




REVIEW

A Systematic Review to Compare Chemical Hazard Predictions of the Zebrafish Embryotoxicity Test With Mammalian Prenatal Developmental Toxicity

Sebastian Hoffmann ,^{*,†,1} Bianca Marigliani,[‡] Sevcan Gül Akgün-Ölmez,[§] Danielle Ireland,[¶] Rebecca Cruz,^{||} Francois Busquet,^{|||} Burkhard Flick ,^{|||} Manoj Lalu,[#] Elizabeth C. Ghandakly,^{**} Rob B.M. de Vries,^{*,††} Hilda Witters,^{‡‡} Robert A. Wright,^{§§} Metin Ölmez,^{¶¶} Catherine Willett,^{##} Thomas Hartung,^{***} Martin L. Stephens,^{*} and Katya Tsaion ^{*}

^{*}Evidence-Based Toxicology Collaboration (EBTC), Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA [†]seh consulting + services, 33106 Paderborn, Germany [‡]Department of Science and Technology, Federal University of São Paulo (UNIFESP), São José dos Campos, 12231-280 São Paulo, Brazil [§]Department of Pharmaceutical Toxicology, Faculty of Pharmacy, Marmara University, Istanbul, 34722, Turkey [¶]Department of Biology, Swarthmore College, Swarthmore, Pennsylvania 19081, USA ^{||}Laboratory of Dental Clinical Research, Universidade Federal Fluminense, Niterói, 20520-040 Rio de Janeiro, Brazil ^{|||}Altertox, 1050 Brussels, Belgium ^{|||}Experimental Toxicology and Ecology, BASF SE, 67063 Ludwigshafen am Rhein, Germany [#]Department of Anesthesiology and Pain Medicine, Ottawa Hospital Research Institute, Ottawa, K1H 8L6 Ontario, Canada ^{**}Berman Institute of Bioethics, Johns Hopkins University, Baltimore, Maryland 21205, USA ^{††}Systematic Review Centre for Laboratory Experimentation (SYRCLE), Department for Health Evidence, Radboud Institute for Health Sciences, Radboudumc, 6500HB Nijmegen, The Netherlands ^{‡‡}VITO NV, 2400 Mol, Belgium ^{§§}William H. Welch Medical Library, Johns Hopkins University, Baltimore, Maryland 21205, USA ^{¶¶}Umraniye Family Health Center (No. 44), Turkish Ministry of Health, 34760 Istanbul, Turkey ^{##}Humane Society International, Washington, 20037 District of Columbia, USA and ^{***}Center for Alternatives to Animal Testing (CAAT), Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA

¹To whom correspondence should be addressed at seh consulting + services, 33106 Paderborn, Germany. E-mail: sebastian.hoffmann@seh-cs.com.

ABSTRACT

Originally developed to inform the acute toxicity of chemicals on fish, the zebrafish embryotoxicity test (ZET) has also been proposed for assessing the prenatal developmental toxicity of chemicals, potentially replacing mammalian studies. Although extensively evaluated in primary studies, a comprehensive review summarizing the available evidence for the ZET's capacity is lacking. Therefore, we conducted a systematic review of how well the presence or absence of exposure-related findings in the ZET predicts prenatal development toxicity in studies with rats and rabbits. A two-tiered systematic

review of the developmental toxicity literature was performed, a review of the ZET literature was followed by one of the mammalian literature. Data were extracted using DistillerSR, and study validity was assessed with an amended SYRCLE's risk-of-bias tool. Extracted data were analyzed for each species and substance, which provided the basis for comparing the 2 test methods. Although limited by the number of 24 included chemicals, our results suggest that the ZET has potential to identify chemicals that are mammalian prenatal developmental toxicants, with a tendency for overprediction. Furthermore, our analysis confirmed the need for further standardization of the ZET. In addition, we identified contextual and methodological challenges in the application of systematic review approaches to toxicological questions. One key to overcoming these challenges is a transition to more comprehensive and transparent planning, conduct and reporting of toxicological studies. The first step toward bringing about this change is to create broad awareness in the toxicological community of the need for and benefits of more evidence-based approaches.

Key words: systematic review; zebrafish embryotoxicity test; prenatal developmental toxicity; test method comparison.

Prenatal developmental toxicity is a pivotal concern in chemical hazard and risk assessment. Therefore, it is an integral part of many regulatory frameworks around the globe, which usually require mammalian toxicity data according to the Test Guideline 414 of the Organisation for Economic Co-operation and Development (OECD TG 414). Some regulatory frameworks require studies in 2 mammalian species, such as the European chemical regulation REACH (Registration, Evaluation and Authorization of Chemicals) for high-volume substances. In such studies, a test substance is administered to pregnant animals (most often orally to rats and rabbits) and maternal toxicity as well as fetal structural abnormalities, altered growth, and death are measured (OECD, 2018). However, the OECD TG 414 is laborious, costly, and time consuming. Also, it requires a substantial number of animals and thereby raises ethical concerns. Because of these issues, there is momentum to develop and alternative methods for prenatal developmental safety assessments. For example, the International Council for Harmonization (ICH) guideline on the detection of reproductive toxicity for human pharmaceuticals encourages the use of *in vitro* assays to support the identification of potential hazards to embryo-fetal development (ICH, 2020).

A promising approach to study prenatal developmental effects is the zebrafish embryotoxicity test (ZET). This test is 1 product arising from the increased use of the zebrafish (*Danio rerio*) as a model organism for studying the effects of chemicals and pharmaceuticals. Simple literature searches demonstrate the exponential growth of these uses of zebrafish since the late 1990s (see, eg, for environmental health, Bambino and Chu [2017] and Cassar et al. [2020]). The increased popularity of the zebrafish model for chemical testing has been mainly driven by the zebrafish's breadth of applications, relevance to human health, and compatibility with high-throughput screening (Bambino and Chu, 2017; Cassar et al., 2020; Garcia et al., 2016; Horzmann and Freeman, 2018). In addition, the translucency of the oviparously developing zebrafish embryo, which allows direct microscopic observation throughout the entire developmental process, is an advantage for studying developmental effects.

The ZET has been developed to identify teratogenic and embryotoxic chemicals (Brannen et al., 2010; He et al., 2014; Selderslaghs et al., 2009; Ton et al., 2006; Yang et al., 2009). It focuses on the first days post-fertilization, starting chemical exposure as early as during cleavage (0.7–2.2 hours post fertilization [hpf]) and ending exposure and observations at the early larval period (approximately 72–120 hpf), when morphogenesis is mostly completed (Kimmel et al., 1995). The ZET focuses on toxic effects of test substances related to mortality and general and specific embryotoxicity (Beekhuijzen et al., 2015).

The utility of the ZET for the detection of prenatal developmental effects has been evaluated for specific classes of chemicals (Beker van Woudenberg et al., 2013; Hermesen et al., 2011), and the use of the ZET in combination with other test methods has been suggested and explored (Augustine-Rauch et al., 2016; Kroese et al., 2015; Piersma et al., 2013).

However, broader application of the ZET—when either used alone or in combination with other evidence, for example, from new approach methodologies—has been impeded by substantial differences in published protocols, especially regarding exposure (duration and concentrations); outcomes to be observed; outcome interpretation; and chorionation status (Beekhuijzen et al., 2015; Hamm et al., 2019). Such differences can lead to discrepancies among tests assessing the same substance; thus method harmonization and standardization has been called for (Nishimura et al., 2016). First steps toward the harmonization of the ZET include a promising effort led by the pharmaceutical industry toward standardization and validation (Ball et al., 2014; Gustafson et al., 2012) and the proposal of optimal test conditions (Beekhuijzen et al., 2015). More recently, the U.S. National Toxicology Program contributed to these efforts through the Systematic Evaluation of the Application of Zebrafish in Toxicology program that identified sources of variability in ZET assays (Hamm et al., 2019).

Although the ZET offers a number of compelling advantages as compared with traditional mammalian methods, a systematic assessment of its value for the evaluation of prenatal developmental effects of chemicals is lacking. An obvious choice for moving forward would be a formal validation study conducted according to internationally agreed-upon principles (OECD, 2005). This approach could build on the results obtained by Gustafson et al. (2012) and Ball et al. (2014). However, such a prospective approach entails practical and methodological challenges, such as the requirement for substantial resources and a standardized ZET protocol. To avoid the practical challenges of a prospective approach, retrospective validation has been proposed for test methods, such as the ZET, for which a substantial amount of data is already available (Balls et al., 2006). Balls et al. (2006) also proposed that systematic review methods could be applied to collect and assess existing evidence in this context. Furthermore, one would have to consider the fact that the ZET could be used in combination with other evidence as part of a testing strategy. The construction and assessment of testing strategies entails the integration of various test methods and other information sources, typically combining testing and modelling approaches addressing distinct and complementary mechanisms. Due in no small part to the daunting methodological challenges, assessment approaches for such strategies are

still being discussed (Burgdorf *et al.*, 2019; Hartung *et al.*, 2013; Piersma *et al.*, 2018).

Systematic review techniques have recently attracted substantial attention in the field of chemical risk assessment (Hoffmann *et al.*, 2017; Whaley *et al.*, 2016). Inspired by systematic reviews assessing diagnostic test accuracy (see <https://methods.cochrane.org/sdt/handbook-dta-reviews>; last accessed on June 15, 2021), we applied systematic review methods to retrospectively assess a specific toxicological test method. In the process, we addressed two main objectives: (1) to determine to what extent ZET and mammalian test results agree and (2) to explore the challenges of applying systematic review methodology to toxicological test method assessment. We chose the ZET primarily because we wanted to provide a comprehensive, systematic, and objective evaluation of its potential to inform the assessment of the prenatal developmental toxicity hazard of chemicals. We also expected that sufficient studies would be available to allow for a systematic review. Our systematic review of the ZET and mammalian literatures was guided by the following question: “How well does the presence or absence of treatment-related findings in the ZET predict the presence or absence of prenatal development toxicity in rat and rabbit studies (OECD TG 414 and equivalents)?” A preparatory study addressing this question and documenting initial lessons learned in the application of systematic review methods has been summarized by Stephens *et al.* (2019). Here, we present and discuss the results of the fully realized systematic review documented in our PROSPERO-registered protocol, with some modifications (Tsaoun *et al.*, 2018).

MATERIALS AND METHODS

Adaptations of systematic review methods to the assessment of toxicological test method performance were explored in a preparatory study (Stephens *et al.*, 2019). Based on the findings of this study, a final review protocol was registered, to which we refer for details not reported here (Tsaoun *et al.*, 2018). The protocol was based on the template for systematic reviews of animal intervention studies proposed by de Vries *et al.* (2015). We briefly describe the protocol here, highlighting and justifying any subsequent amendments.

Search strategy. Literature searches were performed using PubMed, Embase (Embase.com), BIOSIS Previews (Clarivate Analytics), and TOXLINE (National Library of Medicine). (TOXNET, which included TOXLINE, was retired on December 16, 2019 [https://www.nlm.nih.gov/pubs/techbull/nd19/nd19_toxnet_new_locations.html; last accessed on June 15, 2021]. Much of TOXLINE’s content has been migrated to PubMed, with archival content available via download [<https://www.nlm.nih.gov/toxnet/toxline-help.html>; last accessed on June 15, 2021].) There were no language or other limitations, except for a date limitation indicated below for the mammalian searches. The search strings included a combination of keywords and terms from controlled vocabularies (ie, MeSH and Emtree) and were constructed to achieve a balance of precision and recall in the results. Search strings were designed for each of the 4 databases to identify ZET and mammalian developmental toxicity studies. These search strings were developed and run in a particular sequence, with the goal of identifying 2 sets of studies—1 for ZET and 1 for mammalian tests—examining the same chemicals.

The zebrafish searches were first run in the 4 databases on June 23, 2016. These searches included concepts for species, developmental stage, and toxicity. The results of these searches

were screened for eligibility and the chemicals examined in the included studies were extracted. The mammalian searches, focused on the chemicals identified by the zebrafish searches, were then run in the 4 databases. Searches in the databases were run on July 13, July 14, and July 15, 2018. These searches covered the earliest dates in each database up to 2016, in order to match the time frame of the zebrafish searches, and included concepts for species, developmental stage, toxicity, and chemicals. For reasons outlined below, only terms for 75 of the 1436 chemicals identified by the zebrafish searches were included in the mammalian searches. Search terms for these 75 chemicals and their synonyms were derived from MeSH, Emtree, and PubChem. These chemical terms are not part of the mammalian searches that are listed in the published protocol (Tsaoun *et al.*, 2018). The final zebrafish searches and the final mammalian searches (with chemical terms) are provided here as [Supplementary Material \(Supplementary Material 1 \[zebrafish\] and Supplementary Material 2 \[mammalian\]\)](#). No additional sources, such as references of eligible studies, were considered.

Screening. Eligibility criteria for ZET studies were identical to those reported in the preparatory study (Stephens *et al.*, 2019), with the exception of studies exposing zebrafish embryos 144 hpf, in which only the observations until 120 hpf were considered eligible. Outcome measures were assigned to 3 types: mortality, general embryotoxicity, or specific embryotoxicity (Table 1). Note that we excluded behavior-related outcomes, which are frequently addressed in ZET studies (Dach *et al.*, 2019), because functional deficits are usually not investigated in mammalian prenatal developmental toxicity studies (OECD, 2018). Rather than defining eligibility by specific outcomes, ZET studies were included if outcome measures of all 3 types were observed.

The eligibility criteria for mammalian studies have been amended from those reported previously (Stephens *et al.*, 2019). The time frame for eligible exposures, which were defined based on most frequently used exposure windows (rat: gestation days [GDs] 5–15; rabbit: GDs 6–18), was expanded to the entire gestational period, as this was imposing an unnecessary restriction. Mammalian outcomes were grouped under 4 types: growth retardation, external abnormalities, soft tissue abnormalities, and skeletal abnormalities. Prenatal mortality was not considered, as the cause can often not be determined unambiguously (OECD, 2008).

