OXFORD

SOT | Society of Toxicology
www.toxsci.oxfordjournals.org

TOXICOLOGICAL SCIENCES, 171(1), 2019, 56–68

doi: 10.1093/toxsci/kfz128
Advance Access Publication Date: June 13, 2019
Review

REVIEW

Adaptation of the Systematic Review Framework to the Assessment of Toxicological Test Methods: Challenges and Lessons Learned With the Zebrafish Embryotoxicity Test

Martin L. Stephens,[a] Sevcan Gül Akgün-Ölmez,[b] Sebastian Hoffmann,[a,c] Rob de Vries,[a,d] Burkhard Flick,[e] Thomas Hartung,[f,g] Manoj Lalu,[h,i,j] Alexandra Maertens,[f] Hilda Witters,[k] Robert Wright,[l] and Katya Tsaioun[a,1]

[a]Evidence-Based Toxicology Collaboration (EBTC), Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205; [b]Department of Pharmaceutical Toxicology, Faculty of Pharmacy, Marmara University, 34722 Istanbul, Turkey; [c]Seh Consulting+Services, 33106 Paderborn, Germany; [d]SYRCLE (SYstematic Review Centre for Laboratory Animal Experimentation), Department for Health Evidence (Section HTA), Radboud Institute for Health Sciences, Radboud University Medical Center, 6525 GA Nijmegen, The Netherlands; [e]BASF SE, 67063 Ludwigshafen am Rhein, Germany; [f]Center for Alternatives to Animal Testing (CAAT), Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205; [g]University of Konstanz, CAAT-Europe, 78464 Konstanz, Germany; [h]Department of Anesthesiology and Pain Medicine; [i]Department of Cellular and Molecular Medicine, University of Ottawa; [j]Clinical Epidemiology and Regenerative Medicine Programs, Ottawa Hospital Research Institute Ottawa, Canada K1H 8L6; [k]VITO NV, 2400 Mol, Belgium; and [l]William H. Welch Medical Library, Johns Hopkins University, Baltimore, Maryland 21205

[1]To whom correspondence should be addressed. Fax: 410-614-2871. E-mail: ktsaiou1@jhu.edu.

ABSTRACT

Systematic review methodology is a means of addressing specific questions through structured, consistent, and transparent examinations of the relevant scientific evidence. This methodology has been used to advantage in clinical medicine, and is being adapted for use in other disciplines. Although some applications to toxicology have been explored, especially for hazard identification, the present preparatory study is, to our knowledge, the first attempt to adapt it to the assessment of toxicological test methods. As our test case, we chose the zebrafish embryotoxicity test (ZET) for developmental toxicity and its mammalian counterpart, the standard mammalian prenatal development toxicity study, focusing the review on how well the ZET predicts the presence or absence of chemical-induced prenatal developmental toxicity observed in mammalian studies. An interdisciplinary team prepared a systematic review protocol and adjusted it throughout this piloting phase, where needed. The final protocol was registered and will guide the main study (systematic review), which will execute the protocol to comprehensively answer the review question. The goal of this preparatory study was to translate systematic review

methodology to the assessment of toxicological test method performance. Consequently, it focused on the methodological issues encountered, whereas the main study will report substantive findings. These relate to numerous systematic review steps, but primarily to searching and selecting the evidence. Applying the lessons learned to these challenges can improve not only our main study, but may also be helpful to others seeking to use systematic review methodology to compare toxicological test methods. We conclude with a series of recommendations that, if adopted, would help improve the quality of the published literature, and make conducting systematic reviews of toxicological studies faster and easier over time.

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

Toxicology is undergoing a paradigm shift, in which new test methods are being developed that may contribute to replacing existing methods that have been used routinely for decades (Andersen and Krewski, 2009). Over time, researchers exploring a new method may standardize the protocol and accumulate data on large numbers of chemicals. An important question to ask is how well the method serves its intended purpose. For potential replacement tests, this question typically is addressed by comparing the results of the new test with those of the standard test used routinely for the toxicity of interest, ie, the reference test method (Hoffmann *et al.*, 2008). When there is a sufficient body of published literature, questions about test method performance could potentially be addressed by assessing the available evidence through a literature review (Balls *et al.*, 2006; Corvi *et al.*, 2008). Such retrospective assessments could be conducted in lieu of, or as a justification for, prospective validation trials.

Whether concerned with test method performance or other issues, reviews of the toxicological literature are typically carried out in the form of expert narratives. Typically, narrative reviews have no clear objective, articulated search strategy, predefined inclusion and exclusion criteria, protocol, or study quality and risk-of-bias assessment. Such reviews have the benefit of being relatively economical in manpower and resources in the short term, but their methodology tends to be fairly *ad hoc* and nontransparent, and, therefore, the reviews are potentially difficult to appraise, interpret, and reproduce (de Vries *et al.*, 2014). Expert narrative reviews have their place in the continuum of types of published literature summaries, for example, by generating hypotheses or presenting speculative mechanistic insight that could stimulate creativity and new ideas. However, where human health or environmental regulatory decisions are concerned, such as the selection of clinical trial design, relevant outcome or biomarker, or selection of a test method to determine toxicity of a new chemical to be used in the environment, a more systematic approach is warranted.

Here we explore systematic review methodology as a means of evaluating the performance of a given test method. In contrast to narrative reviews, systematic reviews consist of a formal series of steps that starts with the formulation of a specific question and culminates in the synthesis of the relevant data from the included papers. Systematic reviews originated in clinical medicine have been standardized for application to clinical and health care over several decades by Cochrane (Higgins and Green, 2011), and are now applied in many areas.

Within toxicology, systematic review methodology has begun to be applied to the hazard identification and risk assessment of chemicals (Birnbaum *et al.*, 2013; EFSA, 2010; Johnson *et al.*, 2014; Rooney *et al.*, 2014), but it has not yet been applied to the assessment of toxicological test methods (Stephens *et al.*, 2016). Translating this methodology to other use contexts inevitably involves some adaptation, while retaining the general approach and adhering to core principles such as transparency, comprehensiveness, and objectivity (Hoffmann *et al.*, 2017). Here we present the specific adaptations we found necessary in the present context of toxicological test methods.

For our test case of applying systematic review methods to test method performance, we chose the zebrafish embryotoxicity test (ZET) and its ability to predict the outcomes of the standard mammalian test for prenatal developmental toxicity. In the ZET, freshly fertilized zebrafish embryo eggs are exposed to various concentrations of a test substances added to their aqueous environment, usually for up to 120 h post fertilization. During this period, embryos are examined for various toxic effects that relate to mortality, general embryotoxicity (such as hatching rate and body shape), and specific embryotoxicity (Beekhuijzen *et al.*, 2015; He *et al.*, 2014; Selderslaghs *et al.*, 2009). Although the principles of the ZET are largely agreed, protocol details may vary considerably among studies, which results in a lack of harmonization and standardization (Hamm *et al.*, 2018). The standard mammalian test has been standardized as the Organisation for Economic Co-operation and Development's Test Guideline 414 (TG 414), which was first published in 1981 and revised in 2001 and in 2018 (OECD, 2018), and similar national guidelines. In brief, pregnant mammals, most often rats or rabbits, are administered test articles and toxic effects on fetuses are observed at the end of the gestation period. The developmental effects observed can be grouped as growth, external, skeletal, or soft tissue. In addition, variations, commonly defined as effect diverging beyond the usual range of structural constitution, which may not adversely affect survival or health, are discriminated from malformations, commonly defined as permanent structural changes, which may adversely affect survival, development, or function (Chahoud *et al.*, 1999; Solecki *et al.*, 2003). A further complication is added when maternal toxicity has been observed, which may have been causing or contributing to the effects observed in the fetuses. This mammalian prenatal developmental toxicity test has drawbacks including low throughput, long duration, considerable expense, and large numbers of animals used (Sipes *et al.*, 2011). These challenges are being addressed by exploring alternative approaches, either alone or in combination (Augustine-Rauch *et al.*, 2016; Ball *et al.*, 2014; Kroese *et al.*, 2015; Panzica-Kelly *et al.*, 2015).

