OXFORD
SOT | Society of Toxicology
academic.oup.com/toxsci

TOXICOLOGICAL SCIENCES, 188(2), 2022, 198–207

https://doi.org/10.1093/toxsci/kfac053
Advance Access Publication Date: 27 May 2022
Research Article

Exploring the Influence of Experimental Design on Toxicity Outcomes in Zebrafish Embryo Tests

Jui-Hua Hsieh ,*,[1] Mamta Behl,† Frederick Parham,* and Kristen Ryan*

*Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, USA and †Neurocrine Biosciences Inc., San Diego, California, 92130, USA

[1]To whom correspondence should be addressed. E-mail: jui-hua.hsieh@nih.gov.

ABSTRACT

Compound toxicity data obtained from independent zebrafish laboratories can vary vastly, complicating the use of zebrafish screening for regulatory decisions. Differences in the assay protocol parameters are the primary source of variability. We investigated this issue by utilizing data from the NTP DNT-DIVER database (https://doi.org/10.22427/NTP-DATA-002-00062-0001-0000-1, last accessed June 2, 2022), which consists of data from zebrafish developmental toxicity (devtox) and locomotor response (designated as “neurotox”) screens from 3 independent laboratories, using the same set of 87 compounds. The data were analyzed using the benchmark concentration (BMC) modeling approach, which estimates the concentration of interest based on a predetermined response threshold. We compared the BMC results from 3 laboratories (A, B, C) in 3 toxicity outcome categories: mortality, cumulative devtox, and neurotox, in terms of activity calls and potency values. We found that for devtox screening, laboratories with similar/same protocol parameters (B vs C) had an active call concordance as high as 86% with negligible potency difference. For neurotox screening, active call concordances between paired laboratories are lower than devtox screening (highest 68%). When protocols with different protocol parameters were compared, the concordance dropped, and the potency shift was on average about 3.8-fold for the cumulative devtox outcome and 5.8-fold for the neurotox outcome. The potential contributing protocol parameters for potency shift are listed or ranked. This study provides a quantitative assessment of the source of variability in zebrafish screening protocols and sets the groundwork for the ongoing Systematic Evaluation of the Application of Zebrafish in Toxicology effort at the National Toxicology Program.

Key words: zebrafish; developmental toxicity; developmental neurotoxicity; computational toxicology.

Zebrafish embryos are known to have advantages such as transparent and rapid development and high scalability, and thus, have been widely used as toxicity screening tools, especially for developmental toxicity (devtox) and the light/dark induced locomotor response (neurotox) screenings (Achenbach *et al.*, 2020; Bailey *et al.*, 2013; d’Amora and Giordani, 2018; He *et al.*, 2014; Sipes *et al.*, 2011). As a result of their wide use, these screening approaches have been adopted in various government-directed campaigns. Two public datasets have come from these screening efforts: the ToxCast dataset (Toxicity Forecasting) from the Environmental Protection Agency (EPA) (Judson *et al.*, 2010) and DNT-DIVER (Developmental NeuroToxicity Data Integration and Visualization Enabling Resource) from the Division of the National Toxicology Program (DNTP) (Data Release: Developmental NeuroToxicity Data Integration and Visualization Enabling Resource (DNT-DIVER), 2018; Behl *et al.*, 2019). There are some differences between the data available in these 2 datasets. In the ToxCast Phase I dataset, 306 compounds were screened in the zebrafish devtox assay by a laboratory in the EPA (Padilla *et al.*, 2012). In the ToxCast Phase II dataset, a library of 1060 compounds were screened using both the devtox and neurotox assay by a laboratory at Oregon State University (Thomas *et al.*, 2019; Truong *et al.*, 2014; Zhang *et al.*, 2017). The ToxCast Phase II dataset was also screened by the same EPA laboratory but only using the devtox assay (data available in U.S. EPA CompTox Chemicals Dashboard). On the other hand, in the DNT-DIVER dataset, although there are only 87 compounds, all were screened in both devtox and neurotox assays independently conducted by 3 laboratories with their own assay protocols and experimental designs. The screening data were harmonized and were analyzed using a data analysis pipeline (Hsieh *et al.*, 2019). The data meet the FAIR principles (findability, accessibility, interoperability, and reusability) and the results can be visualized on web applications particularly designed for DNT data (Behl *et al.*, 2019).

The screening results for the same set of compounds in DNT-DIVER can vary vastly. For example, a greater than 10-fold potency difference was observed in zebrafish larvae exposed to "carbaryl" between 2 devtox screens (Supplementary Figure 1). Because all the screening compounds were ensured to have acceptable compound quality by DNTP, the observed data difference is primarily attributed to the difference in the screening protocols and experimental designs. It is well recognized in the zebrafish community that there is a need to characterize the source of variability in the zebrafish embryonic screening protocols (Hamm *et al.*, 2019; Nishimura *et al.*, 2016; Ogungbemi *et al.*, 2019; Planchart *et al.*, 2016; Tal *et al.*, 2020). In Hamm *et al.*, information regarding various devtox protocols was gathered after interviewing zebrafish researchers. The authors placed the protocol parameters into 2 categories in terms of their degree of concern in affecting the results. The parameters of lesser concerns include source and strain of fish used, feed, water, embryo exposure conditions, physio-chemical properties, endpoints, and data collection. On the other hand, the dosing scenario (static vs static renewal, aka repeated exposure) and chorion status (with vs without) are specifically highlighted because of their higher potential to influence the devtox outcome. In Ogunbemi *et al.*, they also highlighted the importance of the exposure (duration, concentration) in zebrafish embryonic behavioral assays but with limited focus on the chorion status (Ogungbemi *et al.*, 2019).

Building on our previous work, we used the public DNT-DIVER dataset to investigate the influence of protocol parameters in zebrafish DT (devtox) and NT (neurotox) screening on compound toxicity outcomes. The influence was quantitatively assessed using a linear mixed effects model (LMM) and potential sources of variability are proposed. This study serves as the groundwork for the ongoing DNTP Systematic Evaluation of the Application of Zebrafish in Toxicology (SEAZIT) (Hamm *et al.*, 2019) and Organisation for Economic Co-operation and Development zebrafish light/dark induced locomotor response protocol harmonization efforts (Hessel *et al.*, 2021).

MATERIALS AND METHODS

Assay and compound data. Three data sources (Lab-A, B, C) used their in-house zebrafish DT and NT assays to screen compounds provided by NTP using blinded chemical codes. The tabulated comparison of the protocol parameters and assay designs across 3 data sources is provided in Tables 1 and 2 and Supplementary Table 1. In total, 87 unique compounds were tested in all 3 sources, with 4 of the compounds tested in duplicate. The tested compounds cover diverse chemical classes including pesticides, polycyclic aromatic hydrocarbons, flame retardants, drugs, and compounds considered negative in most toxicity assays (Behl *et al.*, 2019). The identifiers (CAS registry number and use category) of the 87 compounds and their related physicochemical properties (from EPA Chemical Dashboard) are available in the Supplementary File (excel spreadsheets: chem_list, and chem_physichem).

