Application of Benchmark Concentration (BMC) Analysis on Zebrafish Data: A New Perspective for Quantifying Toxicity in Alternative Animal Models

Jui-Hua Hsieh,^{*,1} Kristen Ryan,[†] Alexander Sedykh,[‡] Ja-An Lin,[§]
Andrew J. Shapiro,[†] Frederick Parham,[†] and Mamta Behl[†]

^{*}Kelly Government Solutions, Durham, North Carolina, 27709, USA; [†]Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, 27709, USA; [‡]Sciome, Durham, North Carolina, 27709, USA; and [§]Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27516, USA

¹To whom correspondence should be addressed. Jui-Hua Hsieh, Kelly Government Solutions, 530 Davis Dr., Bldg: Keystone Building Rm: 2019, Durham, NC 27709. E-mail: jui-hua.hsieh@nih.gov.

ABSTRACT

Over the past decade, the zebrafish is increasingly being used as a model to screen for chemical-mediated toxicities including developmental toxicity (DT) and neurotoxicity (NT). One of the major challenges is lack of harmonization in data analysis approaches, thereby posing difficulty in comparing findings across laboratories. To address this, we sought to establish a unified data analysis strategy for both DT and NT data, by adopting the benchmark concentration (BMC) analysis. There are two critical aspects in the BMC analysis: having a toxicity endpoint amenable for BMC and selecting a proper benchmark response (BMR) for the endpoint. For the former, in addition to the typical endpoints in NT assay (eg, hyper/hypo-response quantified by distance moved), we also used endpoints that assess the differences in movement patterns between chemical-treated embryos and control embryos. For the latter, we standardized the selection of BMR, which is analogous to minimum activity threshold, based on intrinsic response variations in the endpoint. When comparing our BMC results with a traditionally used LOAEL method (lowest-observed-adverse-effect level), we found high active compound concordance (100% for DT vs 74% for NT); generally, the BMC was more sensitive than LOAEL (no. of BMC more sensitive/no. of concordant active compounds, 43/50 for DT vs 16/26 for NT). Using the BMC with standardized toxicity endpoints and an appropriate BMR, we may now have a unified data-analysis approach to comparing results across different zebrafish datasets, for a better understanding of strengths and challenges when using the zebrafish as a screening tool.

Key words: Benchmark concentration, Zebrafish, Alternative animal models, Developmental toxicity screening, Neurotoxicity screening.

This article is published as part of the NTP Neurotoxicology Screening Strategies Initiative.

The zebrafish (*Danio rerio*), is a well-recognized model for toxicology research due to its relatively low cost, metabolic competence, and high degree of homology to mammalian systems. (Hill *et al.*, 2005; Horzmann and Freeman, 2018; Howe *et al.*, 2013; Otte *et al.*, 2017). Among the research areas, zebrafish have been used to investigate chemical-mediated developmental toxicity (DT) and neurotoxicity (NT) (Bailey *et al.*, 2013; He *et al.*, 2014; McCollum *et al.*, 2011), which are found to have close

resemblance to corresponding toxicities in mammals (Sipes *et al.*, 2011). In addition, the zebrafish can be applied in medium to high-throughput toxicity screening platforms (Padilla *et al.*, 2012; Truong *et al.*, 2014; Zhang *et al.*, 2017). Therefore, the zebrafish is proposed as an alternative animal model for twenty-first century toxicity testing. However, a broader adoption of the zebrafish model as a toxicity screening tool in the regulatory context has not yet been implemented due to lack of

standardized documentation of experimental protocol designs, lack of understanding of toxicokinetics of chemicals in zebrafish, and inconsistency in data analysis strategies to classify chemical activity outcomes. The National Toxicology Program (NTP) and other global groups are working toward addressing some of these specific issues (Beekhuijzen et al., 2015; Gustafson et al., 2012; SEAZIT, 2018); herein, we focus our efforts on providing a strategy toward consistent processing and analysis of data obtained from zebrafish DT and NT assays.