Given the present novel application of systematic review methodology to the setting of toxicological test method assessment, we first explored methodological issues in this preparatory study. The available chapters of Cochrane's systematic review methodology for the assessment of diagnostic test accuracy (DTA) were used as a starting point (Cochrane, 2018). We focus here on the challenges encountered and lessons learned from this methodological translation. The insights will be applied to the conduct of the systematic review to comprehensively answer the review question of how predictive the ZET is of mammalian results, which will be reported elsewhere.

Acknowledging that systematic review methodology works best on narrowly defined questions, we focused our comparisons of effects on developing zebrafish (lethality, general, and specific embryotoxicity) and mammalian prenatal developmental toxicity described in external, soft tissue, and skeletal fetal examinations. Moreover, we translated data on the nature and severity of different findings into qualitative outcomes of the presence or absence of prenatal developmental toxicity. Given these considerations, we designed our systematic review to answer 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 rats and rabbit studies (OECD TG 414 and equivalents)?*

MATERIALS AND METHODS

Review team formation and protocol preparation. This study was initiated and coordinated by the Evidence-based Toxicology Collaboration (EBTC) (http://www.ebtox.org/; last accessed June 9, 2019), which is based at the Johns Hopkins Bloomberg School of Public Health. The EBTC is an international multi-stakeholder organization that seeks to facilitate the application of evidence-based approaches—including systematic review—to toxicology. The EBTC staff invited individuals with relevant expertise from an existing EBTC working group to join the review team, which undertook the present systematic review. Members were recruited who could provide the necessary and diverse expertise on mammalian reproductive toxicology, zebrafish developmental toxicity, systematic review methodology, and information science. Care was taken to recruit individuals from relevant sectors, including academia, government, industry, and nongovernmental organizations, which would help to ensure that diverse perspectives were represented. The members of the review team served as individual scientists, not as representatives of their organizations or sectors. Their initial charge was to prepare a protocol describing how the various steps in the review would be carried out.

Search strategy. The goal of the literature search strategy was to find the full set of chemicals (and their associated studies) that had been tested in both the ZET and the mammalian prenatal developmental toxicity test. The review team was familiar enough with the literature on prenatal developmental toxicity testing to realize that there was only limited published evidence directly comparing the results from the ZET and guideline studies. Consequently, rather than synthesizing pre-existing test comparisons in the literature, we identified primary studies of chemicals tested in the ZET or the mammalian assays. Specifically, the review team developed a 2-stage strategy. We first searched for ZET studies; after screening and the application of our inclusion and exclusion criteria, the resulting included studies yielded the identities of the chemicals that were tested in zebrafish embryos. In the second stage, we searched the mammalian literature for prenatal developmental toxicity studies on the same set of chemicals identified in the first stage. We designed the strategies for both stages to achieve a balance of precision and comprehensiveness in the results. Search elements included controlled vocabulary terms (ie, MeSH [Medical Subject Headings] and Emtree terms), as well as keywords applied to relevant search fields (title, abstract, descriptors, etc.).

We searched PubMed, Embase (Embase.com), BIOSIS Previews (Clarivate Analytics), and TOXLINE (National Library of Medicine) from the earliest available dates to the dates of the searches (see below). No language limits were applied in the search.

The zebrafish search consisted of a zebrafish concept, a developmental stage concept, and a toxicity concept. We ran this search in all 4 databases on April 24, 2014. The mammalian search consisted of a rat and rabbit concept, an embryo and maternal concept, and a chemicals concept consisting of the compounds identified through the zebrafish search. We ran this search in all 4 databases on March 7, 2016. The results of the zebrafish and mammalian searches were entered into EndNote for the identification and removal of duplicates. The complete search strategies are provided as a Supplementary Material.

Screening of zebrafish studies for inclusion and exclusion. Inclusion and exclusion criteria relating to technical aspects of the zebrafish studies were difficult to formulate as there has been limited standardization of the ZET. We used the following inclusion criteria:

- The study reported original data.
- The study was conducted on wild-type zebrafish (*Danio rerio*) embryos (strain reported).
- Zebrafish embryos were exposed to an individual chemical with clear identification (eg, chemical name).
- At least 3 chemical concentrations were tested in addition to a negative/vehicle control group.
- Exposure began no later than 6 h post fertilization (hpf).
- The study was performed for a duration of 48–120 hpf.
- The reported outcomes included mortality, general toxicity (ie, outcomes related to hatching, cell viability, body shape [general], edema, the cardiovascular system [heartbeat and blood flow], and the yolk sac), and specific embryotoxicity (outcomes related to body shape [specific], fins, skin, the cardiovascular system [specific, eg, alteration to blood vessels], the central nervous system, sensory organs, head, the digestive system, and trunk).
- The study included at least 10 eggs per concentration.
- The study was reported in English.

For the purposes of this preparatory study, we randomly selected a subset of 50 ZET studies, based on the assumption that this number would yield at least some eligible zebrafish studies. This was considered sufficient to optimize the search strategy and the selection process, as well as to develop extraction tables for the subsequent application in the main study, which was expected to include thousands of zebrafish studies.

The 50 randomly selected studies were subjected to title and abstract screening (level 1 screening). The studies that met the prespecified inclusion criteria and studies for which an inclusion/exclusion decision could not be made from the information in the title and abstract alone were carried forward to full text screening (level 2 screening). At both levels, all studies were screened by 2 reviewers independently. Conflicts were resolved between the screeners or, if they could not reach agreement by themselves, by a third reviewer.

Extraction of data from included zebrafish studies. The chemicals tested in the included ZET studies were identified and extracted into the Microsoft Excel table. For each chemical, the information specified in the review protocol was extracted into the same table. This included bibliographic details (first author and year of publication), study design characteristics (eg, the type of included controls, species, and strain), intervention characteristics (eg, the chemical name, concentrations tested, and start and duration of exposure), and outcomes (related to mortality and morphological alterations). The developmental effects to be assessed are not yet harmonized in the ZET community (Hamm *et al.*, 2018). We, therefore, decided on several commonly reported outcomes, such as mortality, hatching rate and delay, body shape, edema, and other alterations, to include and

extract, and these were considered sufficient for this adaptation of systematic review methodology.

Screening of the mammalian studies for inclusion and exclusion. The literature search for mammalian studies was based on the chemicals that were extracted from the included zebrafish studies. The search strings are provided in the Supplementary Material. We screened the titles and abstracts (level 1 screening) of the resulting mammalian studies.

The studies that met the inclusion criteria and studies for which an inclusion/exclusion decision could not be made from the information in the title and abstract alone were carried forward to full text screening (level 2 screening). At both levels, all studies were screened by 2 reviewers independently. Conflicts were resolved between the screeners or, if they could not reach agreement by themselves, by a third reviewer. Reasons for exclusion were documented. Both levels of screening were carried out using cloud-based SWIFT-Active Screener software (https://www.sciome.com/swift-activescreener/; last accessed June 11, 2019).

We used the following inclusion criteria in screening the mammalian studies:

- The study reported original data.
- The study was conducted on wild-type rats or rabbits (strain reported).
- Rats/rabbits were exposed to an individual chemical from the included zebrafish studies.
- At least 3 doses were administered orally in addition to a negative/vehicle control group.
- At least 4 pregnant females were treated and reported per group.
- The developing fetuses were examined for death, structural malformations and variations (external, visceral, and skeletal), and altered growth, as defined and classified by others (Chahoud *et al.*, 1999; Solecki *et al.*, 2003; Wise *et al.*, 1997).
- The study was reported in English.

Extraction of data from included mammalian studies. For each study-chemical combination that met the inclusion criteria, data were extracted from the full texts into a Microsoft Excel extraction table, according to the protocol. In addition, information on maternal toxicity was collected in order to allow for a discussion of primary embryotoxic effect versus secondary effects potentially caused by maternal toxicity, if warranted.

Risk of bias. At the time this study was conducted, no risk of bias guidance or tool that focused on toxicological animal studies was available. Consequently, we considered tools for animal studies in general (Krauth *et al.*, 2014) and for preclinical animal studies, ie, SYRCLE's risk of bias tool (Hooijmans *et al.*, 2014). We aimed for a set of risk of bias (and methodological and reporting) criteria that would apply equally to both the ZET and the mammalian prenatal developmental toxicity tests, notwithstanding that the ZET is clearly not a classical animal test and that some of its design features are characteristic of *in vitro* or ecotoxicological studies, eg, the use of microtiter plates or the "immersed" exposure in an aqueous environment.