Zebrafish data in DNT-DIVER. The DNT-DIVER Data Release (https://doi.org/10.22427/NTP-DATA-002-00062-0001-0000-1) includes both the concentration-response data and the activity data. The activity data were derived using the benchmark concentration (BMC) modeling approach reported in our recent publication (Hsieh *et al.*, 2019). A detailed description of the data analysis process is provided in the Supplementary Methods, Supplementary Figure 2, as well as the R source code to generate activity data of zebrafish datasets in the DNT-DIVER Data Release.

The activity data from the DNT-DIVER Data Release include both binary activity call (active/inactive) and the potency values (ie, BMC, a type of point of departure for the toxicity). For the devtox assay, activity data from 2 endpoints are available, which are created based on reported incidences: "percent of mortality" (aka percent of total number of dead embryos) and "percent of affected" embryos (aka cumulative devtox; this is the percent of total number of dead or malformed embryos). For the neurotox assay, activity data from 2 types of endpoints are available. The first type is the quantity-type endpoints, where the response is the area-under-the-curve value from the recorded total distance moved per embryo at each individual light and dark time phase (Table 2). The second type is the similarity-type endpoints, where the response is the similarity of the movement pattern across the whole experiment time between the embryo treated with the study chemical and the embryo treated with vehicle control only. The activity data from the DNT-DIVER Data Release are provided at the compound level. The comparison of the BMC values of the duplicates in the zebrafish datasets is shown in the Supplementary Figure 6.

The normalized response data (for normalization, please refer to the Supplementary Methods) of the vehicle control across plates are pooled to derive the standard deviation (SD). The benchmark responses (BMR) per endpoint for the zebrafish datasets were provided in the Supplementary File (excel sheet: BMR_endpoints).

Activity calls. The toxicity endpoints were classified into 3 outcome categories: mortality, cumulative devtox, and neurotox (Hsieh *et al.*, 2019). If multiple time points were available for each endpoint, the closest matched time points across data sources were preserved. For example, because cumulative devtox endpoints (ie, malformations) were only recorded at 4 days post-fertilization (dpf) in Lab-A, the 4 dpf data were used to compare the 5 dpf data from Lab-B and Lab-C. There is only one endpoint in the mortality and cumulative devtox endpoint categories ("percent of mortality at 4/5 dpf" or "percent of affected at 4/5 dpf"). For the neurotox endpoint category, for Lab-A/Lab-C, 7 endpoints were used, including 4 quantity-type endpoints (endpoints for measuring the distance moved in the 2 light phases + endpoints for measuring the distance moved in the 2 dark phases) and 3 similarity-type endpoints (endpoints for measuring the movement pattern similarity between treated and untreated embryos using either Pearson's correlation, Spearman's correlation, and cosine similarity). For Lab-B, an additional 2 quantity-type endpoints were used because there was an extra light-dark cycle in the experimental design. When summarizing the activity for each category per compound, a compound was considered active if it was active in any of the endpoints, and the most potent BMC value was reported. The summarized data used in this study is available in the Supplementary File (excel sheet: all_bmc).

Table 1. Summary of the Protocol Parameters (Both Devtox and Neurotox Assays) From 3 Data Sources

	Lab-A	Lab-A	Lab-B	Lab-C
Assay type	Devtox	Neurotox	Devtox/neurotox	Devtox/neurotox
Fish strain	AB Tg(Cmlc2: copGFP)	AB wild-type	5D Tropical	5D Tropical
Embryo dechorionated (at time point)	No	No	Yes (4 hpf)	Yes (4 hpf)
Exposure volume	1000 µl	200 µl	100 µl	200 µl
Exposure at time point	3–5 and 48 hpf	72 hpf	6 hpf	6 hpf
Exposure scenario	Static renewal	Static	Static	Static
Effect measured at time point	96 hpf	120 hpf	120 hpf	120 hpf
Plate format	24-well	96-well	96-well	96-well
No. of embryos per well	5	1	1	1
Vehicle control (DMSO) concentration	0.50%–1% (max)	0.50%	0.64%	0.50%
Max tested concentration	Based on MTC	Based on MTC or LOAEL in DT	Fixed (67 µM)	Fixed (30 µM)
Number of concentrations per plate	5–8	5	7	5
Concentration spacing (log10)	0.26	0.3	0.3	0.48
Plate replicate	1	1	3	1
Number of replicate embryos per concentration	15	16	12	8

Abbreviations: MTC, maximal tolerated concentration; hpf, hour post fertilization; LOAEL, lowest-observed-adverse-effect level.

Table 2. Summary of the Experimental Design from 3 Data Sources in the Neurotox Assay

	Lab-A	Lab-B	Lab-C
Acclimation	5 min (L) -> 5 min (D)	3 min (L1) -> 3 min (D1)	5 min (L)
Summary of testing	2 L-D	3 L-D	2 L-D
Details of testing	10 min (L1) -> 10 min (D1) -> 10 min (L2) -> 10 min (D2)	3 min (L1) -> 3 min (D1) -> 3 min (L2) -> 3 min (D2) -> 3 min (L3) -> 3 min (D3)	5 min (L1) -> 5 min (D1) -> 5 min (L2) -> 15 min (D2)
Total testing time	40 min	18 min	30 min
Measurement	Total distance moved, velocity, duration, etc.[a]	Total distance moved	Total distance moved

Abbreviations: L, light; D, dark.

[a]Only total distance moved is used in the analysis.

Activity call concordance. Two concordance values were reported: the active concordance and the inactive concordance. The active concordance was calculated as the number of active concordant pairs divided by the sum of the number of discordant pairs and active concordant pairs. The inactive concordant rate was calculated as the number of inactive concordant pairs divided by the sum of the number of discordant pairs and inactive concordant pairs. An active or inactive concordant pair was defined as a compound active or inactive in all the data sources. To define a discordant pair, in addition to the activity, the highest tested concentration was also considered. The consideration was included because all 3 data sources used different testing ranges. For example, for a discordant pair, if a compound was active in source-a at a concentration that was out of the testing range of source-b, this compound might be active if source-b extended its testing range, thus this pair should not be included as "discordant" in the analysis. The mock example is shown in Supplementary Figure 3. The number of concordant/discordant pairs is provided in Supplementary File (excel sheet: devtox_bmc_3lab_compare, neurotox_bmc_3lab_compare, mortality_bmc_3lab_compare). The 95% confidence interval (CI) of the active/inactive concordance was also calculated using the percentile method based on 2000 bootstrap samples (R package *boot*) (Canty and Ripley, 2021).