The title and abstract screening and full-text screening of the zebrafish and mammalian studies were each carried out by 2 reviewers, who resolved conflicts through discussion or, if needed, by involving a third reviewer. In addition, title and abstract screening was aided by automated machine-learning tools: zebrafish studies were excluded when 1 reviewer confirmed exclusion suggested by the automatic exclusion functionality of SWIFT-Active Screener (Sciome LLC, <https://www.sciome.com/swift-activescreener/>; last accessed on June 15, 2021), and mammalian studies were included or excluded when 1 reviewer confirmed the respective suggestion obtained by applying the automated reviewer functionality of DistillerSR’s AI toolkit (Evidence Partners Inc., <https://www.evidencepartners.com>; last accessed on June 15, 2021).

Selection of chemicals. A total of 1436 chemicals were tested in the included ZET studies, with a majority of these chemicals (1060) tested using a high-throughput system (Truong *et al.*, 2014). This large number of chemicals presented challenges for developing

Table 1. Summary of ZET Outcome Measures by Outcome Group and Type

Outcome Type	Outcome Group	Outcome Measure
Mortality	—	Heartbeat severely reduced, coagulation
General embryotoxicity	Hatching	Unhatched, partially hatched
	Cell viability	Overall degeneration, coagulation (local)
	Body shape (general)	Arrest, retardation
	Edema	Cranial, pericardium, or yolk edema
	Cardiovascular system	Heartbeat or blood flow decreased
	Yolk	Yolk sac or yolk sac extension still present
Specific embryotoxicity	Body shape (specific)	Curved, short, or kinked tail; short body
	Fins	Dorsal, ventral, pectoral, or caudal fin alterations
	Skin	Pigmentation alterations
	Cardiovascular system	Heart, aorta, vein, or vessel alterations
	CNS and sensory organs	Brain or nasal cavity impaired; eye or otic vesicle alterations
	Head	Mouth opening or jaw impaired
	Digestive system	Anterior, mid, posterior intestine, or anus alterations
	Trunk	Somites, spinal cord, or notochord impaired

the mammalian searches. As each chemical has multiple synonyms, even with the use of a URL-based API (Application Programming Interface) for searching PubChem, the search and data clean-up for generating the synonyms for 1436 chemicals would have been very labor- and time-intensive. A related challenge would have been the length of the resulting search strings. Very long search strings can present problems for databases, resulting in the need to split searches into multiple parts. This can lead to more than usual duplication in search results, which then needs to be removed at a later step. Furthermore, had these searching-based hurdles been overcome, it was likely that the resulting set of mammalian studies requiring screening would have been unmanageable, based on project resources.

In light of these challenges, we reduced the number of chemicals from 1436 to 75. Although possibly introducing bias, an informed, nonrepresentative selection of chemicals was preferred over a random selection, primarily because it would likely result in a set of chemicals better balancing mammalian prenatal developmental toxicants and nontoxicants. The 75 chemicals were chosen because they are represented in at least one of the following sources identified by the review team as relevant: 2 lists of reference substances (Brown, 2002; Daston et al., 2014), an assessment of a human embryonic stem cell-based assay for developmental toxicity screening (Palmer et al., 2013), the EPA ToxRefDB database (available at <https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data>; last accessed on June 15, 2021), and in other relevant resources (eg, Kleinstreuer et al., 2011; Malir et al., 2013; Palmer et al., 2017). The list of 75 chemicals and the resources are provided as [Supplementary Material 4](#).

Data extraction. Specific data extraction forms addressing both study characteristics and outcome data focused on outcome types were devised for ZET and mammalian studies in DistillerSR. Note that from studies exposing zebrafish embryos 144 hpf only eligible observations, that is, until 120 hpf were extracted. For ZET studies testing more than one chemical and for mammalian studies that tested a chemical on both rats and rabbits, data were extracted separately for each chemical and each species (using the clone functionality of DistillerSR). In order to address the fact that more than one set of data may be extracted from a study, we refer to datasets (rather than studies) from here onwards. Data were extracted by one reviewer, and quality control was ensured by a second reviewer by checking

all extracted data. Conflicts were resolved by the 2 reviewers through discussion.

Critical appraisal. We critically appraised the included studies regarding their reporting completeness, their risk of bias (RoB), that is, systematic errors in study design or conduct that may lead to either an overestimation or an underestimation of the true effect (Higgins et al., 2021). Because, to our knowledge, a specific tool for potential biases in toxicological studies that is based on empirical evidence is not available, we applied the RoB tool developed by the SYstematic Review Center for Laboratory animal Experimentation (SYRCLE) (Hooijmans et al., 2014). Based on the Cochrane RoB tool (Higgins et al., 2011), the SYRCLE tool has been developed for application to preclinical animal studies and addresses the classical biases related to selection, performance, detection, attrition, and reporting, to both mammalian and ZET studies with some modifications. We omitted the criterion addressing selective outcome reporting due to the multitude of potential outcomes and the “catch-all” criterion on biases not covered by the other domains in the tool. When applying the tool to ZET datasets, we replaced the criterion on randomized housing, which cannot be applied to zebrafish embryos, with a criterion on homogeneity of test conditions.

In addition, and deviating from the protocol, we included 3 criteria addressing reporting completeness and a set of “other” appraisal criteria not related to RoB, but considered important for data analysis, for example, dose-response and concentration-response plausibility, and issues with negative control data, such as high mortality. Plausibility of the dose-/concentration-response was determined by evaluating the change in response over time (ZET datasets) and over increasing concentrations (ZET and mammalian datasets) for each outcome, flagging nonmonotonous patterns. The “other” criteria relate to the concept of study sensitivity, that is, the ability to detect a true effect, described by Cooper et al. (2016).

For studies with more than one dataset, reporting and RoB criteria were assessed for the study as a whole, but the “other” criteria were applied to each dataset. All studies and datasets were appraised by one reviewer, and quality control was ensured by a second reviewer by checking all appraisals. Conflicts were resolved by the 2 reviewers through discussion.

An overview of all criteria including supportive instruction for reviewers is included in [Supplementary Material 4](#).

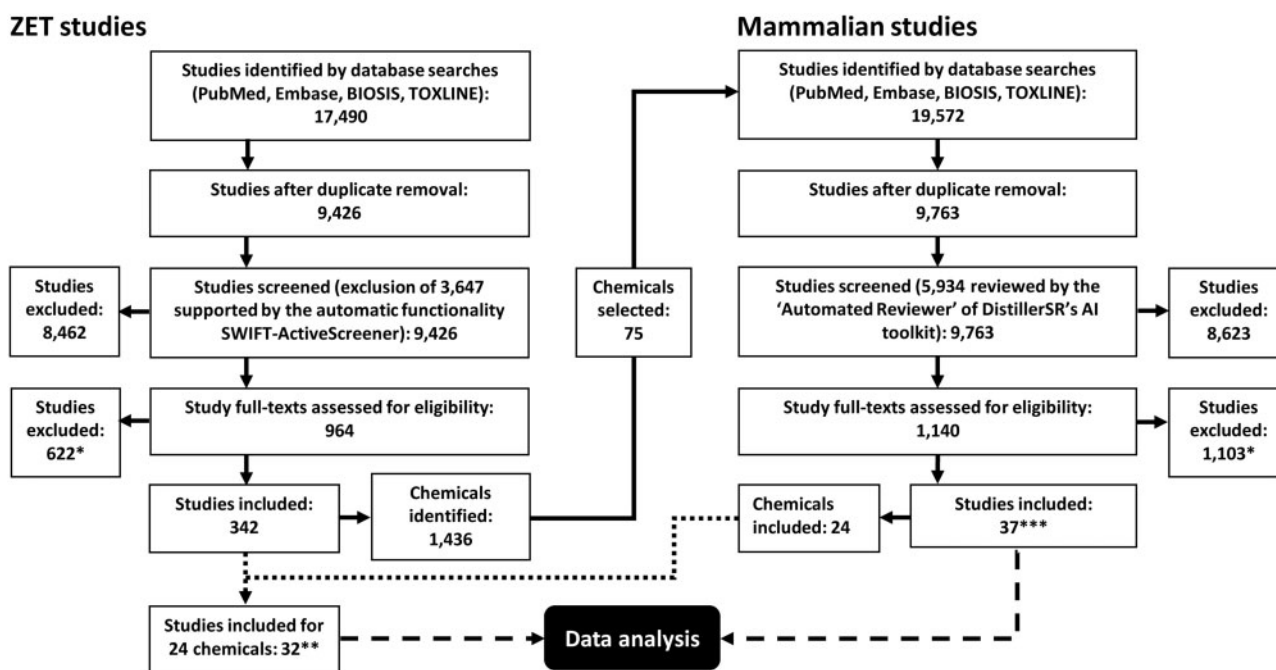


Figure 1. Modified PRISMA diagram (*see Table 2 for reasons for exclusion; **74 datasets; ***40 datasets).

Data analysis. Data analysis was conducted in a 3-step process as outlined in detail in the published protocol (Tsaïoun *et al.*, 2018).

First, we concluded for each dataset whether the results were positive (effect(s) present), negative (no effect(s) present), or inconclusive. In brief, a ZET dataset was considered positive for embryotoxicity if any outcome of general or specific embryotoxicity was observed at any concentration and any time point. ZET datasets not meeting these criteria were considered negative or, in specific cases, for example, when the maximum test concentration was considered too low (ie, did not induce mortality or was below 1000 μM), inconclusive. A mammalian dataset was considered positive if (1) an increased number of malformations or a significant increase in variations (compared with control) were observed for at least 1 outcome and (2) these malformations or variations occurred at a dose equal to or lower than the dose causing maternal toxicity. Mammalian datasets not meeting these criteria were considered negative or, in specific cases, for example, when the maximum dose was considered too low, inconclusive.

Second, we identified the chemicals with discordant results across ZET studies or across mammalian datasets (ie, negative in some ZET/mammalian studies and positive in other ZET/mammalian studies). The respective datasets were examined to identify potential experimental reasons for the differences.

Third, the results from ZET studies were compared with the results from mammalian studies across all chemicals using contingency tables.

RESULTS

Summary of Searching and Screening Steps

The ZET searches generated a total of 17 490 publications. Duplicate removal reduced these to 9426 results, from which 1654 out-of-scope references (books, book chapters, meeting

abstracts, non-English, patents, and research proposals) were excluded by sorting and searching reference type fields in EndNote. The remaining 7772 references were further reduced to 964 after title and abstract screening. Full-text screening for eligibility yielded 342 included studies. At this stage, studies were excluded primarily because no original data were reported (26.1%), the exposure was not started within 0–6 hpf (18.5%), less than 3 concentrations were used (17.0%), or no developmental toxicity outcomes were investigated (12.7%). A complete overview of reasons for exclusion is presented in Table 2. The 342 included ZET studies tested a total of 1436 chemicals (Figure 1). More than 1000 of these chemicals were tested in a single high-throughput study, most of them exclusively (Truong *et al.*, 2014). The majority of studies ($193/342 = 56\%$) investigated 1 substance, whereas 15 studies (4.4%) tested more than 10 substances.

The mammalian searches generated a total of 19 572 publications. Duplicate removal reduced these to 9763 results, from which 983 out-of-scope references (non-English and research proposals) were excluded by sorting and searching reference type fields in EndNote. The remaining 8780 references were further reduced to 1140 in the title and abstract screening. Full-text screening for eligibility yielded 37 included studies (Figure 1). During full-text screening, almost half of the studies (49.1%) were excluded because no original data were reported, especially in conference abstracts (Table 2). Exclusion also occurred for the following main reasons: exposures were not eligible (including nonoral administration routes) (13.6%), group sizes were smaller than 16 (13.5%), and less than 3 doses were tested (7.8%). Because 3 of the 37 eligible studies tested a chemical in both rats and rabbits, 40 mammalian datasets were included. Twenty-four unique chemicals were represented in these 40 datasets.

In a final step, we determined which of the 342 included ZET studies tested at least 1 of the 24 chemicals from the 37 included mammalian studies. This resulted in a final included set of 32 ZET studies with 74 datasets.

Table 2. Frequencies of Exclusion for Different Criteria During Full-Text Screening of Zebrafish and Mammalian Studies

Zebrafish Studies			Mammalian Studies		
Exclusion Criterion	No.	%	Exclusion Criterion	No.	%
Population: modified zebrafish	42	6.8	Population: modified rat or rabbit	20	1.8
Exposure: not single chemical exposure	33	5.3	Exposure: not oral route or not single chemical exposure	150	13.6
Outcomes: no developmental toxicity	79	12.7	Outcomes: no developmental toxicity	48	4.4
Language: not English	12	1.9	Language: not English	38	3.4
No original data reported	162	26.1	No original data reported	542	49.1
Less than 10 eggs per concentration	18	2.9	Less than 16 animals per group	149	13.5
Less than 3 concentrations	106	17.0	Less than 3 doses	86	7.8
Exposure not within 0–6 hpf	115	18.5	Other (eg, nonincluded chemical)	70	6.3
Time point of outcome assessment	30	4.8			
Other (eg, duplicates, full text unavailable, etc.)	25	4.0			
Total	622	100		1103	100

The entire evidence retrieval process is summarized in [Figure 1](#) as a PRISMA flow diagram by [Moher et al. \(2009\)](#) adapted to our review approach.

Characterization of the Included Studies

The 32 included ZET studies were published between 1993 and 2016. Twenty-five studies had 1 eligible dataset (ie, for 1 chemical), 5 studies had 3–6, 1 study had 9, and one study had 21 eligible datasets. Of the 24 included chemicals, 10 chemicals had 1 dataset (ie, tested in one ZET study), 7 chemicals had 2 or 3, and the remaining 7 chemicals had 5–8 datasets. The summary of the extracted data presented in [Supplementary Table 1](#) shows heterogeneity in the experimental design and the reporting of results. For example, the number of test concentrations ranged from 3 to 10, exposure ended between 48 and 144 hpf, and the way the results were presented ranged from detailed information (ie, each outcome at each timepoint) to summary measures integrating the data across timepoints and outcomes. In addition, information relevant for the data extraction, for example, the zebrafish strain and the dechoriation status, was not reported in some cases. However, the test concentration ranges of datasets for the same substance usually overlapped. Four studies did not observe or report results for all outcome types, but were considered eligible based on embryotoxicity observed in either general or specific outcomes.