We selected a set of 11 criteria (Table 2). Eight were drawn from the SYRCLE risk of bias tool (Hooijmans *et al.*, 2014) and 3—reporting of randomization, blinding, and sample size calculation—were the most commonly identified criteria in the Krauth *et al.* (2014) systematic review of risk of bias and methodological quality instruments for animal studies. Two reviewers applied the tool to each study independently, with the assessment options "yes," "no," and "unknown" for the risk-of-bias criteria, and the options "yes" and "no" for the reporting criteria. Any disagreements were resolved by discussion between the 2 reviewers, or by third reviewer when needed.

As information pertinent to these 11 criteria is rarely reported (Avey *et al.*, 2016; Drucker, 2016; Kilkenny *et al.*, 2009; Leung *et al.*, 2018), we approached the assessment in a manner we considered most efficient. When several chemicals and/or species were tested in a given study, it is recommended to assess the risk of bias for each species-chemical combination, as aspects pertinent to the assessment, such as the outcomes reported or the attrition rate, may differ among such combinations. However, we assumed that reporting would be consistent for each species-treatment combination in a study and assessed each study as a whole, instead of evaluating each species-treatment combination per study.

Data evaluation. Prenatal developmental toxicity hazard, ie, the potential of a chemical to cause adverse effect that is relevant for hazard assessment, was considered to be a binary outcome. Consequently, we tailored evaluation procedures for the ZET and the mammalian tests according to whether the chemical was negative (ie, not embryotoxic in the ZET or absence of adverse findings in fetal examination of the mammalian studies, respectively) or positive (ie, embryotoxic in the ZET or presence of adverse findings in the fetal examinations of the mammalian studies, respectively) in a given study.

Data analysis and the presentation of preliminary findings on test method performance were not the focus of this preparatory study. We refer interested readers to the protocol (Tsaioun *et al.*, 2018), which specifies the evaluation in detail.

RESULTS

Protocol Preparation and Amendments

The review team produced a working draft of the review protocol. Numerous amendments proved necessary, given that this preparatory study was pioneering the application of systematic review methodology to the new context of toxicological test method comparison. All amendments were tracked and incorporated into the final protocol, which is being executed in the main study, and was registered in PROSPERO, an international prospective register of systematic reviews (CRD42018096120) (Tsaioun *et al.*, 2018).

Search Results

The results of our literature search for ZET studies are summarized in Figure 1, which follows the PRISMA format (Moher *et al.*, 2009); 11 741 studies were retrieved from our search. This number was reduced to 5074 after using EndNote functionality to remove duplicates and, for the purpose of this preparatory study, documents clearly out-of-scope (eg, papers indexed as non-English and documents without original data, such as research proposals and meeting abstracts). As planned, 50 of these studies were randomly selected to explore the applicability of the adapted methodology. After screening the titles and abstracts of these studies against our inclusion/exclusion criteria (level 1 screening), 8 papers remained included. After retrieving full texts of these papers and applying the same criteria (level 2 screening), 1 paper was left. Papers were excluded at each screening level for a variety of reasons, eg, reporting no prenatal developmental toxicity outcomes or presenting no original data (Figure 1).

The one included paper by Gao *et al.* (2014), reported on a ZET study that assessed 7 chemicals for developmental toxicity effects, including structural malformations. These 7 chemicals

Table 1. Summary of Included Mammalian Studies

Study ID	Species	Strain	Chemical	Effect(s)	Overall Assessment of Prenatal Developmental Toxicity
Zhao et al. (2010)	Rat	Sprague Dawley	Gambogic acid	Decrease in fetal weight in the presence of maternal toxicity. No fetal malformations. Increase in fetal variations: rudimentary cervical ribs and retarded ossification in skull, sternebra, and vertebra	Positive
Obbink and Dalderup (1963)	Rat	Wistar albino	Thalidomide	No maternal toxicity. No effect on fetal weight. Decreased litter size based on increased number of resorptions and stillborn. No fetal malformations. Increased number of abnormal fifth sternal ossification centrum	Positive
Staples and Holtkamp (1963)	Rabbit	Dutch-belted		Tail malformations; malrotated (clubbed) limbs	Positive
Dwornik and Moore (1965)	Rat	Holtzman albino		Increased number of abnormal vertebral centra and vertebrae, increased incidence of absent fifth sternebra and miscellaneous abnormalities like poor ossification of some or all bones of the pelvis	Positive
Fratta *et al.* (1965)	Rabbit	New Zealand		Dysmelia	Positive
Schumacher *et al.* (1968)	Rabbit	New Zealand		Increased number of limb abnormalities and rib abnormalities (no detailed effect description, but assumed to be malformations)	Positive
Lehmann and Niggeschulze (1971)	Rabbit	Himalayan rabbits "Biberach"		Dose-dependent incidence of malformations, increased cleft palate	Positive
McBride (1974)	Rabbit	New Zealand white		Fetuses with multiple external malformations (at high doses; no malformations in control)	Positive
Flohé *et al.* (1981)	Rabbit	New Zealand white		Increased number of malformed fetuses (no more details, but reference to another paper)	Positive
Matsubara *et al.* (1983)	Rabbit	Japanese white; JW-NIBS rabbits		Increased hydrocephalus; microphthalmia	Positive
Sterz *et al.* (1987)	Rabbit	Himalayan rabbits		Dysmelia	Positive
Kawamura *et al.* (2014)	Rabbit	Kbl: JW rabbits		Malrotated paws; ectrodactyly, brachydactyly	Positive

were Auranofin (CAS-no. 34031-32-8), Curcumin (CAS-no. 458-37-7), Gambogic acid (CAS-no. 2752-65-0), Mycophenolic acid (CAS-no. 24280-93-1), Taxol (CAS-no. 33069-62-4), Thalidomide (CAS-no. 50-35-1), and Triptolide (CAS-no. 38748-32-2). These compounds were then used as the chemical concept in the subsequent systematic search of the mammalian literature (see Supplementary Material). This resulted in 1442 papers being retrieved, after removing duplicates; 263 papers remained included after level 1 (title and abstract) screening, and 12 of these papers met our inclusion/exclusion criteria after the level 2 (full text) review. The reasons for exclusion are reported in Figure 2. These 12 papers were included in the final analysis.

Test Results

For the 7 chemicals assessed by the included zebrafish study, acute toxicity, and cardiovascular toxicity, as well as developmental toxicity, were evaluated by Gao *et al.* (2014). Mammalian prenatal developmental toxicity data were found on 2 of these chemicals, namely, gambogic acid and thalidomide. These 2 chemicals caused treatment-related findings in the zebrafish, ie, missing pectoral fins for both chemicals and reduced pigmentation for gambogic acid (Gao *et al.*, 2014).

Of the 12 included mammalian studies, 11 assessed thalidomide, and 1 gambogic acid. In these studies, thalidomide was tested in both rats (2 studies) and rabbits (9 studies), whereas gambogic acid was tested only in rats. Treatment-related adverse findings of prenatal developmental toxicity were described for thalidomide in both rats and rabbits, and for gambogic acid in rats, for which no rabbit data were available (Table 1). Among the reported fetal malformations for thalidomide in rabbits were increased cleft palate, hydrocephalus, microphthalmia, dysmelia, malrotated limbs, and spina bifida. In rats, thalidomide was found to have resulted in prenatal developmental toxicity, such as abnormalities of the vertebral centrum and the fifth sternal ossification centrum. For gambogic acid, the prenatal developmental toxicity in the rat was manifested in an increase of fetal skeletal alterations. The variations reported for this chemical were rudimentary cervical ribs and delayed skull and sternebral ossifications, as well as retarded ossifications of vertebra (Table 1).

Given that our search of the mammalian literature yielded studies with exposures to only 2 out of 7 of these chemicals, ie, thalidomide and gambogic acid, we can compare the mammalian prenatal developmental toxicity results with the ZET results only for these, which showed treatment-related findings in both species.