Linear mixed effect model. The LMM was applied to understand the sources of BMC variation for each endpoint category across data sources, particularly the difference in the protocol parameters and the experimental designs. The model was fit using log10-transformed BMC values as the outcome variable. Compound IDs were set as random effects in the model and variables reflecting differences between laboratories were set as the fixed effects. In the LMM analyses, the *p*-values were estimated via *t* tests using the Satterthwaite approximations to degrees of freedom (R package, *lmerTest*) (Kuznetsova *et al.*, 2017) to estimate if the potency difference between classes is significant. The R package, *psycho* (Makowski, 2018), was used for calculating conditional R^2 (total explanatory power) and marginal R^2 (for fixed effect), and the bootstrap 95% CI of the beta estimate.

RESULTS

First, we tabulated the devtox and neurotox screening protocol of 3 labs based on parameters we captured (Tables 1 and 2 and Supplementary Table 1). The protocol parameters that are differential between labs are summarized in Table 3. Second, we investigated if there is difference in baseline responses (ie, without chemical treatment) between labs. Third, we investigated whether the results with chemical treatment are different among the labs in terms of the activity calls and potency values, either by concordance test or LMM analyses. If the results with chemical treatment are different among the labs, we investigated whether we could identify or rank the contributing protocol parameters.

Protocol Comparison

In Table 1, we provide the summary of protocol parameters from 3 data sources in a tabular form. Overall, of the parameters that we captured, the protocols used in Lab-B and Lab-C were similar to each other than to Lab-A, especially in the devtox assay. In the devtox assay, Lab-B and Lab-C used the same fish strain (5D Tropical), and the same exposure scenario (static). In addition, both measured the effect after a 5-day compound exposure, used a smaller exposure volume (96-well), and dechorionated the embryos prior to compound exposure. In contrast, each of these parameters was different for Lab-A. For the neurotox assay, the chosen protocol parameters and the assay design across data sources are more divided. Unlike Lab-A, Lab-B and Lab-C used the same fish strain (5D Tropical), dechorionated the embryos, and exposed the embryos at an earlier time point (6 hpf). However, for the neurotox assay design (Table 2), Lab-B is different from Lab-A and Lab-C. Lab-B used shorter light and dark segments in addition to a shorter total testing time and smaller exposure volume. The protocol parameters that are differential between labs are summarized in Table 3. The information in this table will be used in the LMM analyses described below.

Baseline Response Variation Comparison

The response variation from embryos treated only with vehicle control represents the baseline response variation, which was assessed using the SD metric (Figs. 1D–F and Supplementary Figs. 3C and 3D). In addition, we also compared the BMR because it captures the intrinsic response variation based on our definition, and represents minimum activity threshold (Hsieh *et al.*, 2019) (Figs. 1A–C and Supplementary Figs. 3A and 3B). For the SD and BMR metrics, the difference among 3 labs is below 10% for the mortality and cumulative devtox endpoints, but for the neurotox (similarity) endpoints, the difference is on average larger than 20%. The BMR values are around 25%–35% for the mortality and cumulative devtox endpoint category, but for neurotox(similarity) endpoints, the BMR values in Lab-B can be up to 45%–65%, indicating a larger change of responses (vs vehicle control) across concentrations is needed to be recognized as effects.

For the neurotox (similarity) endpoints, the movement pattern similarity value between embryos treated only with vehicle control per plate (ie, the denominator in the normalization) was plotted (Figure 1G and Supplementary Figs. 3E and 3F). The value ranges from 0 to 1, where 1 is the maximum value (a complete match of the movement pattern). Lab-B had relatively lower values and a wider data distribution compared with the other 2 data sources. The median value of the distribution is 0.28 (vs 0.75 for Lab-A and 0.64 for Lab-C). For all the 3 metrics, Lab-A ranked as the highest in almost all the endpoints. Overall, the difference between 3 data sources was larger for the neurotox endpoint category. The difference in the baseline response result from the difference in the protocols and the assay design and can affect the response variation from the embryos treated with compounds.

Activity Call and Potency Comparison

The concordance of the active call and inactive call are both considered important. Therefore, both the active concordance value and inactive concordance value are reported. Comparing across all 3 categories (mortality, cumulative devtox, neurotox), the 3-lab concordance is lower than the concordance between Lab-B and Lab-C, indicating the distinctness of Lab-A results. The best 3-lab active concordance was in the cumulative devtox endpoint category, reaching 72.5%, with CI: [58.1%–85%]. Lab-B and Lab-C concordance is highest in all categories, with the highest concordance seen in the cumulative devtox category (90% [80%–97.7%], for inactive concordance) and the lowest concordance seen in the neurotox endpoint category (67.9% [50%–84.6%], for the active concordance) (Figure 2). The concordance between other pairs (A <-> C, A <-> B) is lower, with values of approximately 54% in both mortality and neurotox endpoint category for active concordance but reaching up to approximately 70% for the cumulative devtox endpoint category. Inactive concordance for labs (A <-> C, A <-> B) ranged from approximately 60% to 70%.

The potency differences of matched actives between pairs (A <-> B, A <-> C, B <-> C) are plotted for each endpoint category (Figure 3). The compounds that show more than 10-fold difference are labeled. The BMC difference in the mortality category is smaller than the difference in the cumulative devtox and neurotox categories. For cumulative devtox and neurotox categories, the BMC values from Lab-A tend to be more potent than Lab-B and Lab-C; the BMC values from Lab-B and Lab-C are more similar to each other. For example, lindane, heptachlor, and dieldrin were found to be at least 10-fold more potent in Lab-A than in Lab-B and C in the neurotox endpoint category. In the cumulative devtox endpoint category, 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD), dibenz(a, h)anthracene, and dieldrin were found to be at least 10-fold more potent in Lab-A than in Lab-B and C. The analyses demonstrate that there are differences seen in the activity call and potency between data sources. The activity outcome and potency difference can be related to the difference in protocols and assay designs.

Data Sources Affect the BMC Outcome

The BMC outcome may be related to the data source (ie, laboratory). Therefore, we applied a statistical approach, LMM, to investigate whether the data sources effect on the BMC outcome is significant. The LMM is similar to the linear regression approach but is applied when the data are nonindependent, in this case, the BMC data from each chemical. In the setting of this LMM, the fixed effect is the data source (Lab-A, B, C), the random effect is the compound ID, and the outcome is the BMC value for each endpoint category.