The 37 included mammalian studies were published between 1978 and 2015. Three studies tested chemicals in both rats and rabbits: [Infurna et al. \(1988\)](#) (atrazine), [SDS-Biotech \(1997\)](#) (cyproconazole), and [Kennedy and Kaplan \(1984\)](#) (hexazinone). Six studies were submitted to the Office of Toxic Substances of the U.S. Environmental Protection Agency between 1990 and 1992. Fifteen of the 24 included chemicals were tested in rats only, 2 in rabbits only, and 7n in both species ([Supplementary Table 2](#)). Eight chemicals had more than one dataset. The following were tested multiple times in rats: caffeine (3×), camphor (2×) cyproconazole (2×), ethylene glycol (3×), lovastatin (2×), 2-phenylphenol (2×), and valproic acid (2×). Atrazine was tested twice in rabbits.

Most rat studies (19/31) exposed the pregnant females from GD 6 to GD 15, which is the duration recommended in the OECD TG 414. One study had a shorter exposure duration, and 10 studies had longer exposure durations. Most rabbit studies administered tested chemicals for 13 or 14 days, starting on GD 6 or GD 7. The one exception administered thalidomide for 4 days, from GD 8 to GD 11 ([Sterz et al., 1987](#)).

Results of the Critical Appraisal

Using 14 criteria, the reporting completeness and RoB of the included studies were critically appraised along with specific aspects important for data analysis. Details for all included studies (ZET and mammalian) are provided in [Supplementary Material 5](#).

Reporting in the 32 included ZET studies was very poor. Twenty-seven studies failed all 3 reporting criteria and 30 studies reported insufficient information to evaluate the RoB of 6 criteria, that is, allocation sequence, allocation concealment, blinding of investigators, random outcome assessment, blinded outcome assessment, and completeness of reported outcomes. For the baseline similarity criterion, 16 studies had low RoB, 1 had a high RoB and for 15 insufficient reporting resulted in unclear RoB. The criterion addressing homogeneity of test conditions could not be assessed for 6 studies. For the remaining 26 studies (81%) a low RoB was concluded. In summary, on average the RoB of 6.5 (of 8) criteria could not be appraised due to poor reporting. Therefore, we considered all ZET studies to be at high RoB.

Information to enable assessments of whether exposures were sufficiently high or concentration-responses were plausible was usually reported in the included ZET studies. Control data issues could not be assessed due to insufficient reporting for 43% of the datasets, the majority of which were from 4 studies ([Gustafson et al., 2012](#); [Hermsen et al., 2011](#); [Piersma et al., 2013](#); [Selderslaghs et al., 2012](#)). Control data issues were identified for 9 datasets from the only included high-throughput study ([Truong et al., 2014](#)). This same study had issues for 6 datasets regarding the highest test concentration and for 9 datasets regarding the plausibility of the concentration-response. The concentration-response was also found to be not plausible for 5 datasets from other studies. The impact of these issues on the data analysis is discussed below.

Reporting in the 37 mammalian studies was better than for the ZET studies: randomization was mentioned in 62% of the studies and blinding in 27% of the studies, but power calculation was not mentioned in any of the studies. However, reporting across all studies was such that the RoB could be assessed for only 24% of all criteria. Reporting was particularly poor regarding the criteria addressing allocation sequence, allocation concealment, random housing, blinding of investigators, and random outcome assessment. Reporting was sufficiently detailed to conclude low RoB for 21 studies for “baseline similarity” (57%), for 10 studies for “blinded outcome assessment”

and for 25 studies for “complete outcomes” (68%). A high RoB was primarily identified for “completeness of reported outcomes” (9 studies). This resulted in an average of 1.8 criteria with a low RoB per study, so that all studies were considered to be at high RoB.

The information needed to assess the other criteria informing the data analysis was usually reported in mammalian studies. There were no issues identified for 24 of the 40 datasets, 1 dataset had 3 potential issues and 15 datasets had 1 potential issue. The impact of these issues on the data analysis is discussed below.

Data Analysis

Analysis of individual datasets. The first step in the data analysis was to conclude for each dataset if the tested chemical was positive, negative, or inconclusive based on the extracted data and the procedures specified in the protocol. This step took into account issues identified by the “other” criteria, where applicable. ZET results are presented in Table 3 and mammalian results in Table 4. Both tables are sorted by chemical name and briefly summarize the experimental findings driving the results.

Of the 74 ZET datasets, 57 were positive, 8 were negative, and 9 inconclusive. All inconclusive datasets did not observe general or specific embryotoxicity, but also did not test sufficiently high doses, all being below 1000 μM . Eight of these datasets were from the only high-throughput study (Truong et al., 2014), which used a default test concentration range with 64 μM being the highest test concentration. Inconclusive datasets were excluded from further analysis, reducing the number of chemicals with at least 1 conclusive ZET dataset to 19 (see Table 5). Of these 19 chemicals, 5 had 1 conclusive dataset, 8 had 2 or 3 conclusive datasets, and 6 had 5–8 conclusive datasets.

All mammalian datasets were conclusive. Of the 25 positive datasets, 21 were conducted with rats and 4 with rabbits. Of the 15 negative datasets, 10 were conducted with rats and 5 with rabbits. Two rat datasets did not report visceral outcomes but were considered eligible based on the effects for other outcomes: Collins et al. (1987) focused in this follow-up study of Collins et al. (1983) on the most sensitive outcome and confirmed skeletal effects observed earlier, and SDS-Biotech (1997) tested rabbits in parallel, for which visceral outcomes were reported, so that we assumed that no visceral effects were observed. This protocol deviation did not introduce bias as both chemicals tested in the datasets showed skeletal effects and were therefore considered positive. Two rabbit datasets did not report growth outcomes but were considered eligible based on other outcomes and information: Sterz et al. (1987) observed all types of malformations at the lowest dose tested, and SDS-Biotech (1997) tested rabbits in parallel, for which growth outcomes were reported, so that we assumed that no growth effects were observed. This protocol deviation did not introduce potential bias for Sterz et al. (1987), whereas for SDS-Biotech (1997) the test chemical may have been positive instead of negative, which would have had only a marginal effect on the data analysis.

Evaluation of inconsistent results. Inconsistent results (in terms of negative/positive) were evaluated in detail for the respective chemicals. For the ZET datasets inconsistent results were present for rotenone, tetrabromobisphenol A, and thalidomide. Although Truong et al. (2014) observed no effects other than mortality for rotenone concentrations of 0.64 μM and higher after 120 hpf, 2 studies observed effects on pigmentations at concentrations below 0.64 μM up to the last observation time

points, that is, 80 and 96 hpf (Melo et al., 2015; Pinho et al., 2013). Similarly, although Truong et al. (2014) observed no effects other than mortality for tetrabromobisphenol A at concentrations of 6.1 and 61 μM , 7 studies observed embryotoxic effects at concentrations between 0.5 and 2 μM (see Table 4). The negative results for rotenone and tetrabromobisphenol A obtained by Truong et al. (2014) may be explained by the experimental conditions used, in particular the use of the tropical 5D zebrafish strain and the use of only one early, here not eligible and one late, here eligible assessment time point (120 hpf). Thalidomide produced the most heterogeneous results. It was positive at low concentrations in the Gao et al. (2014) study, where absent pectoral fins were observed at 2.76 μM . It was also positive in 4 datasets from an interlaboratory study (Gustafson et al., 2012), which measured embryotoxic concentrations ranging from 0.1 to 1000 μM . However, thalidomide was also found to be negative for 1 dataset in the Gustafson et al. (2012) study, in the Selderslaghs et al. (2012) study, which tested up to 150 μM due to solubility, and the Truong et al. (2014) study, which was difficult to interpret due to a high negative control mortality and an unclear concentration-related mortality. Although there was no obvious explanation for these heterogeneous results, we judged thalidomide to be positive overall. In doing so, we deviated slightly from the procedure specified in the protocol, according to which a bootstrap resampling procedure should have been applied in case inexplicable discordant results were obtained for more than 5% of the chemicals included in the comparative data analysis. As such results were observed for 1 (thalidomide) of 19 included chemicals as listed in Table 5, that is, 5.3%, this procedure would have been triggered. We considered this a minor deviation from the protocol, even though it biased the overall results toward a slightly increased concordance between the ZET and the mammalian studies.

Regarding the mammalian datasets, caffeine was the only chemical showing discordant results within species, with 2 positive rat studies and 1 negative rat study. As this difference can be explained by different methods of administration (intubation vs. drinking water) (Collins et al., 1983), caffeine was overall considered positive. Cyproconazole, ethylene glycol, and 2-phenylphenol showed discordant results between mammalian species, all being positive in the rat and negative in the rabbit (Table 5). These results may be due to species differences in maternal and prenatal-developmental toxicity or due to experimental differences, for example, in the determination of the dosing regimen or the choice of vehicle (Theunissen et al., 2016).

Chemicals with consistent datasets results were not analyzed further in this regard, because the type of outcome is of less relevance for our hazard-focused review question.

Concordance of ZET and mammalian results. Deriving overall dichotomized results for all chemicals and species allowed us to conduct the planned concordance analysis, which is presented in Table 6. The total number of chemicals that could be compared was low. Because only 8 chemicals were available for a comparison of ZET studies with prenatal developmental toxicity studies in rabbits (Table 6b), these results were not considered further. Seventeen chemicals, that is, 24% of the 75 chemicals initially selected, qualified for a comparison of ZET studies with prenatal developmental toxicity studies in rats (Table 6a). The ZET studies tended to overpredict rat negative results as positive (5 out of 6 chemicals). In addition, 2 out of 3 chemicals that were negative in the ZET (ethylene glycol and fluazinam) were positive in the rat. Consequently, concordant results were obtained for 10 of the

Table 3. ZET Outcomes and Results

Chemical Name (Reference)	Lowest Test Conc. (in μM) With Observed Outcomes (earliest timepoint, hpf)		Result	Justification
	Mortality	General Embryotoxicity		
Acetaminophen (Selderslaghs et al., 2012)	10 (20)	NO	+	Skeletal deformities and lack of pigmentation at 6230 μM
Acetaminophen (David and Pancharatna, 2009)	33.1 (n.r.)	33.1 (n.r.)	+	Reduced hatching, body mass, and body length at 33.1 μM ; tail deformities and lack of pigmentation at 66.2 μM
Acetaminophen (Truong et al., 2014)	NO	NO	Inc.	No effects observed up to the maximum test concentration of 64 μM
All-trans-retinoic acid (Truong et al., 2014)	6.4 (120)	0.064 (120)	+	Caudal fin effects and reduced trunk length at 0.0064 μM
All-trans-retinoic acid (Selderslaghs et al., 2009)	0.0266 (48)	0.0266 (48)	+	Kinked tail at 0.000213 μM and other effects at higher concentrations
All-trans-retinoic acid (Wang et al., 2014)	0.01 (120)	0.004 (120)	+	Several general and specific effects at 0.004 and 0.008 μM
All-trans-retinoic acid (Vandersea et al., 1998)	n.r.	n.r.	+	Pectoral fin effects at 0.5 μM
All-trans-retinoic acid (Teixido et al., 2013)	YES*	YES*	+	Short tails at 0.005 μM
All-trans-retinoic acid (Selderslaghs et al., 2012)	1.47 (72)	YES*	+	A wide spectrum of general and specific embryotoxicity outcomes at 0.00491 μM
All-trans-retinoic acid (Piersma et al., 2013)	YES*	YES*	+	A teratogenicity index (BMC 20) that included general and specific embryotoxic effects was 0.063 μM
Atrazine (Wiegand et al., 2001)	46.25 (24)	92.5 (24)	+	Effects on pigmentation and otoliths at 46.25 μM
Atrazine (Weber et al., 2013)	NO	NO	+	Increased head length at 0.0014 μM
Atrazine (Ton et al., 2006)	500 (48)	200 (96)	+	Heart and trunk edema, and underdeveloped jaw at 200 μM
Atrazine (Perez et al., 2013)	NO	NO	Inc.	No effects observed up to the maximum test concentration of 185.5 μM
Atrazine (Truong et al., 2014)	NO	NO	Inc.	No effects observed up to the maximum test concentration of 64 μM
Butylparaben (Truong et al., 2014)	64 (120)	64 (120)	+	Jaw effects at 0.64 and 64 μM (but not at 6.4 μM)
Caffeine (Zhu et al., 2015)	5150 (54)	YES*	+	A summary index of general and specific embryotoxicity indicated embryotoxicity at sublethal concentrations
Caffeine (Chakraborty et al., 2011)	NO	YES*	+	Decreased hatching rate at 500 μM , but likely also lower concentrations (reporting is unclear)
Caffeine (Selderslaghs et al., 2009)	3220 (24)	1610 (48)	+	Kinked tail at 390 μM and other effects at higher concentrations
Caffeine (Teixido et al., 2013)	YES*	YES*	+	Short tails at 500 μM
Caffeine (Selderslaghs et al., 2012)	3480 (72)	YES*	+	Short tails at 390 μM
Caffeine (Truong et al., 2014)	0.064 (120)	6.4 (120)	+	Skeletal deformities observed at 390 μM

Table 3. (continued)