Risk of Bias

The risk of bias and the reporting quality of the included zebrafish ($N = 1$) and mammalian ($N = 12$) papers were assessed (Table 2). For the 13 studies, the vast majority of the risk-of-bias criteria was rated as "unknown" (indicated as yellow in Table 2), as the necessary information was not reported. The same holds true for the 3 reporting criteria: with a very few exceptions, none of the relevant

Table 2. Assessment of Risk-of-Bias and Reporting Quality of Included Mammalian and Zebrafish Studies

Study Type	Study ID	Risk-of-Bias Criteria (Hooijmans *et al.*, 2014)								Reporting criteria		
		Were the Groups Similar at Baseline or Adjusted for Confounders?	Was the Allocation Sequence Adequately Generated and Applied?	Was the Allocation Adequately Concealed?	Were the Animals Randomly Housed During the Experiment?	Were the Caregivers/ Investigators During the Course of the Experiment Adequately Blinded?	Were Animals Selected at Random During Outcome Assessment?	Was the Outcome Assessment Adequately Blinded?	Were Incomplete Outcome Data Adequately Addressed?	Is It Mentioned That the Experiment Was Randomized?	Is It Mentioned That the Experiment Was Blinded?	Is a Power/ Sample Size Calculation Shown?
Mammalian	Obbink and Dalderup (1963)									[a]		
	Staples and Holtkamp (1963)											
	Dwornik and Moore (1965)											
	Fratta *et al.* (1965)											
	Schumacher *et al.* (1968)											
	Lehmann and Niggeschulze (1971)											
	McBride (1974)											
	Flohé *et al.* (1981)											
	Matsubara *et al.* (1983)											
	Sterz *et al.* (1987)											
	Zhao et al. (2010)											
	Kawamura *et al.* (2014)											
Zebrafish	Gao *et al.* (2014)											

Green, yes; red, no; yellow, unknown.

[a]Randomization is mentioned for rats, but not for rabbits.

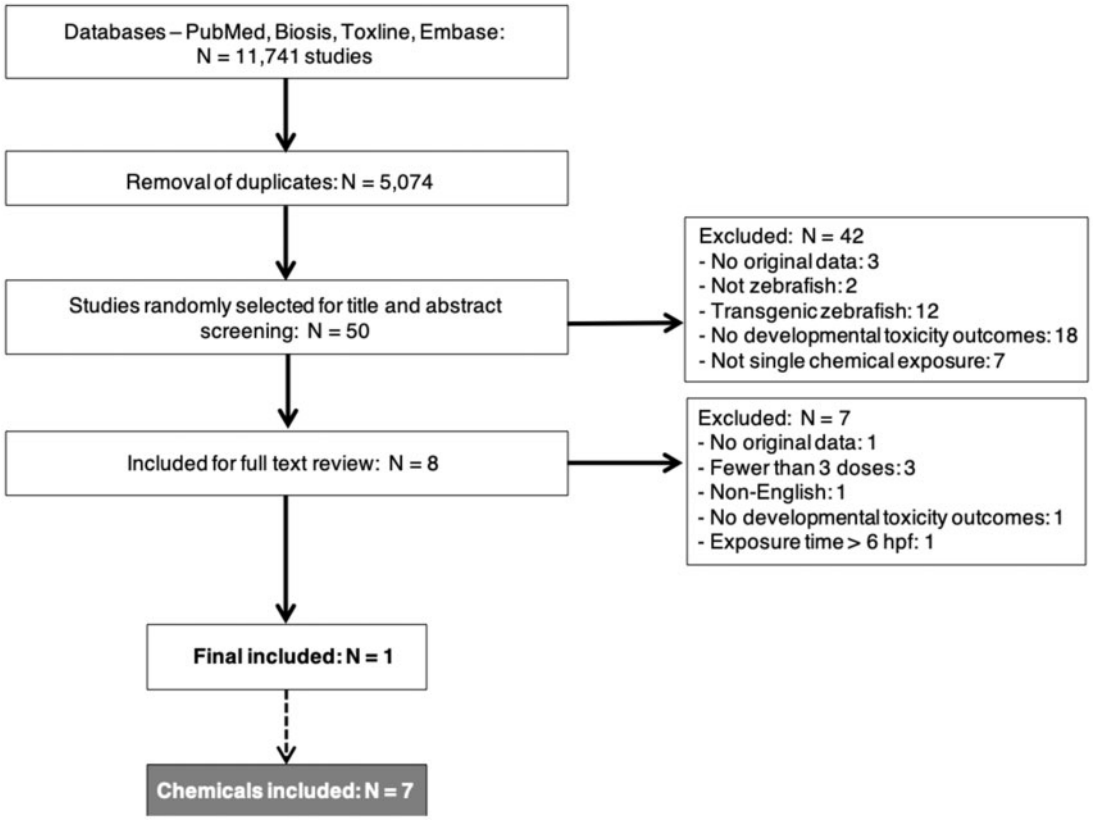

Figure 1. Preferred reporting items for a systematic review and meta-analysis (PRISMA) flow diagram for the zebrafish studies retrieved from the literature search (hpf: hours post fertilization).

information was reported (indicated as red in Table 2). There was no substantive difference observed between the risk of bias in the mammalian studies and the zebrafish study. Although the number of studies was small, it is worth noting that there was no obvious trend in the more recent mammalian literature toward more detailed reporting, as the older studies were equally likely (or not) to contain the information.

Challenges and Lessons Learned

In addition to the preparation of the protocol that adapts systematic review methodology to toxicological test method assessment, we consider the identification of the challenges encountered and the lessons learned as the primary result of this preparatory study. Here we provide an annotated listing of these challenges and lessons learned, organized under the headings of the typical steps of a systematic review.

Formulating the question.

- As outcomes in zebrafish and mammals cannot be compared directly due to differences in anatomy and embryogenesis, we chose embryotoxicity in zebrafish and prenatal developmental toxicity in mammals as nonspecific outcomes that subsume various effects. This broadness in the review question was necessary to render outcomes of the test methods comparable, but it also presented a challenge in several of the subsequent review steps.

Searching the evidence.

- In the absence of studies that tested the same chemicals in parallel in both tests, a novel 2-stage strategy was devised, first to identify studies that tested chemicals in the ZET and, second, to identify studies that tested the same chemicals in mammalian prenatal developmental toxicity studies.
- Fine-tuning of the search strategy was made difficult by the fact that MeSH terms in MEDLINE/PubMed are oriented towards clinical medicine, and do not capture the relevant fields for toxicology (see eg, https://www.nlm.nih.gov/mesh/meshhome.html/ last accessed June 11, 2019).

Selecting the evidence.

- Due to the lack of structured abstracts in the screened literature, it was often challenging to identify the information pertinent to the inclusion and exclusion criteria, which considerably slowed down the efficiency of identifying relevant studies.
- Substantial diversity in reported ZET outcomes, likely a consequence of the lack of protocol harmonization in the field, complicated the identification of relevant studies.

Extracting data.

- Data extraction was made difficult by the lack of a commonly accepted ontology for adverse outcomes in zebrafish studies and, in general, by a failure to use a controlled vocabulary for reporting study information.

Analyzing data.

- Several challenges, some of which have been mentioned above, compelled us to adopt a simplified approach to data analysis, focusing on the presence or absence of any kind of structural alterations. All efforts to make these considerations as transparent and rational as possible were undertaken and are explained in the protocol (Tsaioun *et al.*, 2018).

Reporting.