Data analysis methods for zebrafish DT and NT toxicity endpoints are under active development (Jarema et al., 2015; Padilla et al., 2012; Truong et al., 2014; Zhang et al., 2017). As a result, current data analysis strategies are not only different between domains of DT and NT but are different across laboratories even within each domain. For the DT data, both the EC_{50}/LC_{50} (concentration at half maximal effect/lethal concentration) and LOAEL (lowest-observed-adverse-effect level) strategies are used; for NT data, the LOAEL strategy is primarily used. The LOAEL value may or may not be derived using statistical tests, and if statistical tests are applied, they could be applied on different toxicity endpoints (eg, percentage of affected embryos [Padilla et al., 2012] vs incidences at each individual malformation sites [Truong et al., 2014]). Therefore, it is challenging to compare activity outcomes across laboratories and outcome interpretation is complex. In terms of metric selection, the EC_{50} is not optimal for comparing activity of chemicals which elicit varying maximal effect. The LOAEL approach also has some shortcomings including (1) lack of consideration of concentration-response trend, (2) limiting LOAEL options to tested concentrations, (3) inability to provide uncertainty factors for the reported potency (Benchmark dose modeling – Introduction, 2018). Hence, both metrics are not optimal for comparing a diverse set of chemicals with varying concentration-response profiles tested using varied concentrations from different laboratories.

To address the above-mentioned limitations, we adopted the benchmark dose (BMD) analysis approach (or benchmark concentration [BMC] for zebrafish data). The BMD approach utilizes a user-defined benchmark response (BMR) relative to the background response to derive a BMC, and hence, enables an objective comparison of results across laboratories. The BMD was originally developed for modeling *in vivo* animal data for quantitative risk assessment (eg, Crump, 1984; U.S. EPA, 2015), but the NTP has been implementing this approach/concept to *in vitro* and zebrafish DT data (Behl et al., 2015; Hsieh et al., 2015; Ryan et al., 2016). In this study, we expanded the application of the BMC approach to zebrafish NT data and are consolidating our BMC approach by standardizing the BMR selection procedure. To obtain confidence around the BMC estimate, the lower and upper 95% confidence intervals (BMC_L and BMC_U , respectively) of the BMC were derived by analyzing concentration-response curves (at the selected BMR), which were generated by bootstrapping responses at each concentration. Furthermore, we could quantify the “selectivity” of the DT or NT effect (ie, a potency difference between a primary effect such as malformations and a proposed secondary nonspecific toxic effect such as mortality), by using the ratio of BMC of 2 effects. A highly potent active compound with higher selectivity to the primary effect should be first prioritized.

To apply our BMC approach for zebrafish DT and NT data, we need first to transform the raw data to toxicity endpoints that have specified directionality (ie, increasing/decreasing) with an assumption of concentration-response relationship. In addition to the typical endpoints, we also used endpoints that assess the differences in movement patterns between chemical-treated embryos and control embryos. By using data

generated from a 91-compound screen, we found the newly adopted endpoints can capture additional active compounds and our BMC results were highly concordant to the LOAEL results (Quevedo et al., unpublished data), with the BMC value being generally more sensitive, which is favorable for initial prioritization of compounds for further testing. With BMC_L and selectivity index, the 91 chemicals can be prioritized for their potential of DT and/or NT.

Currently we are working on applying the BMC method on other zebrafish datasets. The BMC method developed in this study can potentially provide a unified data analysis strategy for analyzing alternative animal model data and may maximize the alternative animal model data usage in the chemical risk assessment framework.