Chemical Name (Reference)	Lowest Test Conc. (in μM) With Observed Outcomes (earliest timepoint, hpf)			Result	Justification
	Mortality	General Embryotoxicity	Specific Embryotoxicity		
Camphor (Yim et al., 2014)	395 (48)	395 (72)	395 (96)	+	Mortality at 0.064 and 64 μM ; yolk sac and pericardial edema, as well as pectoral fin effects, at 6.4 μM Bent spine, yolk sac edema, and decreased hatching rate at 395 μM
Camphor (Selderslaghs et al., 2012)	4000 (72)	NO	1230 (72)	+	Otolith abnormalities observed at 1230 μM
Clopyralid (Truong et al., 2014)	0.64 (120)	NO	NO	Inc.	No embryotoxic effects up to the maximum test concentration of 64 μM
Cyproconazole (Truong et al., 2014)	NO	64 (120)	64 (120)	+	A wide spectrum of embryotoxic effects at 64 μM
Cyproconazole (Hermesen et al., 2011)	YES (n.r.)	31.6 (72)	31.6 (72)	+	Score used, but teratogenic effects observed at 31.6 μM
Dimethyl phthalate (Truong et al., 2014)	NO	NO	NO	Inc.	No effects observed up to the maximum test concentration of 64 μM
Ethylene glycol (Truong et al., 2014)	0.064 (120)	NO	NO	-	No embryotoxic effects observed up to the maximum test concentration of 64 μM , but mortality
Fluazinam (Truong et al., 2014)	0.64 (120)	NO	NO	-	No embryotoxic effects up to the maximum test concentration of 64 μM , but mortality
Genistein (Ren et al., 2012)	20 (120)	20 (120)	20 (120)	+	Increased malformation rate, for example, ocular effects, and high mortality observed at 20 μM
Genistein (Truong et al., 2014)	0.64 (120)	64 (120)	64 (120)	+	Yolk sac and pericardial edema, jaw effects, and high mortality observed at 64 μM
Hexazinone (Truong et al., 2014)	6.4 (120)	NO	NO	-	No embryotoxic effects observed up to the maximum test concentration of 64 μM , but mortality
Lovastatin (Gustafson et al., 2012) ^y	0.1 (120)	0.1 (120)	0.1 (120)	+	NOAEL for malformations lower than LC25 (mortality)
Lovastatin (Gustafson et al., 2012) ^y	0.1 (120)	0.1 (120)	0.1 (120)	+	NOAEL for malformations lower than LC25 (mortality)
Lovastatin (Gustafson et al., 2012) ^y	0.1 (120)	0.1 (120)	0.1 (120)	+	NOAEL for malformations lower than LC25 (mortality)
Lovastatin (Gustafson et al., 2012) ^y	0.03 (120)	0.001 (120)	0.001 (120)	+	NOAEL for malformations lower than LC25 (mortality)
Lovastatin (Truong et al., 2014)	64 (120)	0.64 (120)	0.64 (120)	+	Wide spectrum of general and specific embryotoxicity outcomes at 0.64 μM
Methoxyacetic acid (Teixido et al., 2013)	YES*	YES*	2000 (52)	+	Short tails at 2000 μM .
Methoxyacetic acid (Hermesen et al., 2011)	YES (n.r.)	3160 (72)	1000 (72)	+	Score used, but teratogenic effects observed at 1000 μM
Methoxyacetic acid (Hermesen et al., 2011)	NO	YES*	YES*	+	

Table 3. (continued)

Chemical Name (Reference)	Lowest Test Conc. (in μM) With Observed Outcomes (earliest timepoint, hpf)			Result	Justification
	Mortality	General Embryotoxicity	Specific Embryotoxicity		
<i>n</i> -Methylpyrrolidone (Zhang et al., 2013)	8800 (72)	2640 (72)	2640 (72)	+	A teratogenicity index (BMC 20) that included general and specific embryotoxic effects was 6310 μM
2-Phenylphenol (Truong et al., 2014)	64 (120)	NO	NO	Inc.	Small heads and eyes, retarded growth, and pericardial edema at 2640 μM No embryotoxic effects observed up to the maximum test concentration of 64 μM , but mortality at 64 μM
Rotenone (Pinho et al., 2013)	0.316 (32)	0.316 (56)	0.0316 (32)	+	Abnormal embryos at 0.316 μM
Rotenone (Melo et al., 2015)	0.025 (48)	0.025 (72)	0.018 (48)	+	Lack of pigmentation at 0.018 μM
Rotenone (Truong et al., 2014)	0.64 (120)	NO	NO	-	No embryotoxic effects observed up to the maximum test concentration of 64 μM , but mortality at 0.64 μM
Tetrabromobisphenol A (Yang et al., 2015)	0.92 (96)	0.92 (48)	0.92 (96)	+	Malformations at 0.92 μM
Tetrabromobisphenol A (Song et al., 2014)	2.76 (48)	0.92 (72)	NO	+	Reduced hatching rate at 0.92 μM
Tetrabromobisphenol A (Noyes et al., 2015)	6.4 (120)	6.4 (120)	0.64 (120)	+	Jaw malformations at 0.64 μM
Tetrabromobisphenol A (McCormick et al., 2010)	3 (48)	0.75 (48)	0.75 (48)	+	Edema, hemorrhage, and tail malformations at 0.75 μM
Tetrabromobisphenol A (Hu et al., 2009)	2.76 (60)	1.38 (36)	1.38 (36)	+	Malformations at 1.38 μM
Tetrabromobisphenol A (Carlsson and Norrgren, 2014)	1.84 (24)	1.84 (48)	n.r.	+	Edema and mortality at 1.84 μM
Tetrabromobisphenol A (Baumann et al., 2016)	n.r.	n.r.	0.5 (120)	+	Decrease in size and pigmentation of retinal cells at 0.5 μM
Tetrabromobisphenol A (Truong et al., 2014)	6.1 (120)	NO	NO	-	No embryotoxic effects observed up to the maximum test concentration of 64 μM , but mortality at 6.1 μM
Thalidomide (Gao et al., 2014)	27 (96)	13.8 (n.r.)	2.76 (n.r.)	+	Pectoral fins missing at 2.76 μM
Thalidomide (Selderslaghs et al., 2012)	NO	NO	NO	-	No effects observed up to the maximum soluble concentration of 150 μM
Thalidomide (Truong et al., 2014)	6.4 (120)	NO	NO	-	No embryotoxic effects observed up to the maximum test concentration of 64 μM , but mortality
Thalidomide (Gustafson et al., 2012) ^o	10 (120)	0.01 (120)	0.01 (120)	+	NOAEL for malformations lower than LC25 (mortality)
Thalidomide (Gustafson et al., 2012) ^o	NO	NO	NO	-	No effects observed up to the maximum test concentration of 1000 μM
Thalidomide (Gustafson et al., 2012) ^o	1000 (120)	1 (120)	1 (120)	+	NOAEL for malformations lower than LC25 (mortality)
Thalidomide (Gustafson et al., 2012) ^o	1000 (120)	100 (120)	100 (120)	+	NOAEL for malformations lower than LC25 (mortality)
Thalidomide (Gustafson et al., 2012) ^o	100 (120)	1 (120)	1 (120)	+	NOAEL for malformations lower than LC25 (mortality)
Triadimefon (Truong et al., 2014)	NO	64 (120)	64 (120)	+	Several general and specific effects at 64 μM

Table 3. (continued)

Chemical Name (Reference)	Lowest Test Conc. (in μM) With Observed Outcomes (earliest timepoint, hpf)			Result	Justification
	Mortality	General Embryotoxicity	Specific Embryotoxicity		
Triadimefon (Hermsen et al., 2011)	YES*	31.6 (72)	10 (72)	+	Score used, but embryotoxic effects at the sublethal concentration of 10 μM
Triclopyr (Truong et al., 2014)	NO	NO	NO	Inc.	No effects observed up to the maximum test concentration of 64 μM
Triethylene glycol (Truong et al., 2014)	NO	NO	NO	Inc.	No effects observed up to the maximum test concentration of 64 μM
Valproic acid (Herrmann, 1993)	3000 (24)	100 (20)	30 (24)	+	Short or bent tail at 30 μM
Valproic acid (Beker van Woudenberg et al., 2014)	730 (n.r.)	150 (72)	150 (n.r.)	+	Brain and eye effects, as well as pericardial edema, at or approximately at 150 μM
Valproic acid (Selderslaghs et al., 2009)	1500 (24)	750 (48)	1500 (48)	+	Blood circulation effects observed at 750 μM
Valproic acid (Teixido et al., 2013)	YES*	50 (n.r.)	300 (n.r.)	+	Increased head-trunk angle at 50 μM (sublethal) and short tails at 300 μM (sublethal)
Valproic acid (Selderslaghs et al., 2012)	1570 (72)	YES*	550 (72)	+	A wide spectrum of general and specific embryotoxicity outcomes observed
Valproic acid (Truong et al., 2014)	NO	NO	NO	Inc.	No effects observed up to the maximum test concentration of 64 μM
Valproic acid # (Piersma et al., 2013)	NO	YES*	YES*	+	A teratogenicity index (BMC 20) that included general and specific embryotoxic effects was 1585 μM
Valproic acid # (Lee et al., 2013)	6.25 (72)	NO	6.25 (72)	+	Growth retardation at the sublethal concentration of 6.25 μM

hpf, hours post fertilization; n.r., not reported; *, data reported differently, not allowing to determine the lowest test concentration for respective outcomes; +, positive; -, negative; Inc., inconclusive; #, sodium salt; NOAEL, no observed adverse effect level; LC25, lethal concentration inducing 25% mortality; ", same order as in Supplementary Table 1.

Table 4. Mammalian Outcomes and Results

Chemical Name (Reference)	Species Retarded Growth * Variations or Malformations* (at lowest dose) Maternal Toxicity* Result				Justification
	External	Visceral	Skeletal		
Acetaminophen (Burdan et al., 2001)	Rat	350	—	—	+ Retarded growth at nonmaternally toxic dose (350 mg/kg bw)
All- trans-retinoic acid (Seegmiller et al., 1997)	Rat	—	10	5	+ Significant increase in incomplete ossification and superumerary ribs at nonmaternally toxic dose (5 mg/kg bw) and in cleft palate at nonmaternally toxic dose (10 mg/kg bw)
Atrazine (Infurna et al., 1988)	Rat	700	—	70	+ Significant increase in incomplete ossification at lowest maternally toxic dose (70 mg/kg bw)
Atrazine (Infurna et al., 1988)	Rabbit	75	—	75	+ Significant increase in nonossification at lowest maternally toxic maternal dose (75 mg/kg bw)
Butylparaben (Daston, 2004)	Rat	—	—	1000	- No developmental effects observed at maternally toxic dose (1000 mg/kg bw)
Caffeine (Collins et al., 1981)	Rat	40	80	40	- No developmental effects observed at maternally toxic dose (12 mg/kg bw)
Caffeine (Collins et al., 1983)	Rat	100.8	144	50.8	+ Significant increase in skeletal variation at nonmaternally toxic dose (50.8 mg/kg bw)
Caffeine (Collins et al., 1987)	Rat	—	n.a.	48.8	+ Significant increase in sternebral variation at lowest maternally toxic dose (48.8 mg/kg bw)
Camphor (Navarro et al., 1992a)	Rat	—	—	400	- No developmental effects observed at maternally toxic dose (400 mg/kg bw)
Camphor (Navarro et al., 1992b)	Rabbit	—	—	400	- No developmental effects observed at maternally toxic dose (400 mg/kg bw)
Camphor (Leuschner, 1997)	Rat	—	—	1000	- No developmental effects observed at maternally toxic dose (1000 mg/kg bw)
Clopyralid (Hayes et al., 1984)	Rat	—	—	250	- No developmental effects observed at maternally toxic dose (250 mg/kg bw)
Cyproconazole (Machera, 1995)	Rat	20	20	50	+ Retarded growth and visceral malformations at nonmaternally toxic dose (20 mg/kg bw)
Cyproconazole (SDS-Biotech, 1997)	Rat	24	n.a.	24	+ Retarded growth and skeletal malformations at lowest maternally toxic dose (24 mg/kg bw)
Cyproconazole (SDS-Biotech, 1997)	Rabbit	n.a.	—	50	- No developmental effects observed at maternally toxic dose (50 mg/kg bw)
Dimethyl phthalate (Field et al., 1993)	Rat	—	—	200	- No developmental effects observed at maternally toxic dose (200 mg/kg bw)
Ethylene glycol (Maronpot et al., 1983)	Rat	—	—	1000	+ Significant increase in poorly ossified and nonossified vertebral centra at nonmaternally toxic dose (1000 mg/kg bw)
Ethylene glycol (Neeper-Bradley, 1990)	Rat	1000	2500	1000	+ Retarded growth and skeletal malformations at nonmaternally toxic dose (1000 mg/kg bw)
Ethylene glycol (Price et al., 1992)	Rat	2500	5000	2500	+ Increase in (litters with) visceral malformations at lowest maternally toxic dose (1250 mg/kg bw)

Table 4. (continued)