- We observed that the published study reports in our sample were inadequately reported to fully assess risk-of-bias and

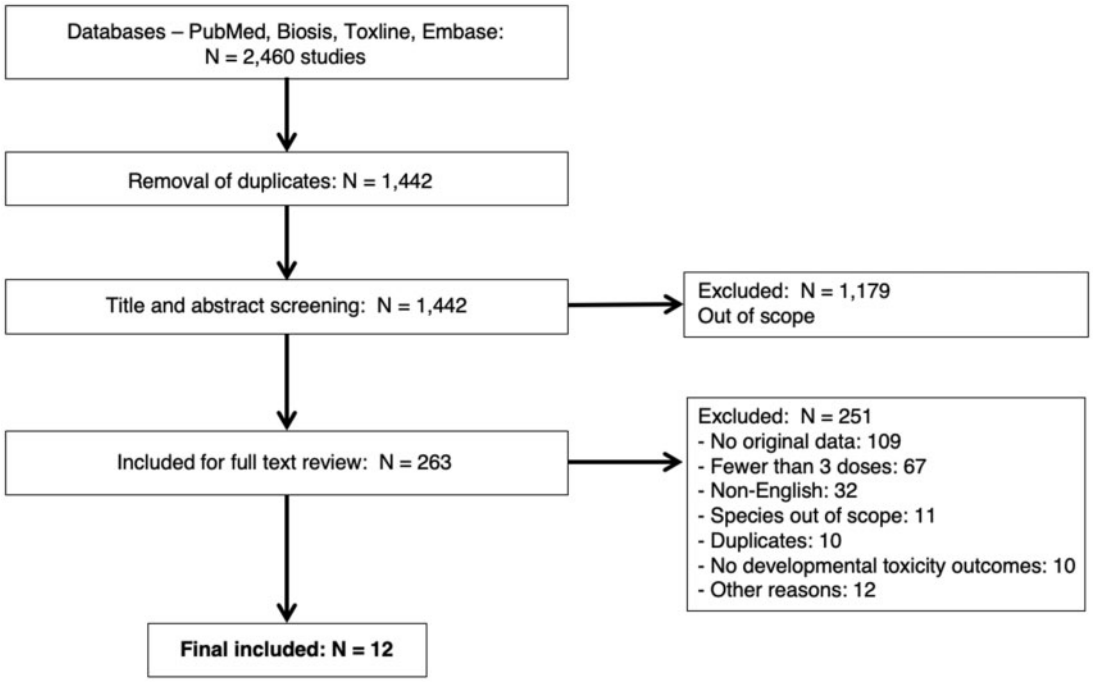

Figure 2. Preferred reporting items for a systematic review and meta-analysis (PRISMA) flow diagram for the mammalian studies retrieved from the literature search.

methodological quality (Table 2), which suggests that this may be the case throughout this literature. This would be consistent with the finding of poor quality of reporting of experimental animal studies, both observed by other reviews related to toxicology (Koustas *et al.*, 2014; NTP, 2016) and in the literature of preclinical animal studies (Freedman *et al.*, 2017; Sena *et al.*, 2014; Tihanyi *et al.*, 2019).

DISCUSSION

Clinical medicine has pioneered systematic review methodology (Chandler and Hopewell, 2013). This methodology is being adopted and adapted to other scientific disciplines such as environmental health science (Rooney *et al.*, 2014, 2016 and Sutton, 2014), preclinical animal studies (Hirst *et al.*, 2014; Yauw *et al.*, 2015), and other fields (Okoli and Schabram, 2010) because of its structured framework, transparency, comprehensiveness, reproducibility, and objectivity. This preparatory study's goal was the translation of systematic review methodology to a new use context: the assessment of test method performance in toxicology. Consequently, we focused on the methodological issues encountered, whereas the main study will report substantive findings.

The methodological translation proved feasible but raised a number of issues and presented numerous challenges. Before discussing these, we briefly comment on the relevance of this preparatory study to the review question, concerning the ability of the ZET to predict prenatal developmental toxicity in the mammalian guideline studies. First, the results of the literature search indicate that there is a substantial literature on the testing of chemicals in both sets of tests (Figs. 1 and 2). Extrapolating the inclusion ratio of 1 in 50 studies from this preparatory study to the main study with several thousand ZET studies, it can be expected that approximately 100 studies may be included in the main study. As these will have tested several hundred chemicals, many of which will also have eligible mammalian prenatal developmental toxicity studies, it can be expected that the main study will be based on a substantial amount of evidence. A second noteworthy finding was the poor reporting, resulting in an unknown risk of bias in the included individual studies (Table 2). Projected to the main study, this may limit the level of confidence in the review's conclusions. Since the time this study was conducted, several critical appraisal tools focused on toxicology/environmental health have been be published (Beronius *et al.*, 2018; Rooney *et al.*, 2014; Woodruff and Sutton, 2014). To the extent that these tools assess risk of bias, they basically address the same bias domains as were assessed here and would have led to largely the same conclusions regarding risk of bias in the studies included here. As a consequence, the risk-of-bias assessment will be reconsidered in relation to other factors that potentially influence the confidence in results of individual studies, such as the physical-chemical properties of chemicals, eg, low water solubility, that may lead to reduced exposure in the ZET, or the interpretation of adverse effects in the mammalian studies in the presence of maternal toxicity. Third, using the nonspecific outcome of embryotoxicity in the ZET and prenatal developmental toxicity in mammals, results of the ZET and the mammalian prenatal developmental toxicity tests were comparable. However, the small number of chemicals, which resulted from our methodological focus, did not—as expected—allow us to draw any sound conclusions regarding the performance of the ZET. This preparatory study, in particular the developed protocol, has provided the methodology to be used for the main study, which will evaluate all evidence obtained from the literature search. This is expected to result in a sufficiently large number of chemicals to compare the test methods and calculate predictive performance parameters.

As this study was the pioneering effort to adapt the systematic review framework to the context of test method assessment, we made a few decisions to facilitate—and learn from—this transition. First, we focused our comparisons on

embryotoxicity *in vitro* and prenatal developmental toxicity *in vivo* only, setting aside potential comparisons of other adverse developmental effects, such as developmental neurotoxic outcomes. In the future, the test methods could be more fully compared in a comprehensive systematic review with subgroup analyses, eg, focusing on specific outcomes or group of outcomes. Second, within our chosen domain, we translated data on the nature and severity of embryotoxicity in the ZET and prenatal developmental toxicity in mammals into qualitative outcomes of the presence or absence of treatment-related alterations. And finally, we limited the mammalian prenatal developmental toxicity studies to those involving rats or rabbits, given that these species have been more commonly used than other species.

These decisions resulted in the following main study review 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 rats and rabbit studies (*OECD TG 414 and equivalents*)? In systematic review terminology, this is fundamentally a PECO question design (eg, Morgan *et al.*, 2018; Woodruff and Sutton, 2014)—that is, the question addresses the Populations (exposed zebrafish embryos, rats, and rabbits fetuses), Exposure (to individual chemicals), Comparison (comparator test: rats/rabbits prenatal developmental toxicity test), and Outcome (embryotoxicity). Simply put, the review compares the embryotoxicity hazards of chemicals in the ZET with the prenatal developmental toxicity observed in rats and rabbits. The outcomes included in this question were broad. They covered many morphological alterations of the embryos in the ZET as well as many morphological alterations in external, soft tissue, and skeletal fetal examinations specific for mammalian prenatal development, which are due to species differences not directly comparable. In contrast, questions should be narrow to make them amenable to systematic review (Hoffmann *et al.*, 2017). Therefore, the comparison of test methods that provide information on a broad range of outcomes that all inform the same hazard is a fundamental challenge that causes problems in the subsequent systematic review steps, eg, for the study selection (What is the minimum set of outcomes that a study must report to be eligible?) and for the data analysis (How to summarize repeat studies of the same chemical?). These aspects should be thoroughly considered when embarking on a systematic review to compare toxicological test methods.

We faced some challenges when translating our PECO question into database search strategies, notably when generating toxicology-related search terms from PubMed's controlled vocabulary MeSH. Although we identified a number of relevant MeSH terms for our search (eg, "Toxicity Tests"[Mesh]), we found a general lack of robustness in MeSH's coverage of toxicology. We addressed this challenge by carefully including relevant keywords in our PubMed search and by running our search in multiple databases. Because of their unique features, these other databases added results not found by PubMed. For instance, Embase has its own controlled vocabulary that addresses toxicology with more robustness, and TOXLINE has an explicit focus on the toxicological literature.

