., Nau, H., Vanparys, P., Piersma, A., Sintes, J.R., Stuart, J., 2006. The practical application of three validated in vitro embryotoxicity tests. The report and recommendations of an ECVAM/ZEBET workshop (ECVAM workshop 57). Altern. Lab. Anim. 34, 527–538.
Spielmann, H., Pohl, I., Döring, B., Liebsch, M., Moldenhauer, F., 1997. The embryonic stem cell test (EST), an in vitro embryotoxicity test using two permanent mouse cell lines: 3T3 fibroblasts and embryonic stem cells. In Vitro Toxicol. 10, 119–127.
Spinu, N., Cronin, M.T.D., Enoch, S.J., Madden, J.C., Worth, A.P., 2020. Quantitative adverse outcome pathway (qAOP) models for toxicity prediction. Arch. Toxicol. 94, 1497–1510.
Staal, Y.C.M., Pennings, J.L.A., Hessel, E.V.S., Piersma, A.H., 2017. Advanced toxicological risk assessment by implementation of ontologies operationalized in computational models. Appl in vitro toxicol 3, 325–332.
Stark, M.R., Ross, M.M., 2019. The chicken embryo as a model in developmental toxicology. Methods Mol. Biol. 1965, 155–171.
Stephens, M.L., Akgun-Olmez, S.G., Hoffmann, S, de Vries, R, Flick, B, Hartung, T, Lalu, M, Maertens, A, Witters, H, Wright, R, tsaioun, K, 2018. Adaptation of the Systematic Review Framework to the Assessment of Toxicological Test Methods: Challenges and Lessons Learned with the Zebrafish Embryotoxicity Test. Toxicol Sci. 171 (1), 56–68.
Stummann, T.C., Hareng, L., Bremer, S., 2009. Hazard assessment of methylmercury toxicity to neuronal induction in embryogenesis using human embryonic stem cells. Toxicology 257, 117–126.
Sturla, S.J., Boobis, A.R., FitzGerald, R.E., Hoeng, J., Kavlock, R.J., Schirmer, K., Whelan, M., Wilks, M.F., Peitsch, M.C., 2014. Systems toxicology: from basic research to risk assessment. Chem. Res. Toxicol. 27, 314–329.
Suzuki, N., Ando, S., Yamashita, N., Horie, N., Saito, K., 2011. Evaluation of novel high-throughput embryonic stem cell tests with new molecular markers for screening embryotoxic chemicals in vitro. Toxicol. Sci. 124, 460–471.
Takahashi, K., Yamanaka, S., 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126 (4), 663–676.
Theunissen, P.T., Pennings, J.L., van Dartel, D.A., Robinson, J.F., Kleinjans, J.C., Piersma, A.H., 2013. Complementary detection of embryotoxic properties of substances in the neural and cardiac embryonic stem cell tests. Toxicol. Sci. 132, 118–130.
Theunissen, P.T., Schulpen, S.H., van Dartel, D.A., Hermsen, S.A., van Schooten, F.J., Piersma, A.H., 2010. An abbreviated protocol for multilineage neural differentiation of murine embryonic stem cells and its perturbation by methyl mercury. Reprod. Toxicol. 29, 383–392.
Thomas, Russell S, Tina Bahadori, Timothy J Buckley, John Cowden, Chad Deisenroth, Kathie L Dionisio, Jeffrey B Frithsen, Christopher M Grulke, Maureen R Gwinn, Joshua A Harrill, Mark Higuchi, Keith A Houck, Michael F Hughes, E Sidney Hunter, III, Kristin K Isaacs, Richard S Judson, Thomas B Knudsen, Jason C Lambert, Monica Linnenbrink, Todd M Martin, Seth R Newton, Stephanie Padilla, Grace Patlewicz, Katie Paul-Friedman, Katherine A Phillips, Ann M Richard, Reeder Sams, Timothy J Shafer, R Woodrow Setzer, Imran Shah, Jane E Simmons, Steven O Simmons, Amar Singh, Jon R Sobus, Mark Strynar, Adam Swank, Rogelio Tornero-Valez, Elin M Ulrich, Daniel L Villeneuve, John F Wambaugh, Barbara A Wetmore, and Antony J Williams. 2019. 'The next generation blueprint of computational toxicology at the U. S. environmental protection agency', *Toxicol. Sci.*, 169: 317-32.
Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., Jones, J.M., 1998. Embryonic stem cell lines derived from human blastocysts. Science 282, 1145–1147.
Tollefsen, K.E., Scholz, S., Cronin, M.T., Edwards, S.W., de Knecht, J., Crofton, K., Garcia-Reyero, N., Hartung, T., Worth, A., Patlewicz, G., 2014. Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). Regul. Toxicol. Pharm. 70, 629–640.
Toutain, P.L., Ferran, A., Bousquet-Mélou, A., 2010. Species differences in pharmacokinetics and pharmacodynamics. Handb. Exp. Pharmacol., 19–48.
Tronser, T., Demir, K., Reischl, M., Bastmeyer, M., Levkin, P.A., 2018. Droplet microarray: miniaturized platform for rapid formation and high-throughput screening of embryoid bodies. Lab Chip 18 (15), 2257–2269.
Tsankov, A.M., Akopian, V., Pop, R., Chetty, S., Gifford, C.A., Daheron, L., Tsankova, N. M., Meissner, A., 2015. A qPCR ScoreCard quantifies the differentiation potential of human pluripotent stem cells. Nat. Biotechnol. 33, 1182–1192.
Uibel, F., Mühleisen, A., Köhle, C., Weimer, M., Stummann, T.C., Bremer, S., Schwarz, M., 2010. ReProGlo: a new stem cell-based reporter assay aimed to predict embryotoxic potential of drugs and chemicals. Reprod. Toxicol. 30 (1), 103–112.
Uibel, F., Schwarz, M., 2015. Prediction of embryotoxic potential using the ReProGlo stem cell-based Wnt reporter assay. Reprod. Toxicol. 55, 30–49.
van Dartel, D.A., Piersma, A.H., 2011. The embryonic stem cell test combined with toxicogenomics as an alternative testing model for the assessment of developmental toxicity. Reprod. Toxicol. 32, 235–244.
van Dartel, D.A.M., Pennings, J.L.A., de la Fonteyne, L.J.J., van Herwijnen, M.H., van Delft, J.H., van Schooten, F.J., Piersma, A.H., 2010. Monitoring developmental toxicity in the embryonic stem cell test using differential gene expression of differentiation-related genes. Toxicol. Sci. 116, 130–219.
van der Burg, B., Wedebye, E.B., Dietrich, D.R., Jaworska, J., Mangelsdorf, I., Paune, E., Schwarz, M., Piersma, A.H., Kroese, E.D., 2015. The ChemScreen project to design a pragmatic alternative approach to predict reproductive toxicity of chemicals. Reprod. Toxicol. 55, 114–123.
van Gelder, Marleen M.H.J., Iris A.L.M. van Rooij, Richard K. Miller, Gerhard A. Zielhuis, Lolkje T.W. de Jong-van den Berg, and Nel Roeleveld. 2010. Teratogenic mechanisms of medical drugs, *Human Reproduction Update*, 16: 378-94.
van Oostrom, C.T., Slob, W., van der Ven, L.T., 2020. Defining embryonic developmental effects of chemical mixtures using the embryonic stem cell test. Food Chem. Toxicol. 140, 111284.
Wang, J, Zeng, H, 2021. Recent advances in electrochemical techniques for characterizing surface properties of minerals. Adv Colloid Interface Sci 288, 102346.
Warkus, E.L.L., Marikawa, Y., 2017. Exposure-based validation of an in vitro gastrulation model for developmental toxicity assays. Toxicol. Sci. 157, 235–245.
Weitzer, G., 2006. Embryonic stem cell-derived embryoid bodies: an in vitro model of eutherian pregastrulation development and early gastrulation. Handb. Exp. Pharmacol., 21–51.
West, P.R., Weir, A.M., Smith, A.M., Donley, E.L., Cezar, G.G., 2010. Predicting human developmental toxicity of pharmaceuticals using human embryonic stem cells and metabolomics. Toxicol. Appl. Pharmacol. 247, 18–27.
Williams, A.J., Grulke, C.M., Edwards, J., McEachran, A.D., Mansouri, K., Baker, N.C., Patlewicz, G., Shah, I., Wambaugh, J.F., Judson, R.S., Richard, A.M., 2017. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J. Cheminform 9, 61.
Wittwehr, C., Aladjov, H., Ankley, G., Byrne, H.J., de Knecht, J., Heinzle, E., Klambauer, G., Landesmann, B., Luijten, M., MacKay, C., Gavin Maxwell, M.E., Meek, B., Paini, A., Perkins, E., Sobanski, T., Villeneuve, D., Waters, K.M., Whelan, M., 2017. How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology. Toxicol. Sci. 155, 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. Chem. Res. Toxicol. 26, 1840–1861.