The BMC difference observed in the cumulative devtox and the neurotox endpoint category is significant between Lab-A (reference) and Lab-B, Lab-C ($p < .001$ in Table 4). On average, the BMC values from Lab-B and Lab-C are significantly less potent (by 0.59 [log10 unit], 3.9-fold, Table 4) than Lab-A for cumulative devtox. Similarly, for the neurotox endpoint category, the BMC values from Lab-B and Lab-C are significantly less potent than Lab-A for about 5.8-fold and 3.4-fold (0.76 and 0.53 log10 unit, respectively, Table 4). The potency difference between Lab-B and Lab-C is not significant by the LMM approach ($p = .99$ [cumulative devtox], 0.83 [mortality], and 0.21 [neurotox], data not shown).

In terms of percent of data variance explained, it is not surprising to see that the random effect (ie, chemical) dominates the variance, but fixed effect (ie, laboratory) also contributes a significant amount in the neurotox category (19.58%) and cumulative devtox category (8.11%), but not much in the mortality category (1%) (Figure 4A). The summary statistics table is available in Table 4 and the complete statistics output is available in Supplementary Tables 2 and 3. The analyses demonstrate that

Table 3. The Differential Protocol Parameters Included in the Surrogate Factors for LMM Analyses

	DT_Factor1	Fish Strain	Dechorionation	Exposure Volume	Exposure Scenario	Measure Effect at Time
Lab-A	Reference	AB Tg (Cmlc2: copGFP)	No	Larger (1000)	Static renewal	Earlier (96 hpf)
Lab-B	Comparison	5D Tropical	Yes	Smaller (100/200)	Static	Later (120 hpf)
Lab-C						
	NT_Factor1	Fish strain	Dechorionation	Exposure at time		
Lab-A	Reference	AB wild-type	No	Later (72 hpf)		
Lab-B	Comparison	5D Tropical	Yes	Earlier (6 hpf)		
Lab-C						
	NT_Factor2	Exposure volume	No. of L/D cycles	Duration of testing		
Lab-A Lab-C	Reference	Larger (200)	2 L/D	Longer (40/30 min)		
Lab-B	Comparison	Smaller (100)	3 L/D	Shorter (18 min)		

Abbreviations: DT, devtox; NT, neurotox; hpf, hour post fertilization.

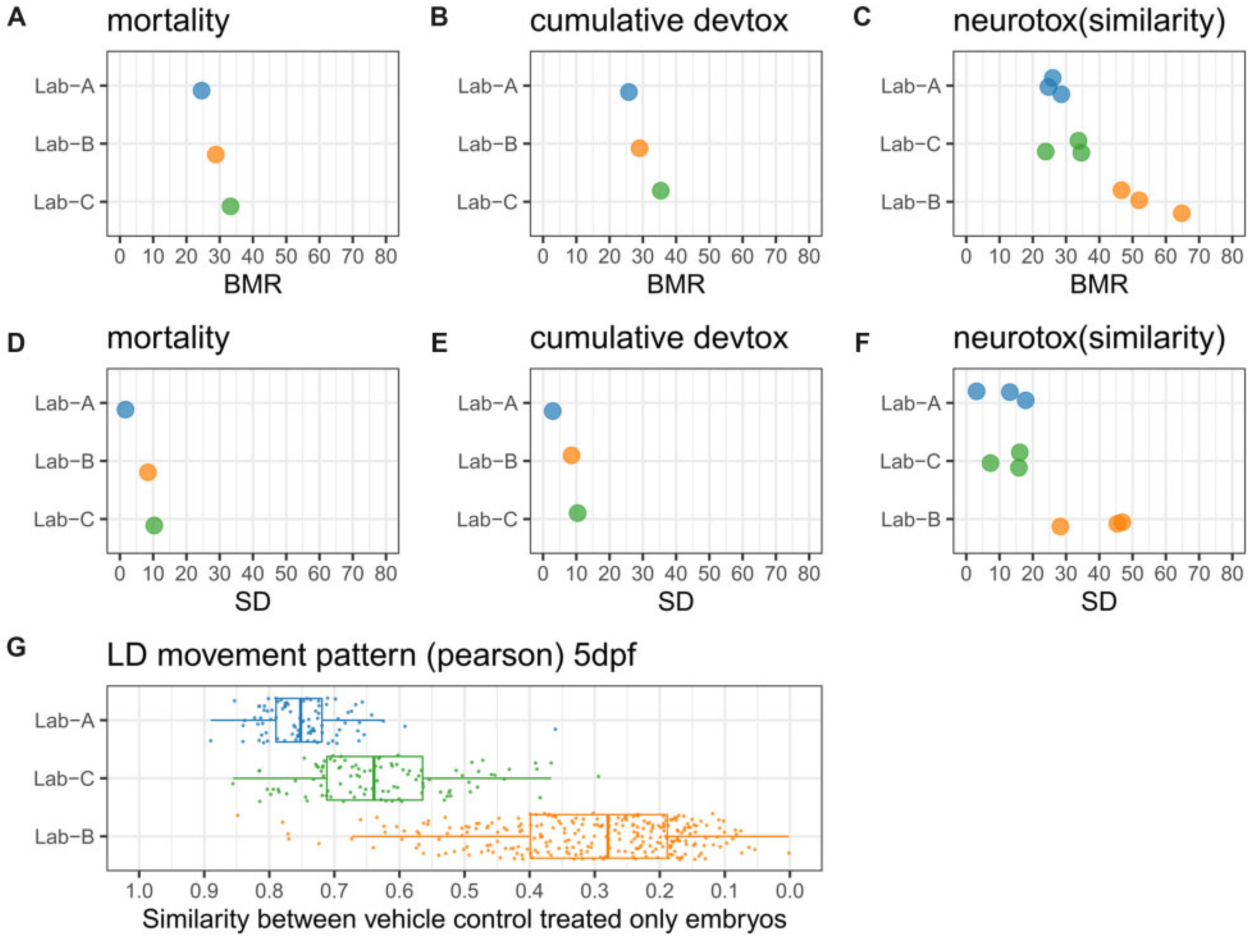

Figure 1. Metrics to evaluate the response variation in wells treated only with vehicle control. The laboratory (Lab-A; Lab-B; Lab-C) with the lowest noise value/highest similarity in the evaluated metrics was sorted to the top: Lab-A is consistently to be the best in all evaluated metrics. A–C, Benchmark response (BMR) in 3 endpoint categories (mortality, cumulative devtox, and neurotox [similarity]). Each dot represents the BMR value (%) from an endpoint; 3 endpoints are available for neurotox(similarity) endpoints. D–F, Standard deviation (SD) values (%) in 3 endpoint categories. The SD values are calculated using 91/273/102 (Lab-A/Lab-B/Lab-C) responses across plates for mortality and cumulative devtox outcome category and 1432/2954/744 (Lab-A/Lab-B/Lab-C) responses across plates for neurotox outcome category. G, Median similarity of movement pattern between pairs of embryos when using Pearson's correlation to evaluate the similarity of Light-dark (LD) movement pattern on 5 day-post- fertilization(dpf). Each dot represents the value from a plate. Boxplot was used to show the data distribution. The variability (represented as SD) is 0.07, 0.15, 0.11, for Lab-A, Lab-B, and Lab-C, respectively. The comparison of the remaining neurotox endpoints is available in the Supplementary Figure 3.