Our starting point in this preparatory study was the methodology for a standard systematic review as used in clinical medicine, as suggested by Hartung (2010). Led by Cochrane, clinical medicine has pioneered systematic review methodology in the context of assessing the effectiveness of "interventions" such as new drugs or surgical techniques (Chandler and Hopewell, 2013). Cochrane is currently translating systematic review methodology to the assessment of DTA, a context that has some parallels to assessing test methods in toxicology (Hoffmann and Hartung, 2005, 2006). Cochrane has been publishing chapters of its *Handbook for DTA Review* online as they become available (Cochrane, 2018). Although intended for a different context, this emerging Handbook was helpful to the present study in a number of areas, especially in protocol preparation and terminology. However, the parallels between this clinical situation (assessing DTA) and the toxicological situation (assessing test performance) are limited in practice. For example, in the clinical situation, the reviewed studies are themselves direct comparisons of the diagnostic tests under review, thus rendering the review essentially a compilation of pre-existing comparisons, as reflected, for example, in the preferred reporting items for a systematic review and meta-analysis (PRISMA) for studies of diagnostic test accuracy (McInnes *et al.*, 2018). In the toxicology context, however, the studies relevant for comparison of 2 toxicological test methods are typically studies of individual chemicals in one or the other test, but not studies comparing both.

In addition, this effort should be put in the larger context of applying systematic review methodology to the evaluation of toxicological studies for environmental health questions and in chemical risk assessment (Rooney *et al.*, 2014; Whaley *et al.*, 2016; Woodruff and Sutton, 2014). This application commonly takes the form of reviewing the evidence that associates chemical exposure with a specific health effect (see eg, Cano-Sancho *et al.*, 2017; Koustas *et al.*, 2014; NTP, 2016). Although the number of such systematic reviews is increasing, the adaptation of the systematic review methodology for this purpose still faces challenges, such as rating the confidence in the body of evidence or integrating the evidence from human, animal, and nonanimal studies (Morgan *et al.*, 2016). Although the PECO questions of the 2 systematic review applications (exposure effects vs test method comparisons) are substantially different, close connections of these 2 applications are evident in other review steps. Literature sources will be similar and some search concepts, eg, for the exposure (chemicals and their synonyms) and for outcomes, are required for both applications. In addition, eligibility criteria related to study design are likely to be similar, as both applications are usually based on studies that fulfill at least basic design requirements, as are approaches to critical appraisal of studies. In contrast, rating the confidence in the body of evidence will differ in some regards, as aspects such as consistency, precision, and effect size are not directly applicable to test method comparisons. In addition and as in the clinical field, data analysis approaches will not have any great similarity.

We conclude with several recommendations that stem from the challenges and lessons identified above.

- *MeSH search terms:* The terminology and hierarchy of MeSH search terms in PubMed should be expanded to provide more utility to toxicology.
- *Structured abstracts:* Toxicology journals should consider requiring structured abstracts that call for critical types of information to be present, labeled, and listed in a certain sequence, as has been called for in clinical studies (Mulrow *et al.*, 1988). An additional advantage of structured abstracts is that they are more amenable to automated approaches, such as machine-reading.
- *Completeness of reporting:* Methodological details and study results should be reported in sufficient detail to permit readers to assess how confident to be in the results and conclusions, as well as how to replicate a given study (Avey *et al.*, 2016; Drucker, 2016; Kilkenny *et al.*, 2009; Leung *et al.*, 2018). Numerous guidelines are

available for reporting quality, see eg, Samuel *et al.* (2016). Also, improved and comprehensive reporting of studies in toxicological databases would be helpful and could ultimately qualify them as eligible evidence sources for systematic review purposes.

- *Risk of bias*: As already called for by Rooney *et al.* (2016), studies providing empirical evidence on the impact of individual biases on toxicological evidence should be conducted. Once a bias has been demonstrated to be influential in the toxicological literature, concrete steps could be taken to minimize such potential biases in experimental studies not only by the researchers, but also by organizations conducting, commissioning and funding the studies, and by regulatory agencies. In addition, the importance of the risk-of-bias assessment needs to be assessed in relation to how other factors, such as external validity, potentially influence conclusion.
- *Ontologies/controlled vocabularies:* An ontology and/or controlled vocabulary can expedite the systematic review process (by streamlining data compilation), and ultimately, can make machine-learning and data-mining approaches possible (Hardy *et al.*, 2012a,b). An ontology and/or controlled vocabulary should be developed (or aligned to an existing) as early in the development of a new test method as is appropriate, which is especially important for test methods with many potential outcomes and those that involve organisms or tissues more distantly related to those used routinely.

If followed, these recommendations would not only facilitate the conduct of systematic reviews, but also promote the broader goal of producing reliable science to inform regulatory decisions.

In conclusion, we argue that systematic reviews of the assessment of toxicological test methods are feasible, although challenging. A practical prerequisite for a definitive systematic review in this context, as in others, is that sufficient relevant evidence is publicly available, which might not be the case for every new toxicological test method. Some systematic reviews have value in documenting the limited extent of available evidence, and thus flagging a data gap. However, it is difficult to assess whether the toxicological community would be well served by a full-blown systematic review of test method comparisons that simply flagged a data gap; other approaches are better suited, eg, evidence maps (Miake-Lye *et al.*, 2016). Moreover, there should be data of sufficient detail and quality in the routinely used species (or human outcomes), retrievable by a systematic literature search, in order to provide a basis for comparison of the new tests. For regulatory evaluations in the future, one could envision test developers submitting detailed information about the mechanistic basis of a new test (Hartung *et al.*, 2013), along with a dataset exhibiting low risk of bias. It could then be compared to the outcomes of the standard guideline test and human outcomes, both obtained through systematic literature searches.

It is acknowledged that the comparison of 2 test methods will be increasingly replaced by comparing combinations of test methods against a single or composite reference standard. Such combinations of various test method and other information, for example in testing strategies and integrated approaches to testing and assessment, will be essential, for example, in implementing *Toxicity Testing in the 21st Century* (NRC, 2007). The challenges in designing and assessing such strategic approaches have been identified, but the discussions of solutions continues (Jaworska and Hoffmann, 2010; Piersma *et al.*, 2018; Rovida *et al.*, 2015). The complexity goes far beyond the direct comparison of 2 test methods for the same purpose as planned in our review, which has successfully been used for so-called one-to-one replacements of *in vivo* test method by a non-animal test method (see eg, Spielmann *et al.*, 2007). However, assuming that a performance assessment will also be required for testing strategies, a systematic review approach could be equally applied. Therefore, our methodological adaptation to one-to-one comparisons will also be of value for more complex situations. For example, test methods addressing the same mechanistic event could be compared systematically or a combination of test methods could be compared to reference results (Kleinstreuer *et al.*, 2018).

Although it can be anticipated that a substantial effort is required to assess toxicological test methods, either individually or in combination, the advent of artificial intelligence (AI) and machine learning (ML) bears the promise of increasing efficiency, ultimately enabling updates of existing systematic reviews in real time. This will make this application much more pragmatic, as compared to prospective studies, eg, when formally validating test methods according to international requirements (Hartung *et al.*, 2004; OECD, 2005). In addition, various methodological challenges remain that call for the adaptation of existing methodology, if not for new approaches. The necessary methodological solutions should adhere to the fundamental evidence-based principles of transparency, objectivity, and consistency, and should be agreed on by all interested stakeholders. As these solutions are developed, systematic review may become a standard tool for the retrospective evaluation of toxicological test methods.

SUPPLEMENTARY DATA

Supplementary data are available at *Toxicological Sciences* online.

DECLARATION OF CONFLICTING INTERESTS

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

ACKNOWLEDGMENTS

The authors wish to thank the following individuals for their assistance: Francois Busquet (for zebrafish expertise); Kate Willet (for supporting protocol development and screening); Bianca Margliani, Michele Palopoli, Tyna Dao, and Xiaoxing Cui (for supporting the literature screening); and Brian Howard and Ruchir Shah (for providing access to SWIFT-Active Screener software, https://sciome.com).

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.

REFERENCES

Andersen, M. E., and Krewski, D. (2009). Toxicity testing in the 21st century: Bringing the vision to life. *Toxicol. Sci.* **107**, 324–330.

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.