Xing, J., Toh, Y.C., Xu, S., Yu, H., 2015. A method for human teratogen detection by geometrically confined cell differentiation and migration. Sci. Rep. 5, 10038.

Zang, R.u., Xin, X., Zhang, F., Li, D., Yang, S.-T., 2019. An engineered mouse embryonic stem cell model with survivin as a molecular marker and EGFP as the reporter for high throughput screening of embryotoxic chemicals in vitro. Biotechnol. Bioeng. 116 (7), 1656–1668.

ICCVAM. 2018. "A Strategic Roadmap for Establishing New Approaches to Evaluate the Safety of Chemicals and Medical Products in the United States. ." In.

zur Nieden, N. I., L. A. Davis, and D. E. Rancourt. 2010. 'Comparing three novel endpoints for developmental osteotoxicity in the embryonic stem cell test', *Toxicol Appl Pharmacol*, 247: 91-7.

zur Nieden, N. I., L. A. Davis, and D. E. Rancourt. 2010. Monolayer cultivation of osteoprogenitors shortens duration of the embryonic stem cell test while reliably predicting developmental osteotoxicity, *Toxicology*, 277: 66-73.

zur Nieden, N. I., L. J. Ruf, G. Kempka, H. Hildebrand, and H. J. Ahr. 2001. 'Molecular markers in embryonic stem cells', *Toxicol In Vitro*, 15: 455-61.

Zeng, W.J., Lu, C, Shy, Y, Wu, C, Chen, X, Li, C, Yao, J, 2020. Initiation of stress granule assembly by rapid clustering of IGF2BP proteins upon osmotic shock. Biochim Biophys Acta Mol Cell Res. 1867 (10), 118795.

zur Nieden, N.I., Kempka, G, Ahr, H.J., 2003. In vitro differentiation of embryonic stem cells into mineralized osteoblasts. Differentiation 71 (1), 18–27.

Zurlinden, T.J., Saili, K.S., Rush, N., Kothiya, P., Judson, R.S., Houck, K.A., Hunter, E.S., Baker, N.C., Palmer, J.A., Thomas, R.S., Knudsen, T.B., 2020. Profiling the ToxCast library with a pluripotent human (H9) stem cell line-based biomarker assay for developmental toxicity. Toxicol. Sci. 174, 189–209.