Chemical Name (Reference)	Species Retarded Growth * Variations or Malformations* (at lowest dose) Maternal Toxicity* Result				Justification		
	External	Visceral	Skeletal				
Ethylene glycol (Tyi et al., 1991)	Rabbit	—	—	—	2000	—	No developmental effects observed at maternally toxic dose (2000 mg/kg bw)
Fluazinam (Tesh et al., 1992)	Rat	50	250	250	50	+	Increase in skeletal malformations and retarded growth at lowest maternally toxic dose (50 mg/kg bw)
Genistein (McClain et al., 2007)	Rat	—	—	—	500	—	No developmental effects observed at maternally toxic dose (500 mg/kg bw)
Hexazinone (Kennedy and Kaplan, 1984)	Rat	—	—	—	22.5	—	No developmental effects observed at maternally toxic dose (22.5 mg/kg bw)
Hexazinone (Kennedy and Kaplan, 1984)	Rabbit	—	—	—	125	—	No developmental effects observed at maternally toxic dose (125 mg/kg bw)
Lovastatin (Lankas et al., 2004)	Rat	100	800	400	200	+	Retarded growth at nonmaternally toxic dose (100 mg/kg bw)
Lovastatin (Minsker et al., 1983)	Rat	800	800	800	800	+	Retarded growth, external malformations, and skeletal malformations at lowest maternally toxic dose (800 mg/kg bw)
Methoxyacetic acid (Carney et al., 2003)	Rabbit	7.5	7.5	7.5	15	+	Retarded growth, various malformations, and increased variations at nonmaternally toxic dose (7.5 mg/kg bw)
n-Methylpyrrolidone (Saillenfait et al., 2002)	Rat	250	500	500	500	+	Retarded growth at nonmaternally toxic dose (250 mg/kg bw)
2-Phenylphenol (Anonymous, 1991)	Rat	—	—	700	700	+	Significantly delayed sternbrae ossification and skull foramen at the lowest maternally toxic dose (700 mg/kg bw)
2-Phenylphenol (Anonymous, 1992)	Rabbit	—	—	—	250	—	No developmental effects observed at maternally toxic dose (250 mg/kg bw)
2-Phenylphenol (Kaneda et al., 1978)	Rat	600	300	—	300	+	Increase in visceral malformations at lowest maternally toxic dose (300 mg/kg bw)
Rotenone (Khera et al., 1982)	Rat	—	—	5	5	+	Significant increase in extra ribs, delayed ossification, and missing sternbrae at lowest maternally toxic dose (5 mg/kg bw)
Tetrabromobisphenol A (Cope et al., 2015)	Rat	—	—	—	—	—	No developmental effects observed at highest and nonmaternally toxic dose (1000 mg/kg bw)
Thalidomide (Sterz et al., 1987)	Rabbit	n.a.	50	50	—	+	External, visceral, and skeletal malformations at nonmaternally toxic dose (50 mg/kg bw)
Triadimefon (Machener, 1992)	Rat	—	—	—	30	—	No developmental effects observed at maternally toxic dose (30 mg/kg bw)
Triclopyr (Hanley et al., 1984)	Rat	—	200	200	200	+	External and skeletal malformations (two fetuses) at lowest maternally toxic dose (200 mg/kg bw)
Triethylene glycol (Ballantyne and Snellings, 2005)	Rat	11 260	—	11 260	11 260	+	

Table 4. (continued)

Chemical Name (Reference)	Species Retarded Growth * Variations or Malformations* (at lowest dose) Maternal Toxicity* Result				Justification	
	External	Visceral	Skeletal			
Valproic acid (Ong et al., 1983)	Rat	600	150	600	+	Significant increase in thoracic centrum skeletal variation and retarded growth at lowest maternally toxic dose (11 200 mg/kg bw) Significant increase in rib variation at nonmaternally toxic dose (150 mg/kg bw) Skeletal malformations at nonmaternally toxic dose (350 mg/kg bw) External and visceral malformations at nonmaternally toxic dose (200 mg/kg bw)
Valproic acid (Pettrere et al., 1986)	Rabbit	—	350	—	+	
Valproic acid (Vorhees, 1987)	Rat	300	200	400	+	

* , unit is mg/kg bw/day; n.a., not available (not observed or not reported); +, positive; —, negative.

17 chemicals (56%). When combining rat and rabbit studies in a conservative way, that is, both have to be negative for an overall negative result, while at least one has to be positive for an overall positive result, 15 chemicals (20%) qualified for the concordance analysis (Table 6c). Of the 13 chemicals that were positive in at least 1 mammalian species, 11 were also positive in the ZET. In addition, one chemical (hexazinone) was negative in all species. In summary, concordant results were obtained for 12 of the 15 chemicals (80%). Statistical significance was not calculated due to the small sample size of included chemicals.

Confidence in results. The two factors impacting on confidence of the entire evidence base, that is, across all chemicals, systematically analyzed were the RoB and the plausibility of concentration-/dose-response. Due to poor reporting, the evidence has high RoB, reducing our general confidence in the evidence used for the determination of concordance. The concentration-/dose-responses, as assessed under the “other” critical appraisal criteria, were considered plausible, with exception of the jaw effects observed by [Truong et al. \(2014\)](#) for butylparaben, which lacked a concentration-response, with effects at 0.64 μM and the lethal concentration of 64 μM, but not at 6.4 μM. This general plausibility increased the confidence in the overall evidence base.

However, on a chemical level, other factors impacting on the confidence were explored. For example, clearly increased severe developmental effects in the absence of general toxicity increased the confidence, for example, as observed for most all-trans-retinoic acid ZET datasets, and in [Seegmiller et al. \(1997\)](#) and [Machera \(1995\)](#).

In other cases, issues identified in the critical appraisal, especially by the “other” criteria that were specifically designed to highlight factors impacting the data analysis reduced the confidence. Among the ZET datasets, we identified, in addition to the above-mentioned butylparaben dataset, 2 datasets with positive results that had such issues. First, [Truong et al. \(2014\)](#) observed three significant outcomes for genistein at 64 μM, which also induced a very high mortality. Second, the decreased hatching rates observed by [Chakraborty et al. \(2011\)](#) with increasing caffeine concentrations could have been due to the difference in embryo ages at baseline. Among the mammalian studies, one dataset had unclear reporting, which reduced the confidence in its negative result. [SDS-Biotech \(1997\)](#) reported no growth outcomes for cyproconazole in the rabbit. Based on other details of the study, we decided that the lack of reporting was due to an absence of effects, although this was not explicitly reported.

Also the above analysis of inconsistent results informed the confidence assessment on the level of the individual chemicals. The frequency of inconsistent results was relatively low (3 out of 14 chemicals for the ZET, 1 of 7 chemicals for rat studies), and, except for the ZET results for thalidomide, potential reasons for the inconsistency of results were identified. Therefore, we considered the overall evidence base as consistent and not as a confidence-reducing factor.

Although we have not planned to integrate those 4 factors, we are confident in that the evidence base allows to draw moderately sound conclusions.

Furthermore, due to the relatively low incidence of chemicals with confidence-reducing issues and due to small sample size, we refrained from a chemical-specific data analysis approach accounting for confidence and weighted all chemicals equally in the concordance analysis.

Table 5. Summary of All Results by Chemical

Chemical Name	ZET Studies		Mammalian Studies	
	Individual Datasets	Overall	Rat	Rabbit
Acetaminophen	+/+ ¹	+	+	nd
All-trans-retinoic acid	+/+/+/+/+/+	+	+	nd
Atrazine	+/+ ²	+	+	+
Butylparaben	+	+	-	nd
Caffeine	+/+/+/+/+	+	-/+/*	nd
Camphor	+/+	+	-/-	-
Cyproconazole	+/+	+	+/+	-
Ethylene glycol	-	-	+/+	-
Fluazinam	-	-	+	nd
Genistein	+/+	+	-	nd
Hexazinone	-	-	-	-
Lovastatin	+/+/+/+	+	+/+	nd
Methoxyacetic acid	+/+	+	nd	+
n-Methylpyrrolidone	+	+	+	nd
Rotenone	+/-	+	+	nd
Tetrabromobisphenol A	+/+/+/+/+/-	+	-	nd
Thalidomide	+/+/+/+/-/-	+	nd	+
Triadimefon	+/+	+	-	nd
Valproic acid	+/+/+/+/+ ¹	+	+/+	+
Clopyralid	1	Inconclusive	-	nd
Dimethyl phthalate	1	Inconclusive	-	nd
2-Phenylphenol	1	Inconclusive	+/+	-
Triclopyr	1	Inconclusive	+	nd
Triethylene glycol	1	Inconclusive	+	nd

+, positive; -, negative; *, considered positive overall; nd, no data; superscript numbers (1 and 2) indicate amount of inconclusive ZET studies.

Table 6. 2 × 2 Contingency Tables Comparing the ZET Results With (a) the Rat Results, (b) the Rabbit Results, and c) the Combined Rat and Rabbit Results (“and”: Negative Results for Both Species; “or”: a Positive Result for at Least One Species)

(a)	Rat		Σ	(b)	Rabbit		Σ	(c)	Rat and/or rabbit		Σ
	-	+			-	+			-(and)	+(or)	
ZET	-	1	2	3	2	0	2	13	1	2	3
	+	5	9	14	2	4	6	13	1	11	12
Σ		6	11	17	4	4	8	2	15		

DISCUSSION

The capacity of the ZET and the mammalian prenatal developmental toxicity test to predict prenatal developmental toxicity hazard of chemicals were systematically reviewed. The potential of the ZET to provide relevant evidence for the assessment of the prenatal developmental toxicity of chemicals has been explored extensively in primary studies. This popularity is evident from our literature search targeted to result in a homogeneous subset of ZET studies, in which we identified 1436 chemicals tested in 342 ZET studies. Informed by an initial scoping exercise (Stephens *et al.*, 2019), we decided to focus on 75 chemicals to stay within feasible dimensions of our review. The search of the mammalian literature identified 37 eligible prenatal developmental toxicity studies for 24 of the 75 chemicals. After we derived conclusions as either positive or negative for each dataset and summarized conclusions for chemicals with more than one dataset, a total of 19 chemicals were available to compare the ZET with the prenatal developmental mammalian test using 2 × 2 contingency tables.

Although the confidence in the evidence was moderate, the confidence in the results of the test method comparison was weakened by the small number of chemicals and also by a higher number of positive results on both sides. However, our review results suggest that the ZET has some potential to identify chemicals that are prenatal developmental toxicants in rats and/or rabbits. Furthermore, our analysis indicated that the ZET is overpredicting chemicals as positive that are negative in the individual mammalian species, and confirmed the need for further standardization of the ZET. To elucidate why the confidence in the test method comparison results remained weak, we discuss potential reasons that limited the evidence and reconsider decisions made when defining the systematic review protocol.

Selection Challenges

The systematic review was designed in such a way that the confidence in its conclusions would be driven to a major extent by the number of chemicals included. By selecting these substances in a nonrandom manner possibly introducing a bias (of

unknown direction), we expected that selecting substances well-studied for prenatal developmental toxicity would result in a high number of chemicals for the concordance analysis. This assumption did not hold true, as we found eligible studies for only 24 of the 75 chemicals.

One factor contributing to the low chemical coverage could have been the stringency of our eligibility criteria, which may have excluded studies relating to any of the other 51 chemicals. However, more relaxed eligibility criteria could have led to other complications. For instance, the criteria addressing group size and number of doses could potentially have been less stringent for mammalian studies, but only for positive chemicals. For negative chemicals, a group size of at least 16 and 3 doses seems to be conventionally required to have sufficient confidence in a negative result. Such a results-based approach would have substantially increased the risk of selection bias because the eligibility of studies could then only have been determined after data analysis. There would also have been complications if the route of exposure criterion had been less stringent. The inclusion of mammalian studies with nonoral administration routes would have further increased the complexity and decreased the interpretability of the data due to route-specific absorption and metabolism.

Another factor contributing to the low number of included chemicals could have been the exclusion of regulatory databases from our set of information sources. However, although regulatory databases are likely to report findings in mammals based on OECD TG 414 and similar tests, these databases may not be publicly available, may not report original data and may not offer comprehensive search options.

Consequently, selection of more than 75 of the 1436 chemicals would have been the most promising way to increase the number of chemicals for the test method comparison. However, a selection process of such dimensions would have required more efficient approaches, for example, aided by artificial intelligence tools that are still being developed and optimized for mining existing evidence for selection purposes.

An increase in the number of included chemicals would also be the only viable approach to obtain a substantial number of chemicals that are negative in the ZET and the mammalian test. The extent of ZET development and standardization is likely an important factor contributing to a high proportion of positive results. Once the general experimental setup of a test method like the ZET has been defined, researchers usually start exploring its application by making sure that reference chemicals with well-known and clear effects are identified. This likely explains, for example, why several ZET datasets for the well-known prenatal developmental toxicants all-trans retinoic acid, thalidomide, and valproic acid were included. In a next step, the interpretation of experimental data is standardized based on the results obtained. With a strong focus on the correct identification of harmful substances, that is, a test methods' sensitivity, exposure conditions and interpretation procedures are often tuned to be sensitive. For example, the effects of embryo dechorionization on ZET outcomes and conclusions have been discussed by Hamm et al. (2019). The risk of such tuning is that a test method will become overly sensitive, indicating harmful effects for most substances tested. This will inevitably lead to a reduced ability to correctly identify nonharmful substances. Our focus on well-known prenatal developmental toxicants and our requirement for a 1000 μ M test concentration for negative conclusions for soluble chemicals likely resulted in the observation that the ZET was positive for 16 of the 19 chemicals with conclusive ZET data. Although we anticipated this lack of

balance and attempted to account for it in the selection process for the 75 chemicals, we did not succeed in avoiding the imbalance, and this reduced the comprehensiveness of our test method comparison. This is an important lesson for researchers planning future systematic reviews comparing toxicological test methods, particularly if the 2 test methods substantially differ in their levels of development and standardization.

Data Extraction and Analysis Challenges

Standardization issues also impacted the data extraction step of our review. Studies of mammalian prenatal developmental toxicity have well-established guidelines for which outcomes should be measured and how outcomes should be measured and assessed, both individually and in combination, particularly fetal and maternal effects induced by the same dose (Chahoud et al., 1999; Danielsson, 2013). In contrast, ZET studies differ substantially in outcomes observed and in how effects are summarized and interpreted (Beekhuijzen et al., 2015). This is reflected, for example, in our data extraction for cases in which we could determine that effects were observed, but not at which concentration and at which timepoint (see Supplementary Table 1). This lack of ZET standardization led to discrepancies between the results of studies, for example, when different outcomes are observed, different concentrations are tested, and different outcome assessment timepoints are used.

Data analysis challenges relate to the discrimination of positive and negative results. This process leads to cases that are clearly positive or negative, but also to borderline cases, which are usually associated with a higher level of uncertainty (Gabbert et al., 2020). Indeed, our conservative interpretation of ZET data led to positive results of such borderline cases. A good example is the positive result determined for the only ZET study with the highly water-soluble chemical *n*-methylpyrrolidone (Zhang et al., 2013), which clearly induced embryotoxic effects at nonlethal, but very high concentrations, that is, $\geq 2640 \mu$ M. Had *n*-methylpyrrolidone been tested only up to 1000 μ M, Zhang et al. (2013) data suggest that no effects would have been observed, which would have led to a negative result according to our data analysis criteria. A similar example is the positive result for the only mammalian study testing triclopyr, which showed a low incidence of malformation at the maternally toxic dose of 200 mg/kg bw (Hanley et al., 1984). Although such malformations were not observed in the other dose groups and the control, a historical database of negative control data may have shown a similarly low background incidence of such malformations, which may have resulted in a negative result.