MATERIALS AND METHODS

Datasets

The data for these analyses were generated by exposing zebrafish embryos to a library of compounds provided by the NTP in DT and NT assays. The 91-compound library was made of diverse chemical classes including pesticides, polycyclic aromatic hydrocarbons, flame retardants, drugs, and compounds considered negative in most toxicity assays and was evaluated in a blinded manner to reduce bias. Additionally, 4 compounds had technical replicates, in order to better understand assay reproducibility. More information regarding the NTP library and compound purity can be reviewed in this study (Behl et al., forthcoming). The zebrafish studies were conducted by BBD-BioPhenix S.L (Biobide, San Sebastián, Guipuzcoa, Spain). The primary source for the assay data and LOAEL analysis can be found in the study (Quevedo et al., unpublished data). Herein, we summarized the assay protocols in the following sections and applied the LOAEL data from the primary source in the comparison.

Experimental Design

Developmental toxicity (DT) assay. Fertilized embryos (from transgenic line [Tg(Cmlc2: CopGFP)] expressing CopGFP under the myocardium specific promoter *cmlc2*, Letamendia et al., 2012) at 3–4 hours post-fertilization (hpf) were placed in 24-well plates (5 embryos per well, 15 embryos per condition, data from three wells were pooled assuming no well effect) with the corresponding chemical concentration (8 concentrations; 5–100 μ M). The concentrations were selected based on a dose-range finding study (Maximum Tolerated Concentration Assay). In the main study, ideally one nontoxic and one lethal (100%) concentration points were included. A group of embryos treated with DMSO (0.5% or 1% maximum, for some compounds with solubility issue or to allow a higher exposure) on a separate 24-well plate was used as a vehicle control. Plates were incubated at 28.5°C for 4 days and embryo media was replaced, and test items added at 2 days post-fertilization (dpf). Detailed analysis of embryo morphology (including malformations in the head, heart, and tail, deformed body shape and the presence of edemas) and lethality was performed at 2 and 4 dpf and analysis was done by several different technicians. Compounds were considered active if at any of the concentrations tested there were at least 20% of embryos showing malformations and the lowest tested concentration at which a significant effect was observed was reported for the active compounds (ie, LOAEL). The percentage of affected (number of dead + number of malformed) and dead embryos was used for effective concentration 50% (EC_{50}) and lethal concentration 50% (LC_{50}) calculations applying a nonlinear

regression test (sigmoidal dose-response curve) using GraphPad Prism (GraphPad Software, San Diego, California). A Teratogenic Index (TI) was estimated as the ratio between the LC_{50} and EC_{50} values. Two TIs were calculated, one per developmental stage (Tzima et al., 2017; Van Voorhis et al., 2016).

Neurotoxicity (NT) assay. In this study, NT is defined as neurotoxic effect that may be caused either due to developmental or acute exposure to the nervous system during embryonic development of the zebrafish.

Wild-type AB embryos were obtained from in-house husbandry and kept at 28.5°C until they reached 3 dpf. At this stage, larvae were dispensed in a 96 squared-well plate (1 embryo per well) and exposed to 5 concentrations per test substance with selection based on results in the DT assay (ie, LOAEL in the DT assay was used as the highest concentration evaluated). A total of 16 embryos were used per condition and a group of vehicle-treated embryos (0.5% DMSO) was used as the control. After 48 h of incubation at 28.5°C, plates were introduced in the Daniovision automated tracking system powered by Ethovision (Noldus). Temperature was set at 28.5°C and after 10 min of habituation, tracking, which consisted of 2 rounds of 10 min light and 10 min dark phases, started. Total duration of the tracking was 40 min. Several parameters were analyzed such as velocity, movement duration, and frequency among others, but the total distance moved (mm) was selected as representative of locomotor activity. The mean of the total distance moved by embryos in each group was measured in 2-min time bins and treated versus control groups were compared using the unpaired Student's *t* test. Compounds were considered active if significant differences were detected ($p < .005$) in more than one point at the same condition and tracking phase; the lowest tested concentration with a significant effect was then reported for the active compounds (ie, LOAEL) (Tzima et al., 2017; Van Voorhis et al., 2016).