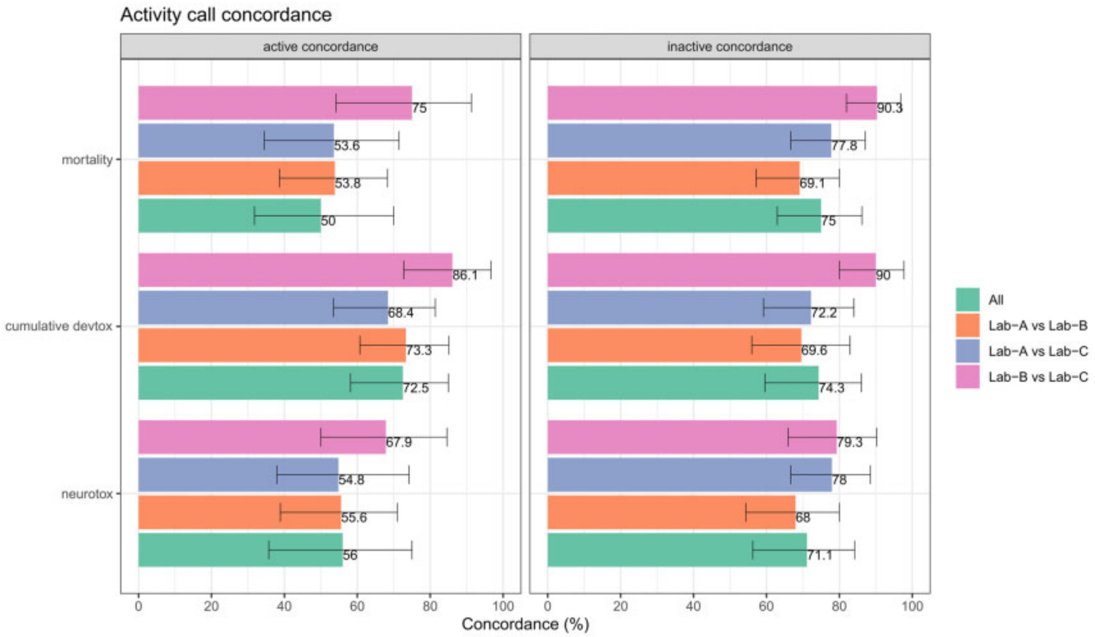

Figure 2. The activity call concordance with confidence interval (CI) as the error bars in endpoints between 3 laboratories. The active/inactive concordance in endpoints between 4 types of comparisons: all (all 3 laboratories), Lab-A versus Lab-B, Lab-B versus Lab-C, and Lab-A versus Lab-C.

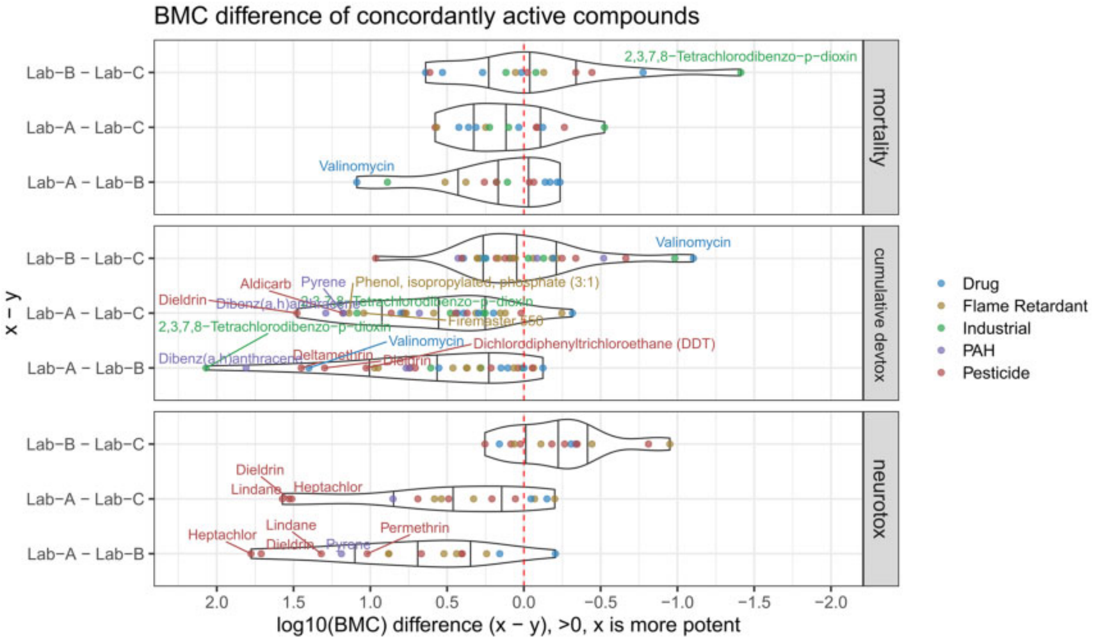

Figure 3. The BMC difference of the concordantly active compounds between pairs of laboratories in 3 categories of endpoints (mortality, cumulative devtox, and neurotox). Each dot represents a compound and is colored based on its category (eg, drug). The violin plot is used to show the distribution of the dots. The 25th, 50th, and 75th percentiles of the distribution are shown as black lines in the violin. The red dashed line is equivalent to 0, meaning there is no potency difference. Compounds that are 10-fold more potent/less potent are labeled. The names in the cumulative devtox panel: 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD), dieldrin, aldicarb, dibenz(a, h)anthracene, pyrene, valinomycin, firemaster 550, dichlorodiphenyltrichloroethane (DDT), phenol, isopropylated, phosphate (3:1), and deltamethrin. A color version of this figure appears in the online version of this article.

data sources affect the BMC outcome in both cumulative devtox and neurotox endpoint categories, but not in the mortality endpoint category.