Detailed Discussion of Two Example Chemicals

Accounting for additional relevant evidence, we evaluated in further detail 2 chemicals, camphor and fluzinam, to better understand the results obtained and potentially decrease uncertainty associated with them.

Camphor was the only chemical without any prenatal developmental effects in both mammalian species (Navarro et al., 1992a,b; Leuschner, 1997). These results are strengthened by another negative rabbit study included in Leuschner (1997) that was considered not eligible in our review because of group sizes smaller than 16. Based on the same studies, the European Food Safety Agency (EFSA) also concluded that camphor is not a prenatal developmental toxicant in mammals (EFSA, 2008). In addition, camphor is easily absorbed in the gastrointestinal tract and is metabolized initially by oxidation, which is possibly species specific. Some human evidence exists that suggests that camphor does not induce prenatal developmental toxicity in

humans (Heinonen *et al.*, 1977). In contrast, our review of the 2 included ZET camphor datasets concluded a positive result for both.

Yim *et al.* (2014) observed coagulation and general embryotoxic effects (yolk sac edema, pericardial edema, and delayed hatching) in a concentration- and time-dependent manner. At 790 μM , both edema types were found in approximately 25% of the embryos and coagulation was found in 20% of the embryos. Two specific embryotoxic effects were also observed. Bent spine was primarily induced by the lowest test concentration of 395 μM and to a minor extent at higher concentrations. Ocular defects were observed at 790 and 1580 μM , the latter concentration leading to 60%–70% coagulation. The interpretation of the data was impaired by the fact that negative control data were reported for coagulation and hatching only. The second ZET camphor study, Selderslaghs *et al.* (2012), observed only 1 embryotoxic effect. At 72 hpf, abnormal otoliths were found in 50% of the embryos treated with 1230 μM , a concentration 3–4-fold lower than the concentration that induced 50% lethality. As effects were reported in terms of LC50 and EC50 only, it cannot be determined at which concentration abnormal otoliths started to occur and if other embryotoxic effects were present in less than 50% of the embryos.

The results of the 2 ZET studies, which tested similar concentrations, but different camphor forms using different vehicles, are difficult to compare, mainly because outcome results were reported differently. However, even though the concentrations inducing about 50% lethality differed in the 2 studies by approximately a factor of 2, no contradictory results were obtained. When the ZET and mammalian results are compared, it is not clear why they are discordant. Assuming that camphor was bioavailable, species-specific metabolism may have caused differences in internal exposure and thus in results, which is supported by a review that identified different metabolites formed by mammalian species (EFSA, 2008). In addition, given that the general biotransformation capacity of zebrafish embryos is still a matter of debate (de Souza Anselmo *et al.*, 2018; Saad *et al.*, 2017), the zebrafish embryo, in contrast to mammals, may not be able to metabolize camphor at all.

The second chemical, fluazinam, was positive in a rat prenatal developmental toxicity study due to an increased number of skeletal malformations and retarded growth at the lowest maternally toxic dose (Tesh *et al.*, 1992). This positive result was confirmed by several unpublished rat and rabbit prenatal developmental toxicity studies, which are summarized in a classification and labeling proposal under the REACH regulation (Anonymous, 2011). In the only ZET study, fluazinam significantly induced mortality at 0.64 and 64 μM (but not at 6.4 μM), but did not induce any other effects in a statistically significant or clearly concentration-dependent manner (Truong *et al.*, 2014). Therefore, it was considered negative. A more recent ZET study showed that fluazinam started to be lethal at 0.3 μM at 96 hpf, killing all embryos at 0.7 μM , and to induce deformities in the same concentration range (Wang *et al.*, 2018). Despite several differences between the 2 ZET studies, such as the zebrafish strain used, the concentration used and the exposure duration, it seems that fluazinam acts through a general (systemic) mechanism and is not specifically embryotoxic, but may induce embryotoxic effects secondary to general effects. Prenatal developmental effects and systemic effects, as determined by maternal toxicity, are also induced by similar doses in rats and rabbits. The level of standardization of mammalian effect interpretation and the nature and severity of effects observed with fluazinam, led to an interpretation as

positive for mammalian tests. The case of fluazinam shows that although differences in the interpretation of effects may explain discordant results between ZET and mammalian tests, there may be other explanations, such as species differences in transformations (hydrolysis and metabolism) or toxicological mechanisms.

Both examples demonstrate that even when data complexity is reduced to dichotomous results through an unambiguous and transparent interpretation, reasons for discordance of results can be manifold. This applies to the concordance of ZET and mammalian results, as well as for the concordance of ZET results from different studies.

In this context, it is important to recall that we are ultimately interested in the potential of a chemical to induce prenatal developmental effects in humans that both the ZET and the mammalian tests attempt to predict. We did not include human evidence in our review, however, primarily because we expected that conclusive human evidence would be available for only a limited number of substances (Clements *et al.*, 2020). Indeed, the lack of reliable human data and the largely unknown relevance of animal prenatal developmental toxicity data for humans are major obstacles to the assessment of the value of new approaches to measuring prenatal developmental toxicity, such as the ZET. This issue is not unique to developmental toxicity. It applies to many, if not all toxicological human health effects, and has been discussed in the broader context of references for the comparison of test methods and strategies (Hoffmann *et al.*, 2008). Strategies for shifting toxicology from a strong reliance on animal data to a more human-relevant and mechanism-based discipline are being proposed and discussed, but require time, resources, and some points of reference to establish confidence (Scialli *et al.*, 2018).

Regarding the methodological challenges of applying systematic review methods to toxicological test method assessment, the conclusions and recommendations of the preparatory study have been confirmed in this full systematic review. Stephens *et al.* (2019) concluded that the application of systematic review methods to toxicological test method assessment is in principle feasible. However, numerous challenges need to be considered in planning and conducting such a review. In retrospect, the most fundamental are the following.

Scoping. The importance of an interdisciplinary review team that covers all needed expertise, especially when adapting systematic review methods to new toxicological or other environmental health applications is stressed. Given that the application of systematic review methods to toxicology is relatively new, the review team should dedicate the necessary time for toxicology domain experts to educate systematic review experts and vice versa. This should take place in the project planning phase in order to optimally scope and frame the review and to understand the requirements and implications of each step in the review. Although we engaged in this process, we nevertheless encountered some challenges, in particular total amount of potentially relevant evidence and its heterogeneity.

Efficiency. In the future, broad review questions, which are required when comprehensive test method comparisons are undertaken, can be expected to be addressed more efficiently with the help of artificial intelligence tools. Although we applied such tools to aid title and abstract screening, tools supporting the review steps of full-text screening, data extraction, and critical appraisal would be of great help. To maximize the potential

of artificial intelligence approaches for systematic reviews in toxicology and environmental health, a fundamental change in the reporting of research is needed. Common ontologies, annotations, and other approaches should be employed to improve the ability of computers to read and process the research literature (Whaley et al., 2020). A key to that change is to increase researchers' awareness of the importance of reporting completeness, which has also been called for in the context of improving reproducibility (Percie du Sert et al., 2020). This change can only be brought about through a combination of efforts, including the appropriate education and training of researchers and the creation of incentives by scientific journals and research funders.

Critical Appraisal

Improved reporting would also facilitate the critical appraisal of studies. In our review, the RoB of approximately 75% of all studies could not be assessed due to inadequate reporting. Although better reporting would help to assess RoB, reducing such bias in future studies will demand more focused efforts. Assuming that poor reporting originates from a fundamental lack of awareness of biases that have the potential to lead to overestimation of effects, education and training of researchers could gradually lead to well-planned, conducted, and reported experimental studies that reduce or eliminate sources of bias.

The concept of study sensitivity, defined as a measure of the ability of a study to detect a true effect or hazard (Cooper et al., 2016), to address important study aspects that would not be identified by a RoB assessment was particularly helpful. A more systematic and empirical exploration of this concept focused on comprehensiveness, applicability, and operationalization has the potential to facilitate and optimize systematic review approaches in environmental health and toxicology.

The application of systematic review approaches to the comparison of 2 toxicological test methods addressing the prenatal developmental effects of chemicals led us to identify contextual and methodological challenges in a transparent and objective manner. One key to overcoming these challenges is a fundamental change in how toxicological studies are planned, conducted, and reported. The first step toward bringing about this change is to create a broad awareness in the toxicological community of the urgent need for and benefits of more evidence-based approaches. This will provide the basis for creating a momentum in the community—from scientists to regulatory agencies and policymakers—to invest in the efforts needed.

We are confident that systematic review methodology will help advance the assessment of toxicological test methods, elucidating their strength and weaknesses in an evidence-based manner. It offers the flexibility to focus on various aspects of test method assessment, such as mechanistic relevance, reproducibility, predictivity, and aspects of applicability. However, advances in adjusting the review methodology for this purpose are required.

SUPPLEMENTARY DATA

Supplementary data are available at Toxicological Sciences online.

ACKNOWLEDGMENTS

The authors wish to thank the following individuals for their assistance: Daniele Wikoff, Julie Goodman, and Elisa Aissa (for discussion of the critical appraisal approach);

Jessica Palmer (for advice on the chemical selection); Seneca Fitch, Amanda McCormack, Yi Zhang, and Michalann Harthill (for supporting the development of the literature screening); and Brian Howard and Ruchir Shah (for providing access to SWIFT-Active Screener software, <https://sciome.com>; last accessed on June 15, 2021).

AUTHOR CONTRIBUTIONS

Project management and supervision: Sebastian Hoffmann, Martin L. Stephens, Katya Tsaïoun, and Thomas Hartung. ZET expertise: Francois Busquet, Catherine Willett, and Hilda Witters.

Reproduction toxicologist: Burkhard Flick.

Systematic review expertise: Manoj Lalu and Rob B.M. de Vries.

Literature search strategy and execution: Robert A. Wright.

Test method assessment and statistical expertise: Sebastian Hoffmann.

Screening and selection: Bianca Marigliani, Sevcan Gül Akgün-Ölmez, Danielle Ireland, Rebecca Cruz, Elizabeth C. Ghandakly, and Metin Ölmez.

FUNDING

The study was funded by Evidence-based Toxicology Collaboration (EBTC), which receives its funding from Center for Alternatives to Animal Testing at Johns Hopkins Bloomberg School of Public Health.

DECLARATION OF CONFLICTING INTERESTS

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

- Anonymous. (1991). Initial submission: The effects of orally administered orthophenylphenol on rat embryonal and fetal development (final report) with cover letter dated 102591 (sanitized). Epa/Ots, #88-92000188S.
- Anonymous. (1992). Initial submission: Ortho-phenylphenol: Gavage teratology study in New Zealand white rabbits (final report) with cover letter dated 062392. Epa/Ots 920004065, #88-920004065.
- Anonymous. (2011). CLH report: Proposal for harmonised classification and labelling based on regulation (EC) no. 1272/2008 (CLP Regulation), Annex VI, Part 2 - Substance name: Fluazinam. Available at: <https://echa.europa.eu/documents/10162/0a0c8670-d270-4f2d-8103-0993df75d8f9>. Accessed October 30, 2020.
- Augustine-Rauch, K., Zhang, C. X., and 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.
- 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., et al. (2014). Fishing for teratogens: A consortium effort for a harmonized zebrafish developmental toxicology assay. *Toxicol. Sci.* **139**, 210–219.