Avey, M. T., Moher, D., Sullivan, K. J., Fergusson, D., Griffin, G., Grimshaw, J. M., Hutton, B., Lalu, M. M., Macleod, M., Marshall, J., *et al.* (2016). The devil is in the details: Incomplete reporting in preclinical animal research. *PLoS One* **11**, e0166733.

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.

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.

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.

Beronius, A., Molander, L., Zilliacus, J., Rudén, C., and Hanberg, A. (2018). Testing and refining the science in risk assessment and policy (SciRAP) web-based platform for evaluating the reliability and relevance of in vivo toxicity studies. *J. Appl. Toxicol.* **38**, 1460–1470.

Birnbaum, L. S., Thayer, K. A., Bucher, J. R., and Wolfe, M. S. (2013). Implementing systematic review at the national toxicology program: Status and next steps. *Environ. Health Perspect.* **121**, a108–a109.

Cano-Sancho, G., Salmon, A. G., and La Merrill, M. A. (2017). Association between exposure to p, p'-DDT and its metabolite p, p'-DDE with obesity: Integrated systematic review and meta-analysis. *Environ. Health Perspect.* **125**, 096002.

Chahoud, I., Buschmann, J., Clark, R., Druga, A., Falke, H., Faqi, A., Hansen, E., Heinrich-Hirsch, B., Hellwig, J., Lingk, W., *et al.* (1999). Classification terms in developmental toxicology: Need for harmonization. *Reprod. Toxicol.* **13**, 77–82.

Chandler, J., and Hopewell, S. (2013). Cochrane methods—Twenty years experience in developing systematic review methods. *Syst. Rev.* **2**, 76.

Cochrane. (2018). *Handbook for DTA Reviews*. Available at: https://methods.cochrane.org/sdt/handbook-dta-reviews. Accessed December 10, 2018.

Corvi, R., Albertini, S., Hartung, T., Hoffmann, S., Maurici, D., Pfuhler, S., van Benthem, J., and Vanparys, P. (2008). ECVAM retrospective validation of in vitro micronucleus test (MNT). *Mutagenesis* **23**, 271–283.

De Vries, R. B., Wever, K. E., Avey, M. T., Stephens, M. L., Sena, E. S., and Leenaars, M. (2014). The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *ILAR J.* **55**, 427–437.

Drucker, D. J. (2016). Never waste a good crisis: Confronting reproducibility in translational research. *Cell Metab.* **24**, 348–360.

Dwornik, J. J., and Moore, K. L. (1965). Skeletal malformations in the Holtzman rat embryo following the administration of thalidomide. *J. Embryol. Exp. Morphol.* **13**, 181–193.

EFSA. (2010). Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA J.* **8**, 1637.

Flohé, L., Draeger, E., Frankus, E., Graudums, I., Günzler, W. A., Helm, F. C., and Kuutti-Savolainen, E. R. (1981). Studies on the hypothetical relationship of thalidomide-induced embryopathy and collagen biosynthesis. *Arzneimittelforschung* **31**, 315–320.

Fratta, I. D., Sigg, E. B., and Maiorana, K. (1965). Teratogenic effects of thalidomide in rabbits, rats, hamsters, and mice. *Toxicol. Appl. Pharmacol.* **7**, 268–286.

Freedman, L. P., Venugopalan, G., and Wisman, R. (2017). Reproducibility2020: Progress and priorities. *F1000Res.* **6**, 604.

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.

Hamm, J. T., Ceger, P., Allen, D., Stout, M., Maull, E. A., Baker, G., Zmarowski, A., Padilla, S., Perkins, E., Planchart, A., *et al.* (2018) Characterizing sources of variability in zebrafish embryo screening protocols. *ALTEX* **36**, 103–120.

Hardy, B., Apic, G., Carthew, P., Clark, D., Cook, D., Dix, I., Escher, S., Hastings, J., Heard, D. J., Jeliazkova, N., *et al.* (2012a). A toxicology ontology roadmap. *ALTEX* **29**, 129–137.

Hardy, B., Apic, G., Carthew, P., Clark, D., Cook, D., Dix, I., Escher, S., Hastings, J., Heard, D. J., Jeliazkova, N., *et al.* (2012b). Toxicology ontology perspectives. *ALTEX* **29**, 139–156.

Hartung, T. (2010). Evidence based-toxicology—The toolbox of validation for the 21st century? *ALTEX* **27**, 241–251.

Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., Hoffmann, S., Roi, A. J., *et al.* (2004). A modular approach to the ECVAM principles on test validity. *Altern. Lab. Anim.* **32**, 467–472.

Hartung, T., Hoffmann, S., and Stephens, M. (2013). Mechanistic validation. *ALTEX* **30**, 119–130.

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.

Higgins, J. P., and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. Available at: http://handbook-5-1.cochrane.org. Accessed November 22, 2018.

Hirst, J. A., Howick, J., Aronson, J. K., Roberts, N., Perera, R., Koshiaris, C., and Heneghan, C. (2014). The need for randomization in animal trials: An overview of systematic reviews. *PLoS One* **9**(6): e98856. https://doi.org/10.1371/journal.pone.0098856.

Hoffmann, S., de Vries, R. B. M., Stephens, M. L., Beck, N. B., Dirven, H. A. A. M., Fowle, J. R., 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.

Hoffmann, S., Edler, L., Gardner, I., Gribaldo, L., Hartung, T., Klein, C., Liebsch, M., Sauerland, S., Schechtman, L., Stammati, 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., and Hartung, T. (2005). Diagnosis: Toxic!—Trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations. *Toxicol. Sci.* **85**, 422–428.

Hoffmann, S., and Hartung, T. (2006). Toward an evidence-based toxicology. *Hum. Exp. Toxicol.* **25**, 497–513.

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.

Jaworska, J., and Hoffmann, S. (2010). Integrated testing strategy (ITS)—Opportunities to better use existing data and guide future testing in toxicology. *ALTEX* **27**, 231–342.

Johnson, P. I., Sutton, P., Atchley, D. S., Koustas, E., Lam, J., Sen, S., Robinson, K. A., Axelrad, D. A., and Woodruff, T. J. (2014). The navigation guide—Evidence-based medicine meets environmental health: Systematic review of human evidence for PFOA effects on fetal growth. *Environ. Health Perspect.* **122**, 1028–1039.

Kawamura, Y., Shirotsuka, Y., Awatsuji, H., Matsumoto, K., and Sato, K. (2014). Common nature in the effects of thalidomide on embryo-fetal development in Kbl: JW and Kbl: NZW rabbits. *Congenit. Anom.* **54**, 41–53.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M. F. W., Cuthill, I. C., Fry, D., Hutton, J., and Altman, D. G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* **4**, e7824.

Kleinstreuer, N. C., Hoffmann, S., Alépée, N., Allen, D., Ashikaga, T., Casey, W., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., *et al.* (2018). Non-animal methods to predict skin sensitization (II): An assessment of defined approaches. *Crit. Rev. Toxicol.* **48**, 359–374.

Koustas, E., Lam, J., Sutton, P., Johnson, P. I., Atchley, D. S., Sen, S., Robinson, K. A., Axelrad, D. A., and Woodruff, T. J. (2014). The navigation guide—Evidence-based medicine meets environmental health: Systematic review of nonhuman evidence for PFOA effects on fetal growth. *Environ. Health Perspect.* **122**, 1015–1027.

Krauth, D., Anglemyer, A., Philipps, R., and Bero, L. (2014). Nonindustry-sponsored preclinical studies on statins yield greater efficacy estimates than industry-sponsored studies: A meta-analysis. *PLoS Biol.* **12**(1): e1001770.

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.

Lehmann, H., and Niggeschulze, A. (1971). The teratologic effects of thalidomide in Himalayan rabbits. *Toxicol. Appl. Pharmacol.* **18**, 208–219.

Leung, V., Rousseau-Blass, F., Beauchamp, G., and Pang, D. S. J. (2018). ARRIVE has not ARRIVEd: Support for the ARRIVE (animal research: Reporting of in vivo experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLoS One* **13**, e0197882.

Matsubara, Y., Goto, M., Mikami, T., Suzuki, Y., and Chiba, T. (1983). Teratogenic effects of thalidomide in the rabbit: Difference in susceptibility between two breeds. *Congenit. Anom.* **23**, 223–229.