Protocols Affect the BMC Outcome

In the following LMM analyses, we will focus only on cumulative devtox and neurotox categories to disentangle the underlying protocol parameters and assay designs that might contribute to the BMC difference seen between data sources because mortality endpoint is not affected by the data source in this dataset shown in the previous LMM analysis. In the next series of LMM, the random effect is still the compound ID but the fixed effects are various surrogate factors that reflect the difference in protocols and experimental designs

Table 4. Potency Difference (Log10 Unit) Obtained From the LMM Analyses

Mortality					
Fixed effect	Data source				
Factor	Data source				
Lab-A	Ref				
Lab-B	[]−0.19				
Lab-C	[]−0.13				
Cumulative devtox					
Fixed effect	Data source	DT_Factor1			
Factor	Data source	DT_Factor1			
Lab-A	Ref	Ref			
Lab-B	[***]−0.59	[***]−0.59			
Lab-C	[***]−0.58				
Neurotox					
Fixed effect	Data source	NT_Factor1	NT_Factor2	NT_Factor1 + NT_Factor2	
Factor	Data source	NT_Factor1	NT_Factor2	NT_Factor1	NT_Factor2
Lab-A	Ref	Ref	Ref	Ref	Ref
Lab-B	[***]−0.76	[***]−0.64	[**]−0.5	[***]−0.53	[+]−0.23
Lab-C	[***]−0.53		Ref		Ref

[significance level (*p*-value): ***, 0–.001; **, 0.001–0.01; +, 0.05–0.1]; ref: reference, Lab-A was used as the reference for the ease of comparison (because the Lab-A results are on average most potent); NT: neurotox; DT: devtox.

between laboratories. The summarized surrogate factors are in Table 3.

For LMM of the cumulative devtox endpoint category, the DT_Factor no. 1 can capture all the variance previously seen in the data source (Figure 4B). The BMC values from the data sources (Lab-B and Lab-C) with DT factor no. 1, which is the composite effect of same fish strain (5D Tropical), dechorionated embryos, static exposure scenario, smaller exposure volume, and effect measured at later time point (5-day), are less potent than Lab-A by about 3.9-fold (0.59 log10 unit). A similar analysis was conducted on the neurotox endpoint category but with 2 NT factors used either separately or jointly. The results are presented in Figure 4C. The BMC values from data sources (Lab-B and Lab-C) with NT factor no. 1, which is the composite effect of same fish strain (5D Tropical), dechorionated embryos, and earlier exposure time (6 hpf), are less potent than Lab-A for about 3.4-fold (0.53 log10 unit). The BMC values from Lab-B with NT factor no. 2, which is the composite effect of smaller exposure volume and shorter behavior testing time, are less potent Lab-A and Lab-C for about 1.7-fold (0.23 log10 unit). The LMM using both NT factors explains the same amount of the variance as the model using the data sources as the factor, because the data source factor are linear combinations of the 2 NT factors. In the model using the 2 NT factors, the NT_Factor1 has lower *p*-value than the NT_Factor2.

DISCUSSION

Several recent efforts have highlighted the variability that currently exists among different laboratories with respect to animal husbandry and protocols for chemical toxicity screening using zebrafish embryos including differences in the strain of fish used, timing and frequency of exposure, status of the chorion, exposure apparatus, endpoints measured, and scoring of phenotypic alterations (Hamm *et al.*, 2019; Ogungbemi *et al.*, 2019; Planchart *et al.*, 2016). Some efforts to evaluate different zebrafish methods for toxicity screening have recommended harmonized devtox protocols (Ball *et al.*, 2014; Beekhuijzen *et al.*, 2015; Gustafson *et al.*, 2012). However, other ongoing efforts such as SEAZIT (https://ntp.niehs.nih.gov/go/seazit) which was designed to address the need for standardized and validated devtox protocols suggest that it is impractical to force researchers to use a single protocol and hence is working toward the identification of key factors that may contribute toward variability in outcomes for developmental toxicity (Hamm *et al.*, 2019).

Taking these efforts into consideration, in this study, we took the advantage of a publicly available dataset (ie, DNT-DIVER interlaboratory data) with developmental toxicants and developmental neurotoxicants to investigate how underlying experimental design may influence toxicity outcomes. We found that independent laboratories with similar devtox assay protocols have both high active call (approximately 86%) and inactive call (approximately 90%) concordance with negligible potency difference in the cumulative devtox outcome category, indicating the assay robustness for toxicity screening. This finding is comparable with the finding made by the zebrafish devtox assay consortium formed within biopharmaceutical companies (Ball *et al.*, 2014; Gustafson *et al.*, 2012), yet their focus is more on predictivity of the assay to teratogenicity. Another study (Busquet *et al.*, 2014) has a similar focus to our interest but was done on the OECD zebrafish embryo acute toxicity test (ZFET), which is more comparable with the mortality endpoints in our analysis. In the paper, it was demonstrated that ZFET shows good intra- and interlaboratory reproducibility when using the harmonized protocol. We also found that data sources do not affect the BMC outcome of mortality in our dataset.

Our assessments have revealed that when devtox assay protocols are different, the concordance drops and on average, the potency shift is around 3.8-fold. The composite effect of several protocol parameters (use of 5D Tropical strain, static exposure with smaller amount of exposure volume, de-chorionation, later-assessed time) can potentially contribute to the decrease of potency. However, we cannot disentangle the effects due to the correlated data structure. The chorion-on versus chorion-off and static versus repeated exposure are highlighted because both are considered to have higher potential to influence devtox outcomes in the publication (Hamm *et al.*, 2019). The chorion-off

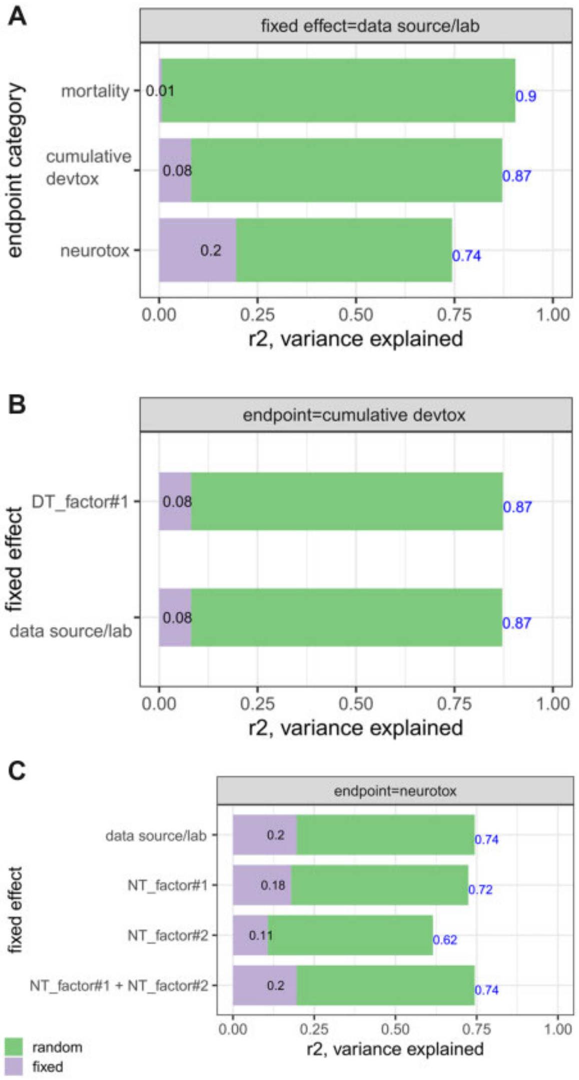

Figure 4. The fraction of variance explained by fixed/random effects in the LMM analyses. A, When fixed effect is the data source on 3 endpoint categories. B, For cumulative devtox and C, neurotox endpoint category, the fixed effect is one or a combination of the surrogate factors (DT_Factor1, NT_Factor1, and NT_Factor2) derived based on Table 3.