- Ballantyne, B., and Snellings, W. M. (2005). Developmental toxicity study with triethylene glycol given by gavage to CD rats and CD-1 mice. *J. Appl. Toxicol.* **25**, 418–426.
- Balls, M., Amcoff, P., Bremer, S., Casati, S., Coecke, S., Clothier, R., Combes, R., Corvi, R., Curren, R., Eskes, C., et al. (2006). The principles of weight of evidence validation of test methods and testing strategies. The report and recommendations of ECVAM workshop 58. *Altern. Lab. Anim.* **34**, 603–620.
- Bambino, K., and Chu, J. (2017). Zebrafish in toxicology and environmental health. *Curr. Top. Dev. Biol.* **124**, 331–367.
- Baumann, L., Ros, A., Rehberger, K., Neuhauss, S. C., and Segner, H. (2016). Thyroid disruption in zebrafish (*Danio rerio*) larvae: Different molecular response patterns lead to impaired eye development and visual functions. *Aquat. Toxicol.* **172**, 44–55.
- Beekhuijzen, M., de Koning, C., Flores-Guillén, M. E., de Vries-Buitenweg, S., Tobor-Kaplon, M., van de Waart, B., and 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.
- Beker van Woudenberg, A., Snel, C., Rijkmans, E., De Groot, D., Bouma, M., Hermsen, S., Piersma, A., Menke, A., and Wolterbeek, A. (2014). Zebrafish embryotoxicity test for developmental (neuro)toxicity: Demo case of an integrated screening approach system using anti-epileptic drugs. *Reprod. Toxicol.* **49**, 101–116.
- Beker van Woudenberg, A., Wolterbeek, A., Te Brake, L., Snel, C., Menke, A., Rubingh, C., de Groot, D., and Kroese, D. (2013). A category approach to predicting the developmental (neuro) toxicity of organotin compounds: The value of the zebrafish (*Danio rerio*) embryotoxicity test (ZET). *Reprod. Toxicol.* **41**, 35–44.
- Brannen, K. C., Panzica-Kelly, J. M., Danberry, T. L., and Augustine-Rauch, K. A. (2010). Development of a zebrafish embryo teratogenicity assay and quantitative prediction model. *Birth Defects Res. B Dev. Reprod. Toxicol.* **89**, 66–77.
- 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.
- Burdan, F., Siezieniewska, Z., Kis, G., and Blicharski, T. (2001). Embryofetotoxicity of acetaminophen (paracetamol) in experimental in vivo model. *Ann. Univ. Mariae Curie-Skłodowska. Sect D Med* **56**, 89–94.
- Burgdorf, T., Piersma, A. H., Landsiedel, R., Clewell, R., Kleinstreuer, N., Oelgeschläger, M., Desprez, B., Kienhuis, A., Bos, P., de Vries, R., et al. (2019). Workshop on the validation and regulatory acceptance of innovative 3R approaches in regulatory toxicology - Evolution versus revolution. *Toxicol. In Vitro* **59**, 1–11.
- Carlsson, G., and Norrgren, L. (2014). Comparison of embryo toxicity using two classes of aquatic vertebrates. *Environ. Toxicol. Pharmacol.* **37**, 24–27.
- Carney, E. W., Pottenger, L. H., Johnson, K. A., Liberacki, A. B., Tornesi, B., Dryzga, M. D., Hansen, S. C., and Breslin, W. J. (2003). Significance of 2-methoxypropionic acid formed from beta-propylene glycol monomethyl ether: Integration of pharmacokinetic and developmental toxicity assessments in rabbits. *Toxicol. Sci.* **71**, 217–228.
- Cassar, S., Adatto, I., Freeman, J. L., Gamse, J. T., Iturria, I., Lawrence, C., Muriana, A., Peterson, R. T., Van Cruchten, S., and Zon, L. I. (2020). Use of Zebrafish in Drug Discovery Toxicology. *Chem. Res. Toxicol.* **33**, 95–118.
- Chahoud, I., Ligensa, A., Dietzel, L., and Faqi, A. S. (1999). Correlation between maternal toxicity and embryo/fetal effects. *Reprod. Toxicol.* **13**, 375–381.
- Chakraborty, C., Hsu, C. H., Wen, Z. H., Lin, C. S., and Agoramoorthy, G. (2011). Effect of caffeine, norfloxacin and nimesulide on heartbeat and VEGF expression of zebrafish larvae. *J. Environ. Biol.* **32**, 179–183.
- Clements, J. M., Hawkes, R. G., Jones, D., Adjei, A., Chambers, T., Simon, L., Stemplewski, H., Berry, N., Price, S., Pirmohamed, M., et al. (2020). Predicting the safety of medicines in pregnancy: A workshop report. *Reprod. Toxicol.* **93**, 199–210.
- Collins, T. F. X., Welsh, J. J., Black, T. N., and Collins, E. V. (1981). A study of the teratogenic potential of caffeine given by oral intubation to rats. *Regul. Toxicol. Pharmacol.* **1**, 355–378.
- Collins, T. F., Welsh, J. J., Black, T. N., and Ruggles, D. I. (1983). A study of the teratogenic potential of caffeine ingested in drinking-water. *Food Chem. Toxicol.* **21**, 763–777.
- Collins, T. F., Welsh, J. J., Black, T. N., Whitby, K. E., and O'Donnell, M. W. (1987). Potential reversibility of skeletal effects in rats exposed in utero to caffeine. *Food Chem. Toxicol.* **25**, 647–662.
- Cooper, G. S., Lunn, R. M., Ågerstrand, M., Glenn, B. S., Kraft, A. D., Luke, A. M., and Ratcliffe, J. M. (2016). Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures. *Environ. Int.* **92–93**, 605–610.
- Cope, R. B., Kacew, S., and Dourson, M. (2015). A reproductive, developmental and neurobehavioral study following oral exposure of tetrabromobisphenol A on Sprague-Dawley rats. *Toxicology* **329**, 49–59.
- Dach, K., Yaghoobi, B., Schmuck, M. R., Carty, D. R., Morales, K. M., and Lein, P. J. (2019). Teratological and behavioral screening of the National Toxicology Program 91-compound library in zebrafish (*Danio rerio*). *Toxicol. Sci.* **167**, 77–91.
- Danielsson, B. R. (2013). Maternal toxicity. *Methods Mol. Biol.* **947**, 311–325.
- Daston, G. P. (2004). Developmental toxicity evaluation of butylparaben in Sprague-Dawley rats. *Birth Defects Res. B Dev. Reprod. Toxicol.* **71**, 296–302.
- Daston, G. P., Beyer, B. K., Carney, E. W., Chapin, R. E., Friedman, J. M., Piersma, A. H., Rogers, J. M., and Scialli, A. R. (2014). Exposure-based validation list for developmental toxicity screening assays. *Birth Defects Res. B Dev. Reprod. Toxicol.* **101**, 423–428.
- David, A., and Pancharatna, K. (2009). Effects of acetaminophen (paracetamol) in the embryonic development of zebrafish, *Danio rerio*. *J. Appl. Toxicol.* **29**, 597–602.
- de Souza Anselmo, C., Sardela, V. F., de Sousa, V. P., and Pereira, H. M. G. (2018). Zebrafish (*Danio rerio*): A valuable tool for predicting the metabolism of xenobiotics in humans? *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **212**, 34–46.
- de Vries, R. B. M., Hooijmans, C. R., Langendam, M. W., van Luijk, J., Leenaars, M., Ritskes-Hoitinga, M., and Wever, K. E. (2015). A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies. *Evidence Based Preclin. Med.* **2**, e00007.
- EFSA (European Food Safety Authority). (2008). Camphor in flavourings and other food ingredients with flavouring properties - Opinion of the Scientific Panel on Food Additives, Flavourings, Processing Aids and Materials in Contact with Food on a request from the Commission. *EFSA J.* **729**, 1–15.
- Field, E. A., Price, C. J., Sleet, R. B., George, J. D., Marr, M. C., Myers, C. B., Schwetz, B. A., and Morrissey, R. E. (1993). Developmental toxicity evaluation of diethyl and dimethyl phthalate in rats. *Teratology* **48**, 33–44.

- Gabbert, S., Mathea, M., Kolle, S. N., and Landsiedel, R. (2020). Accounting for precision uncertainty of toxicity testing: Methods to define borderline ranges and implications for hazard assessment of chemicals. *Risk Anal.* Available at: 10.1111/risa.13648.
- Gao, X. P., Feng, F., Zhang, X. Q., Liu, X. X., Wang, Y. B., She, J. X., He, Z. H., and He, M. F. (2014). Toxicity assessment of 7 anticancer compounds in zebrafish. *Int. J. Toxicol.* **33**, 98–105.
- Garcia, G. R., Noyes, P. D., and Tanguay, R. L. (2016). Advancements in zebrafish applications for 21st century toxicology. *Pharmacol. Ther.* **161**, 11–21.
- 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., et al. (2012). Inter-laboratory assessment of a harmonized zebrafish developmental toxicology assay - Progress report on phase I. *Reprod. Toxicol.* **33**, 155–164.
- Hamm, J. T., Ceger, P., Allen, D., Stout, M., Maull, E. A., Baker, G., Zmarowski, A., Padilla, S., Perkins, E., Planchart, A., et al. (2019). Characterizing sources of variability in zebrafish embryo screening protocols. *Altex* **36**, 103–120.
- Hanley, T. R. J., Thompson, D. J., Palmer, A. K., Beliles, R. P., and Schwetz, B. A. (1984). Teratology and reproduction studies with triclopyr in the rat and rabbit. *Fundam. Appl. Toxicol.* **4**, 872–882.
- Hartung, T., Hoffmann, S., and Stephens, M. (2013). Mechanistic validation. *Altex* **30**, 119–130.
- Hayes, W. C., Smith, F. A., John, J. A., and Rao, K. S. (1984). Teratologic evaluation of 3,6-dichloropicolinic acid in rats and rabbits. *Fundam. Appl. Toxicol.* **4**, 91–97.
- He, J. H., Gao, J. M., Huang, C. J., and Li, C. Q. (2014). Zebrafish models for assessing developmental and reproductive toxicity. *Neurotoxicol. Teratol.* **42**, 35–42.
- Heinonen, O. P., Sloan, D., and Shapiro, S. 1977. *Birth Defects and Drugs in Pregnancy*, pp. 516 + xi. Publishing Sciences Group Inc., Littleton, MAA.
- Hermesen, S. A., Van Den Brandhof, E. J., van der Ven, L. T., and Piersma, A. H. (2011). Relative embryotoxicity of two classes of chemicals in a modified zebrafish embryotoxicity test and comparison with their in vivo potencies-2. *Toxicol. In Vitro* **25**, 745–753.
- Herrmann, K. (1993). Effects of the anticonvulsant drug valproic acid and related substances on the early development of the zebrafish (*Brachydanio rerio*). *Toxicol. In Vitro* **7**, 41–54.
- Higgins, J. P. T., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D., Savovic, J., Schulz, K. F., Weeks, L., Sterne, J. A. C., et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* **343**, d5928.
- Higgins, J.P.T, Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., and Welch, V.A. (editors) (2021). *Cochrane Handbook for Systematic Reviews of Interventions Version 6.2* (updated February 2021), 2nd ed. John Wiley & Sons, Chichester, UK. Available at: www.training.cochrane.org/handbook. Accessed May 5, 2021.
- Hoffmann, S., Edler, L., Gardner, I., Gribaldo, L., Hartung, T., Klein, C., Liebsch, M., Sauerland, S., Schechtman, L., Stamatii, A., et al. (2008). Points of reference in the validation process: The report and recommendations of ECVAM Workshop 66. *Altern. Lab. Anim.* **36**, 343–352.
- Hoffmann, S., de Vries, R. B. M., Stephens, M. L., Beck, N. B., Dirven, H. A. A. M., Fowle, J. R., 3rd, Goodman, J. E., Hartung, T., Kimber, I., Lalu, M. M., et al. (2017). A primer on systematic reviews in toxicology. *Arch. Toxicol.* **91**, 2551–2575.
- Hooijmans, C. R., Rovers, M. M., de Vries, R. B., Leenaars, M., Ritskes-Hoitinga, M., and Langendam, M. W. (2014). SYRCLE's risk of bias tool for animal studies. *BMC Med. Res. Methodol.* **14**, 43.
- Horzmann, K. A., and Freeman, J. L. (2018). Making waves: New developments in toxicology with the zebrafish. *Toxicol. Sci.* **163**, 5–12.
- Hu, J., Liang, Y., Chen, M., and Wang, X. (2009). Assessing the toxicity of TBBPA and HBCD by zebrafish embryo toxicity assay and biomarker analysis. *Environ. Toxicol.* **24**, 334–342.
- ICH. (2020). ICH S5 (R3) guideline on reproductive toxicology: Detection of toxicity to reproduction for human pharmaceuticals - step 5. Available at: <https://www.ema.europa.eu/en/ich-s5-r3-guideline-reproductive-toxicology-detection-toxicity-reproduction-human-pharmaceuticals>. Accessed June 15, 2021.
- Infurna, R., Levy, B., Meng, C., Yau, E., Traina, V., Rolofson, G., Stevens, J., and Barnett, J. (1988). Teratological evaluations of atrazine technical, a triazine herbicide, in rats and rabbits. *J. Toxicol. Environ. Health* **24**, 307–319.
- Kaneda, M., Teramoto, S., Shingu, A., and Shirasu, Y. (1978). Teratogenicity and dominant lethal studies with o phenyl phenol. *J. Pest. Sci.* **3**, 365–370.
- Kennedy, G. L. J., and Kaplan, A. M. (1984). Chronic toxicity, reproductive, and teratogenic studies of hexazinone. *Fundam. Appl. Toxicol.* **4**, 960–971.
- Khera, K. S., Whalen, C., and Angers, G. (1982). Teratogenicity study on pyrethrum and rotenone (natural origin) and ronnel in pregnant rats. *J. Toxicol. Environ. Health* **10**, 111–119.
- Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., and Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203**, 253–310.
- 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., et al. (2011). Identifying developmental toxicity pathways for a subset of ToxCast chemicals using human embryonic stem cells and metabolomics. *Toxicol. Appl. Pharmacol.* **257**, 111–121.
- Kroese, E. D., Bosgra, S., Buist, H. E., Lewin, G., van der Linden, S. C., Man, H. Y., Piersma, A. H., Rorije, E., Schulpen, S. H., Schwarz, M., et al. (2015). Evaluation of an alternative in vitro test battery for detecting reproductive toxicants in a grouping context. *Reprod. Toxicol.* **55**, 11–19.
- Lankas, G. R., Cukierski, M. A., and Wise, L. D. (2004). The role of maternal toxicity in lovastatin-induced developmental toxicity. *Birth Defects Res. B Dev. Reprod. Toxicol.* **71**, 111–123.
- Lee, S. H., Kang, J. W., Lin, T., Lee, J. E., and Jin, D. I. (2013). Teratogenic potential of antiepileptic drugs in the zebrafish model. *Biomed. Res. Int.* **2013**, 726478.
- Leuschner, J. (1997). Reproductive toxicity studies of D-camphor in rats and rabbits. *Arzneimittelforschung* **47**, 124–128.
- Machener, L. (1992). Initial submission: Evaluation of embryotoxic and teratogenic effects on rats following oral administration of triadimefon with cover letter dated 08/12/92. *Epa/Ots 920009403 #88-920009403*.
- Machera, K. (1995). Developmental toxicity of cyproconazole, an inhibitor of fungal ergosterol biosynthesis, in the rat. *Bull. Environ. Contam. Toxicol.* **54**, 363–369.
- Malir, F., Ostry, V., Pfohl-Leszkowicz, A., and Novotna, E. (2013). Ochratoxin A: Developmental and reproductive toxicity—An overview. *Birth Defects Res. B* **98**, 493–502.
- Maronpot, R. R., Zelenak, J. P., Weaver, E. V., and Smith, N. J. (1983). Teratogenicity study of ethylene glycol in rats. *Drug Chem. Toxicol.* **6**, 579–594.