McBride, W. G. (1974). Fetal nerve cell degeneration produced by thalidomide in rabbits. *Teratology* **10**, 283–291.

McInnes, M. D. F., Moher, D., Thombs, B. D., McGrath, T. A., Bossuyt, P. M., Clifford, T., Cohen, J. F., Deeks, J. J., Gatsonis, C., Hooft, L., *et al.* (2018). Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. *JAMA* **319**, 388–396.

Miake-Lye, I. M., Hempel, S., Shanman, R., and Shekelle, P. G. (2016). What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. *Syst. Rev.* **5**, 28.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *J. Clin. Epidemiol.* **62**, 1006–1012.

Morgan, R. L., Thayer, K. A., Bero, L., Bruce, N., Falck-Ytter, Y., Ghersi, D., Guyatt, G., Hooijmans, C., Langendam, M., Mandrioli, D., *et al.* (2016). GRADE: Assessing the quality of evidence in environmental and occupational health. *Environ. Int.* **92–93**, 611–616.

Morgan, R. L., Whaley, P., Thayer, K. A., and Schünemann, H. J. (2018). Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environ. Int.* **121**, 1027–1031.

Mulrow, C. D., Thacker, S. B., and Pugh, J. A. (1988). A proposal for more informative abstracts of review articles. *Ann. Intern. Med.* **108**, 613–615.

NRC. (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. National Academies Press, Washington, DC. Available at: http://books.nap.edu/openbook.php? record_id=11970. Accessed March 27, 2019.

NTP (National Toxicology Program). (2016). *Systematic Literature Review on the Effects of Fluoride on Learning and Memory in Animal Studies*. *NTP Research Report 1*. National Toxicology Program, Research Triangle Park, NC.

Obbink, H. J., and Dalderup, L. M. (1963). Effects of thalidomide in the rat foetus. *Experientia* **19**, 645–646.

OECD. (2005). *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. Environmental Health and Safety Monograph Series on Testing and Assessment No. 34*. OECD Publishing, Paris.

OECD. (2018). *Test No. 414: Prenatal Development Toxicity Study, OECD Guidelines for the Testing of Chemicals*. Section 4. OECD Publishing, Paris. Available at: https://doi.org/10.1787/9789264070820-en. Last accessed June 9, 2019.

Okoli, C., and Schabram, K. (2010). A Guide to Conducting a Systematic LiteratureReview of Information Systems Research. *Sprouts: Working Papers on Information Systems*, **10**(26). http://sprouts.aisnet.org/10-26. Last accessed June 9, 2019.

Panzica-Kelly, J. M., Zhang, C. X., and Augustine-Rauch, K. A. (2015). Optimization and performance assessment of the chorion-off dechorinated zebrafish developmental toxicity assay. *Toxicol. Sci.* **146**, 127–134.

Piersma, A. H., Burgdorf, T., Louekari, K., Desprez, B., Taalman, R., Landsiedel, R., Barroso, J., Rogiers, V., Eskes, C., Oelgeschläger, 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.

Rooney, A. A., Boyles, A. L., Wolfe, M. S., Bucher, J. R., and Thayer, K. A. (2014). Systematic review and evidence integration for literature-based environmental health science assessments. *Environ. Health Perspect.* **122**, 711–718.

Rooney, A. A., Cooper, G. S., Jahnke, G. D., Lam, J., Morgan, R. L., Boyles, A. L., Ratcliffe, J. M., Kraft, A. D., Schünemann, H. J., Schwingl, P., *et al.* (2016). How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. *Environ. Int.* **92–93**, 617–629.

Rovida, C., Alépée, N., Api, A. M., Basketter, D. A., Bois, F. Y., Caloni, F., Corsini, E., Daneshian, M., Eskes, C., Ezendam, J., *et al.* (2015). Integrated testing strategies (ITS) for safety assessment. *ALTEX* **32**, 25–40.

Samuel, G. O., Hoffmann, S., Wright, R. A., Lalu, M. M., Patlewicz, G., Becker, R. A., DeGeorge, G. L., Fergusson, D., Hartung, T., Lewis, R. J., *et al.* (2016). Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: A scoping review. *Environ. Int.* **92–93**, 630–646.

Schumacher, H., Blake, D. A., Gurian, J. M., and Gillette, J. R. (1968). A comparison of the teratogenic activity of thalidomide in rabbits and rats. *J. Pharmacol. Exp. Ther.* **160**, 189–200.

Selderslaghs, I. W. T., 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.

Sena, E. S., Currie, G. L., McCann, S. K., Macleod, M. R., and Howells, D. W. *et al.* (2014). Systematic reviews and meta-analysis of preclinical studies: Why perform them and how to appraise them critically. *J. Cereb. Blood Flow Metab.* **34**, 737–742.

Sipes, N. S., Padilla, S., and Knudsen, T. B. (2011). Zebrafish: As an integrative model for twenty-first century toxicity testing. *Birth Defects Res. C Embryo Today* **93**, 256–267.

Solecki, R., Bergmann, B., Bürgin, H., Buschmann, J., Clark., R.,. Druga, A., van Duijnhoven, E. A. J., Duverger, M., Edwards, J., Freudenberger, H., *et al.* (2003). Harmonization of rat fetal external and visceral terminology and classification. Report of the fourth workshop on the terminology in developmental toxicology. *Reprod. Toxicol.* **17**, 625–637.

Spielmann, H., Hoffmann, S., Liebsch, M., Botham, P., Fentem, J. H., Eskes, C., Roguet, R., Cotovio, J., Cole, T., Worth, A., *et al.* (2007). The ECVAM international validation study on in vitro tests for acute skin irritation: Report on the validity of the EPISKIN and EpiDerm assays and on the skin integrity function test. *Altern. Lab. Anim.* **35**, 559–601.

Staples, R. E., and Holtkamp, D. E. (1963). Effects of parental thalidomide treatment on gestation and fetal development. *Exp. Mol. Pathol. Suppl.* **2**, 81–106.

Stephens, M. L., Betts, K., Beck, N. B., Cogliano, V., Dickersin, K., Fitzpatrick, S., Freeman, J., Gray, G., Hartung, T., McPartland, J., *et al.* (2016). The emergence of systematic review in toxicology. *Toxicol. Sci.* **152**, 10–16.

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.

Tihanyi, D. K., Szijarto, A., Fülöp, A., Denecke, B., Lurje, G., Neumann, U. P., Czigany, Z., and Tolba, R. (2019). Systematic review on characteristics and reporting quality of animal studies in liver regeneration triggered by portal vein occlusion and associating liver partition and portal vein ligation for staged hepatectomy: Adherence to the ARRIVE guidelines. *J. Surg. Res.* **235**, 578–590.

Tsaioun, K., Busquet, F., Flick, B., Hoffmann, S., Lalu, M., Stephens, M., de Vries, R., Witters, H., Wright, R., and Akun Ölmez, 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*. Available at: http://www.crd.york.ac.uk/PROSPERO/display_record.php? ID=CRD42018096120. Last accessed June 9, 2019.

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.

Wise, D., Beck, S., Beltrame, D., Beyer, B., Chahoud, I., Clark, R. L., Clark, R., Druga, A., Feuston, M., Guittin, P., *et al.* (1997). Terminology of developmental abnormalities in common laboratory mammals (Version 1). *Teratology* **55**, 249–292.

Woodruff, T. J., and Sutton, P. (2014). The navigation guide systematic review methodology: A rigorous and transparent method for translating environmental health science into better health outcomes. *Environ. Health Perspect.* **122**, 1007–1014.

Yauw, S. T. K., Wever, K. E., Hoesseini, A., Ritskes-Hoitinga, M., and van Goor, H. (2015). Systematic review of experimental studies on intestinal anastomosis. *Br. J. Surg.* **102**, 726–734.

Zhao, L., Zhen, C., Wu, Z., Hu, R., Zhou, C., and Guo, Q. (2010). General pharmacological properties, developmental toxicity, and analgesic activity of gambogic acid, a novel natural anticancer agent. *Drug Chem. Toxicol.* **33**, 88–96.