procedure in the devtox assay is suggested to have better sensitivity mammalian teratogens (Panzica-Kelly *et al.*, 2015). In another paper (Wilson *et al.*, 2020), the authors demonstrated the dosing scenario affects the most on activity potency by systematically altered the test conditions for 8-compound screening.

Additionally, our analysis was completed using the totality of malformations identified after compound exposure. Individual malformations may be differentially influenced by changes in protocol parameters and a more thorough evaluation could impact our understanding of the concordance. Some of DNTP's ongoing SEAZIT activities are specifically designed to address this research question. For example, a SEAZIT Interlaboratory Study was designed to determine the individual and/or synergistic effect of chorion removal and exposure media renewal by systematically varying these factors. The individual malformations in the context of ontology will be considered when evaluating the toxicity concordance. Also, individual chemicals can have different absorption, distribution, metabolism, and excretion (ADME) properties, thus react differently to the change of the protocol parameters, particularly the chorion-on/off condition. It can be important to measure the internal concentration of chemicals within the embryos to understand outcome difference between conditions (Quevedo *et al.*, 2019).

Assessment of larval behavior, for an indication of chemical-induced of neurotoxicity, can be measured in several approaches. These approaches include the L/D locomotor response test (in our analysis), the spontaneous tail coiling test, and the photomotor response test (Ogungbemi *et al.*, 2019). The latter 2 tests also have been used in the toxicity screening (Reif *et al.*, 2016; Vliet *et al.*, 2017) but may lack the ability to discriminate modes of action for compounds interfering with neurotransmission (Vliet *et al.*, 2017). On the other hand, the induced hypo- (neurotoxic) and hyper (neuroactive) response in the L/D locomotor test can be useful to link with mode of action of the toxicants (Ellis *et al.*, 2012; Kokel *et al.*, 2010). However, it is also shown that the neuroactive/neurotoxic outcome can differ when using different experimental conditions (Ogungbemi *et al.*, 2019). Knowing this, our focus in this study is to investigate if we can get concordant toxicity outcome based on the most sensitive activity in multiple endpoints. This mindset is toward using this assay in the Integrated Approaches for Testing and Assessment for DNT battery screening (H. Hogberg *et al.*, Organophosphorus flame retardants, a case study on the use of IATA for DNT to prioritize a class of compounds., submitted).

For zebrafish neurotox screens evaluated in this study, active call concordances between paired laboratories are lower than devtox screens (highest in neurotox: 68% vs 86% in cumulative devtox category). Further protocol harmonization may help to increase the concordance, which is the ongoing work in OECD Zebrafish Expert group (Hessel *et al.*, 2021). Based on the current dataset, the potency shift on average can be up to 5.7-fold. And we found that the composite effect of several protocol parameters can contribute to decrease of potency. The protocol parameters that have higher contribution to the decrease of potency (approximately 3.4 fold) include use of 5D Tropical strain, dechorionated embryos, and earlier exposure time (or longer exposure duration). The fish strain (exposure time) parameter was also identified to be a lower/higher risk of bias factor in Ogungbemi *et al.* but de-chorionation is not highlighted. The protocol parameters that have lower contribution to the decrease of potency (approximately 1.7 fold) include smaller exposure volumes and shorter behavior testing time. The small exposure volume can affect well concentration and the duration of experimental testing time is shown to affect the behavior (MacPhail *et al.*, 2009).

Our study used a zebrafish public dataset with the same chemical source for analysis, providing both quantitative and qualitative interpretation on how protocol parameters could affect the toxicity outcome. Some limitations of the analyses are noted here. First, we did not factor the plate replicate number (ie, Lab-B did triplicate plate testing, whereas Lab-A/Lab-C did single plate testing) into the analyses and we adopted a plate-centric activity data aggregation strategy for Lab-B (see Supplementary Method) instead of pooling the data from triplicate testing. The procedure can introduce bias to the data analysis. In Supplementary Figure 5, we compared the BMC difference from these 2 strategies. The BMC difference is negligible for the mortality and cumulative devtox outcome category, but the BMC results from the pooled data are about 1.04 (median)/1.12 (mean)-fold more potent than the plate-centric strategy. Also, some outliers (dieldrin, pyrene, dibenz(a, h)anthracene, DDT) are highlighted, where the BMC of dieldrin is approximately 2.5-fold more potent when pooling the data. Second, we did not consider the BMC variation between the duplicate testing of the 4 blinded substances. In DNT-DIVER, all the activity data were summarized, by pooling the data, to the chemical level (ie, CAS registry number) to simplify the comparison. In Supplementary Figure 6, we present BMC variation between the duplicates tested in the zebrafish datasets. The median/mean of SD of log10(BMC) between the duplicate testing are 0.35/0.41 (2.24/2.57-fold), 0.09/0.09 (1.23-fold), 0.07/0.13 (1.17/1.34-fold), for mortality, cumulative devtox, and neurotox outcome categories, respectively. Third, the directionality of responses in the neurotox assay may be useful to inform mode of action. However, in our analyses, we did not consider it. The decision is intentional because our goal in this study is to evaluate this sets of assays as one module in the DNT screening battery, and thus, focusing on the activity call concordance and the amount of potency shift. In addition, one type of neurotox endpoints we used (similarity-type) cannot capture directionality of responses because they only compare the movement pattern to the vehicle control, and regarding the biphasic behavior response (eg, increased movement at lower concentration then decreased movement at higher concentration), special type of parametric models may be needed to fit the data.

We think this study sets the groundwork for the ongoing NTP SEAZIT project as well as the OECD Zebrafish Expert Group effort on the global harmonization of zebrafish protocols related to locomotor response test. As we work toward including zebrafish as a complementary model in drug development and toxicity testing, it is important to understand critical elements that may be most influential in interpreting outcomes and the limitations of the alternative animal model.

SUPPLEMENTARY DATA

Supplementary data are available at *Toxicological Sciences* online.