- McClain, R. M., Wolz, E., Davidovich, A., Edwards, J., and Bausch, J. (2007). Reproductive safety studies with genistein in rats. *Food Chem. Toxicol.* **45**, 1319–1332.
- McCormick, J. M., Paiva, M. S., Haggblom, M. M., Cooper, K. R., and White, L. A. (2010). Embryonic exposure to tetrabromobisphenol A and its metabolites, bisphenol A and tetrabromobisphenol A dimethyl ether disrupts normal zebrafish (*Danio rerio*) development and matrix metalloproteinase expression. *Aquat. Toxicol.* **100**, 255–262.
- Melo, K. M., Oliveira, R., Grisolia, C. K., Domingues, I., Pieczarka, J. C., de Souza Filho, J., and Nagamachi, C. Y. (2015). Short-term exposure to low doses of rotenone induces developmental, biochemical, behavioral, and histological changes in fish. *Environ. Sci. Pollut. Res. Int.* **22**, 13926–13938.
- Minsker, D. H., MacDonald, J. S., Robertson, R. T., and Bokelman, D. L. (1983). Mevalonate supplementation in pregnant rats suppresses the teratogenicity of mevinolinic acid, an inhibitor of 3-hydroxy-3-methylglutaryl-coenzyme a reductase. *Teratology* **28**, 449–456.
- Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* **6**, e1000097.
- Navarro, H. A., Price, C. J., Marr, M. C., Myers, C. B., and Heindel, J. J. (1992a). Developmental toxicity evaluation of 'd'-camphor (CAS no. 464-49-3) administered by gavage to Sprague-Dawley (CDN) rats on gestational days 6 through 15: Final study report and appendix. NTIS Technical Report (NTIS/PB92-170034).
- Navarro, H. A., Price, C. J., Marr, M. C., Myers, C. B., Heindel, J. J., and Schwetz, B. A. 1992b. Final report on the developmental toxicity of 'd'-camphor (CAS no. 464-49-3) in New Zealand White (NZW) (Trade name) rabbits. NTIS Technical Report 123784.
- Neeper-Bradley, T. L. (1990). Developmental toxicity evaluation of ethylene glycol administered by gavage to cd rats: Determination of a no observable effect level with cover letter. Epa/Ots 0990-0323.
- Nishimura, Y., Inoue, A., Sasagawa, S., Koiwa, J., Kawaguchi, K., Kawase, R., Maruyama, T., Kim, S., and Tanaka, T. (2016). Using zebrafish in systems toxicology for developmental toxicity testing. *Congenit. Anom.* **56**, 18–27.
- Noyes, P. D., Haggard, D. E., Gonnerman, G. D., and Tanguay, R. L. (2015). Advanced morphological - behavioral test platform reveals neurodevelopmental defects in embryonic zebrafish exposed to comprehensive suite of halogenated and organophosphate flame retardants. *Toxicol. Sci.* **145**, 177–195.
- OECD. (2005). Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment. OECD Series on Testing and Assessment, Number 34, Paris.
- OECD. (2008). Guidance document on mammalian reproductive toxicity testing and assessment. OECD Series on Testing and Assessment, Number 43, Paris.
- OECD. (2018). Test no. 414: Prenatal developmental toxicity study. In *OECD Guidelines for the Testing of Chemicals, Section 4*, OECD Publishing, Paris. Available at: 10.1787/9789264070820-en.
- Ong, L. L., Schardein, J. L., Petreter, J. A., Sakowski, R., Jordan, H., Humphrey, R. R., Fitzgerald, J. E., and de la Iglesia, F. A. (1983). Teratogenesis of calcium valproate in rats. *Fundam. Appl. Toxicol.* **3**, 121–126.
- Palmer, J. A., Smith, A. M., Egnash, L. A., Colwell, M. R., Donley, E. L. R., Kirchner, F. R., and 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., and 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.
- Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., et al. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLoS Biol.* **18**, e3000410.
- Perez, J., Domingues, I., Monteiro, M., Soares, A. M., and Loureiro, S. (2013). Synergistic effects caused by atrazine and terbuthylazine on chlorpyrifos toxicity to early-life stages of the zebrafish *Danio rerio*. *Environ. Sci. Pollut. Res. Int.* **20**, 4671–4680.
- Petreter, J. A., Anderson, J. A., Sakowski, R., Fitzgerald, J. E., and de la Iglesia, F. A. (1986). Teratogenesis of calcium valproate in rabbits. *Teratology* **34**, 263–269.
- 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., Schulp, S. H., et al. (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., Oelgeschlager, M., et al. (2018). Workshop on acceleration of the validation and regulatory acceptance of alternative methods and implementation of testing strategies. *Toxicol. In Vitro* **50**, 62–74.
- Pinho, B. R., Santos, M. M., Fonseca-Silva, A., Valentao, P., Andrade, P. B., and Oliveira, J. M. (2013). How mitochondrial dysfunction affects zebrafish development and cardiovascular function: An in vivo model for testing mitochondria-targeted drugs. *Br. J. Pharmacol.* **169**, 1072–1090.
- Price, C. J., Tyl, R. W., and Marr, M. C. (1992). Initial submission: Teratologic evaluation of ethylene glycol (CAS no. 107-21-1) administered to cd rats on gestational days 6 through 15 (final report) with attach & letter 030592. Epa/Ots 920002044, #88-920002044.
- Ren, X., Lu, F., Cui, Y., Wang, X., Bai, C., Chen, J., Huang, C., and Yang, D. (2012). Protective effects of genistein and estradiol on PAHs-induced developmental toxicity in zebrafish embryos. *Hum. Exp. Toxicol.* **31**, 1161–1169.
- Saad, M., Matheeußen, A., Bijttebier, S., Verbueken, E., Pype, C., Casteleyn, C., Van Ginneken, C., Apers, S., Maes, L., Cos, P., et al. (2017). In vitro CYP-mediated drug metabolism in the zebrafish (embryo) using human reference compounds. *Toxicol. In Vitro* **42**, 329–336.
- Saillenfait, A. M., Gallissot, F., Langonne, I., and Sabate, J. P. (2002). Developmental toxicity of N-methyl-2-pyrrolidone administered orally to rats. *Food and chemical toxicology: An international journal published for the Br. Ind. Biol. Res. Assoc.* **40**, 1705–1712.
- Scialli, A. R., Daston, G., Chen, C., Coder, P. S., Euling, S. Y., Foreman, J., Hoberman, A. M., Hui, J., Knudsen, T., Makris, S. L., et al. (2018). Rethinking developmental toxicity testing: Evolution or revolution? *Birth Defects Res.* **110**, 840–850.
- SDS-Biotech. (1997). Summary of toxicological studies on cyproconazole. *J. Pest. Sci.* **22**, 263–268.
- Seegmiller, R. E., Ford, W. H., Carter, M. W., Mitala, J. J., and Powers, W. J. (1997). A developmental toxicity study of tretinoin administered topically and orally to pregnant Wistar rats. *J. Am. Acad. Dermatol.* **36**, S60–S66.

- Selderslaghs, I. W., Blust, R., and Witters, H. E. (2012). Feasibility study of the zebrafish assay as an alternative method to screen for developmental toxicity and embryotoxicity using a training set of 27 compounds. *Reprod. Toxicol.* **33**, 142–154.
- Selderslaghs, I. W., Van Rompay, A. R., De Coen, W., and Witters, H. E. (2009). Development of a screening assay to identify teratogenic and embryotoxic chemicals using the zebrafish embryo. *Reprod. Toxicol.* **28**, 308–320.
- Song, M., Liang, D., Liang, Y., Chen, M., Wang, F., Wang, H., and Jiang, G. (2014). Assessing developmental toxicity and estrogenic activity of halogenated bisphenol A on zebrafish (*Danio rerio*). *Chemosphere* **112**, 275–281.
- Stephens, M. L., Akgun-Olmez, S. G., Hoffmann, S., de Vries, R., Flick, B., Hartung, T., Lalu, M., Maertens, A., Witters, H., Wright, R., et al. (2019). 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**, 56–58.
- Sterz, H., Nothdurft, H., Lexa, P., and Ockenfels, H. (1987). Teratologic studies on the Himalayan rabbit: New aspects of thalidomide-induced teratogenesis. *Arch. Toxicol.* **60**, 376–381.
- Teixido, E., Pique, E., Gomez-Catalan, J., and Llobet, J. M. (2013). Assessment of developmental delay in the zebrafish embryo teratogenicity assay. *Toxicol. In Vitro* **27**, 469–478.
- Tesh, J. M., Willoughby, C. R., Lambert, E. P., Wilby, O. K., and Tesh, S. A. (1992). Initial submission: Teratology Study of 3-Chloro-n-(5-chloro-2,6-dinitro-4-trifluoromethyl-phenyl) 5-trifluoro-methyl-pyridinamine in rats with cover letter dated 08/19/1992. Govt. Reports Announcements & Index 91.
- Theunissen, P. T., Beken, S., Beyer, B. K., Breslin, W. J., Cappon, G. D., Chen, C. L., Chmielewski, G., De Schaepdrijver, L., Enright, B., Foreman, J. E., et al. (2016). Comparison of rat and rabbit embryo–fetal developmental toxicity data for 379 pharmaceuticals: On the nature and severity of developmental effects. *Crit. Rev. Toxicol.* **46**, 900–910.
- Ton, C., Lin, Y., and Willett, C. (2006). Zebrafish as a model for developmental neurotoxicity testing. *Birth Defects Res. A Clin. Mol. Teratol.* **76**, 553–567.
- Truong, L., Reif, D. M., St Mary, L., Geier, M. C., Truong, H. D., and Tanguay, R. L. (2014). Multidimensional in vivo hazard assessment using zebrafish. *Toxicol. Sci.* **137**, 212–233.
- Tsaioun, K., Busquet, F., Flick, B., Hoffmann, S., Lalu, M., Stephens, M. L., de Vries, R., Witters, H., Wright, R., and Akgun-Olmez, S. G. 2018. The performance of the zebrafish embryo test (ZET) in predicting the presence and absence of malformations in the studies of prenatal development toxicity in rats and rabbits (OECD TG 414 and equivalents). A systematic review. PROSPERO: International Prospective Register of Systematic Reviews.
- Tyl, R. W., Price, C. J., Marr, M. C., Myers, C. B., and Heindel, J. J. (1991). Final report on the developmental toxicity of ethylene glycol (Cal No. 107-21-1) in New Zealand white rabbits. Volume 1 of 2: Final Study Report and Appendix. Govt. Reports Announcements & Index 110.
- Vandersea, M. W., Fleming, P., McCarthy, R. A., and Smith, D. G. (1998). Fin duplications and deletions induced by disruption of retinoic acid signaling. *Dev. Genes Evol.* **208**, 61–68.
- Vorhees, C. V. (1987). Teratogenicity and developmental toxicity of valproic acid in rats. *Teratology* **35**, 195–202.
- Wang, Y., Chen, J., Du, C., Li, C., Huang, C., and Dong, Q. (2014). Characterization of retinoic acid-induced neurobehavioral effects in developing zebrafish. *Environ. Toxicol. Chem.* **33**, 431–437.
- Wang, X. H., Zheng, S. S., Huang, T., Su, L. M., Zhao, Y. H., Souders, C. L., and Martyniuk, C. J. (2018). Fluazinam impairs oxidative phosphorylation and induces hyper/hypo-activity in a dose specific manner in zebrafish larvae. *Chemosphere* **210**, 633–644.
- Weber, G. J., Sepulveda, M. S., Peterson, S. M., Lewis, S. S., and Freeman, J. L. (2013). Transcriptome alterations following developmental atrazine exposure in zebrafish are associated with disruption of neuroendocrine and reproductive system function, cell cycle, and carcinogenesis. *Toxicol. Sci.* **132**, 458–466.
- Whaley, P., Edwards, S. W., Kraft, A., Nyhan, K., Shapiro, A., Watford, S., Wattam, S., Wolffe, T., and Angrish, M. (2020). Knowledge organization systems for systematic chemical assessments. *Environ. Health Perspect.* **128**, 125001.
- Whaley, P., Halsall, C., Ågerstrand, M., Aiassa, E., Benford, D., Bilotta, G., Coggon, D., Collins, C., Dempsey, C., Duarte-Davidson, R., et al. (2016). Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations. *Environ. Int.* **92–93**, 556–564.
- Wiegand, C., Krause, E., Steinberg, C., and Pflugmacher, S. (2001). Toxicokinetics of atrazine in embryos of the zebrafish (*Danio rerio*). *Ecotoxicol. Environ. Saf.* **49**, 199–205.
- Yang, L., Ho, N. Y., Alshut, R., Legradi, J., Weiss, C., Reischl, M., Mikut, R., Liebel, U., Müller, F., and Strähle, U. (2009). Zebrafish embryos as models for embryotoxic and teratological effects of chemicals. *Reprod. Toxicol.* **28**, 245–253.
- Yang, S., Wang, S., Sun, F., Zhang, M., Wu, F., Xu, F., and Ding, Z. (2015). Protective effects of puerarin against tetrabromobisphenol a-induced apoptosis and cardiac developmental toxicity in zebrafish embryo-larvae. *Environ. Toxicol.* **30**, 1014–1023.
- Yim, E. C., Kim, H. J., and Kim, S. J. (2014). Acute toxicity assessment of camphor in biopesticides by using *Daphnia magna* and *Danio rerio*. *Environ. Health Toxicol.* **29**.
- Zhang, J., Qian, J., Tong, J., Zhang, D., and Hu, C. (2013). Toxic effects of cephalosporins with specific functional groups as indicated by zebrafish embryo toxicity testing. *Chem. Res. Toxicol.* **26**, 1168–1181.
- Zhu, F., Wigh, A., Friedrich, T., Devaux, A., Bony, S., Nuggeoda, D., Kaslin, J., and Wlodkowic, D. (2015). Automated Lab-on-a-chip technology for fish embryo toxicity tests performed under continuous microperfusion (muFET). *Environ. Sci. Technol.* **49**, 14570–14578.