ACKNOWLEDGMENTS

We thank Drs Nisha Sipes (U.S. EPA) and Christopher McPherson (DNTP/NIESH) for their insightful comments.

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

Achenbach, J. C., Leggiadro, C., Sperker, S. A., Woodland, C., and Ellis, L. D. (2020). Comparison of the zebrafish embryo toxicity assay and the general and behavioral embryo toxicity assay as new approach methods for chemical screening. *Toxics* **8**, 126.

d'Amora, M., and Giordani, S. (2018). The utility of zebrafish as a model for screening developmental neurotoxicity. *Front. Neurosci.* **12**, 976.

Bailey, J., Oliveri, A., and Levin, E. D. (2013). Zebrafish model systems for developmental neurobehavioral toxicology. *Birth Defects Res. C Embryo Today* **99**, 14–23.

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.

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.

Behl, M., Ryan, K., Hsieh, J.-H., Parham, F., Shapiro, A. J., Collins, B. J., Sipes, N. S., Birnbaum, L. S., Bucher, J. R., Foster, P. M. D., *et al.* (2019). Screening for developmental neurotoxicity at the national toxicology program: The future is here. *Toxicol. Sci.* **167**, 6–14.

Busquet, F., Strecker, R., Rawlings, J. M., Belanger, S. E., Braunbeck, T., Carr, G. J., Cenijn, P., Fochtman, P., Gourmelon, A., Hübler, N., *et al.* (2014). OECD validation study to assess intra- and inter-laboratory reproducibility of the zebrafish embryo toxicity test for acute aquatic toxicity testing. *Regul. Toxicol. Pharmacol.* **69**, 496–511.

Canty, A., and Ripley, B. (2021) boot: Bootstrap R (S-Plus) functions. R package version 1.3-28.

Data Release: Developmental NeuroToxicity Data Integration and Visualization Enabling Resource (DNT-DIVER) (2018). Division of National Toxicology Program, Research Triangle Park, NC.

Ellis, L. D., Seibert, J., and Soanes, K. H. (2012). Distinct models of induced hyperactivity in zebrafish larvae. *Brain Res.* **1449**, 46–59.

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., *et al.* (2019). Characterizing sources of variability in zebrafish embryo screening protocols. *ALTEX* **36**, 103–120.

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.

Hessel, E., Shafer, T., Padilla, S., Hill, B., Truong, L., Tanguay, R., Hsieh, J., Ryan, K., Behl, M., Ellis, L., *et al.* (2021). An Inter-laboratory Case Study to Determine the Added Value of the Zebrafish Light-Dark Transition Test to Predict Developmental Neurotoxicity: Report from OECD DNT Expert Group, In: The Toxicologist: Supplement to Toxicology Sciences, 180 (1), *Society of Toxicology* (Abstract no. 1203).

Hsieh, J.-H., Ryan, K., Sedykh, A., Lin, J.-A., Shapiro, A. J., Parham, F., and Behl, M. (2019). Application of benchmark concentration (BMC) analysis on zebrafish data: A new perspective for quantifying Toxicity in alternative animal models. *Toxicol. Sci.* **167**, 92–104.

Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., Reif, D. M., Rotroff, D. M., Shah, I., Richard, A. M., *et al.* (2010). In vitro screening of environmental chemicals for targeted testing prioritization: The ToxCast project. *Environ. Health Perspect.* **118**, 485–492.

Kokel, D., Bryan, J., Laggner, C., White, R., Cheung, C. Y. J., Mateus, R., Healey, D., Kim, S., Werdich, A. A., Haggarty, S. J., *et al.* (2010). Rapid behavior-based identification of neuroactive small molecules in the zebrafish. *Nat. Chem. Biol.* **6**, 231–237.

Kuznetsova, A., *et al.* (2017). lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26.

MacPhail, R. C., Brooks, J., Hunter, D. L., Padnos, B., Irons, T. D., and Padilla, S. (2009). Locomotion in larval zebrafish:

Influence of time of day, lighting and ethanol. *NeuroToxicology* **30**, 52–58.

Makowski, D. (2018). The psycho package: An efficient and publishing-oriented workflow for psychological science. *J. Open Source Softw*. **3**, 470.

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.

Ogungbemi, A., *et al*. (2019). Hypo- or hyperactivity of zebrafish embryos provoked by neuroactive substances: A review on how experimental parameters impact the predictability of behavior changes. *Environ. Sci. Eur*. **31**, 88.

Padilla, S., Corum, D., Padnos, B., Hunter, D. L., Beam, A., Houck, K. A., Sipes, N., Kleinstreuer, N., Knudsen, T., Dix, D. J., *et al*. (2012). Zebrafish developmental screening of the ToxCast^{TM} Phase I chemical library. *Reprod. Toxicol*. **33**, 174–187.

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.

Planchart, A., Mattingly, C. J., Allen, D., Ceger, P., Casey, W., Hinton, D., Kanungo, J., Kullman, S. W., Tal, T., Bondesson, M., *et al*. (2016). Advancing toxicology research using in vivo high throughput toxicology with small fish models. *ALTEX* **33**, 435–452.

Quevedo, C., Behl, M., Ryan, K., Paules, R. S., Alday, A., Muriana, A., and Alzualde, A. (2019). Detection and prioritization of developmentally neurotoxic and/or neurotoxic compounds using zebrafish. *Toxicol. Sci*. **168**, 225–240.

Reif, D. M., Truong, L., Mandrell, D., Marvel, S., Zhang, G., and Tanguay, R. L. (2016). High-throughput characterization of chemical-associated embryonic behavioral changes predicts teratogenic outcomes. *Arch. Toxicol*. **90**, 1459–1470.

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 Rev*. **93**, 256–267.

Tal, T., Yaghoobi, B., and Lein, P. J. (2020). Translational toxicology in zebrafish. *Curr. Opin. Toxicol*. **23–24**, 56–66.

Thomas, D. G., *et al*. (2019). Time-dependent behavioral data from zebrafish reveals novel signatures of chemical toxicity using point of departure analysis. *Comput. Toxicol*. **9**, 50–60.

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.

Vliet, S. M., Ho, T. C., and Volz, D. C. (2017). Behavioral screening of the LOPA. C1280 library in zebrafish embryos. *Toxicol. Appl. Pharmacol*. **329**, 241–248.

Wilson, L. B., Truong, L., Simonich, M. T., and Tanguay, R. L. (2020). Systematic assessment of exposure variations on observed bioactivity in zebrafish chemical screening. *Toxics* **8**, 87.

Zhang, G., Truong, L., Tanguay, R. L., and Reif, D. M. (2017). A new statistical approach to characterize chemical-elicited behavioral effects in high-throughput studies using zebrafish. *PLoS One* **12**, e0169408.