Pre-BMC modeling

The raw data from DT and NT assays need to be transformed into toxicity endpoints that are amenable for BMC modeling (data transformation). Then to facilitate comparisons across datasets, toxicity endpoint responses were normalized (data normalization) when needed. The steps were demonstrated using one representative chemical (2, 2', 4, 4'-tetrabromodiphenyl ether, BDE-47) can be found in Figure 1 (a more abstract version can be found in Supplementary Figure 1) and details are provided in the following sections. The definitions of important terms used in this study can be found in Table 1.

Data transformation

The recorded zebrafish DT (binary incidence) data and NT (total amount of distance moved [mm] at a time interval [2 min]) data were transformed into appropriate toxicity endpoints. A toxicity endpoint is defined as a specific directional (increased/decreased) toxic effect (ie, response) with a tendency of monotonic, concentration-dependent response (ie, trend). The toxicity endpoints for quantifying DT or NT effect are explained below.

Developmental toxicity (DT) endpoints. The incidences of developmental effects were summarized as two toxicity endpoints: percent mortality (percentage of dead embryos) and percent affected (percentage of embryos either dead or positive in any malformation site). For this study, our primary purpose is to identify active compounds to prioritize for further testing, so we

do not distinguish between the specific types of malformations. Therefore, the response of "percent affected" endpoint is always greater than or equal that of "percent mortality." The response range of these two endpoints was from 0% to 100% (ie, all the embryos were either malformed or dead).

Neurotoxicity (NT) endpoints. The total amount of distance moved in a time interval per embryo was recorded (in this study, every 2 min). Based on these data, two types of NT endpoints were generated depending on the response they measured: quantity and similarity. The response of the quantity type NT endpoints was the total amount of distance moved per viable embryo, which was estimated by area under the time-response curve (AUC). The AUC was calculated for each light/dark phase (2 light phases + 2 dark phases); thus, in total, 4 quantity type NT endpoints were created. The response of the similarity type of NT endpoints was the degree of the similarity of movement patterns across the whole experiment time between a chemical-treated viable embryo and all the vehicle control-treated viable embryo (Figure 2). In addition, for embryos in the vehicle control, the movement similarity was calculated between themselves (eg, V01 vs V02 in Figure 2). The degree of similarity could be quantified by 3 metrics: cosine similarity (Bruni et al., 2016), Pearson correlation coefficient (Pearson's *r*), and Spearman rank correlation coefficient (Spearman's ρ). A higher positive value represents higher degree of movement pattern similarity between a pair of a vehicle control-treated only viable embryo and a chemical-treated viable embryo. For negative values produced by either Pearson's *r* or Spearman's ρ metric, a value of zero was substituted. The average similarity value between a reference embryo (either a chemical-treated or a vehicle control-treated embryo) and other vehicle control-treated embryos was used as the response value for the reference embryo. In total, 3 similarity type NT toxicity endpoints (ie, 3 similarity metrics) were created. The script for transforming the raw data into similarity indexes is available in GitHubGist (<https://gist.github.com/moggces/d0ef03f91d953742fa0ef03ffb73c0e6>; last accessed October 23, 2018).

Data normalization

To facilitate comparisons across datasets, toxicity endpoint responses were normalized to vehicle control and were shifted so that baseline response was 0. This normalization process was only applied on the NT data because for DT data, the baseline response was already 0 (ie, no dead/malformed embryos; thus, percent of mortality = 0). For the quantity type NT endpoints, first, the calculated AUC for each phase was transformed using the $\log_{10}(AUC + 1)$ function (+1 to avoid infinite values with an AUC of 0). Next, for each plate, response was calculated using the following equation: Response = $(V_{\text{chemical}} - V_{\text{vehicle control}}) * 100$, where V_{chemical} denotes the well values from chemical-treated viable embryos, and $V_{\text{vehicle control}}$ denotes the median of well values from viable embryos in vehicle control-only wells. Therefore, response >0 (<0) means chemical induced hyperactivity (hypo-activity) compared with responses in the vehicle control wells.

For the similarity type NT endpoints, the responses on each plate were normalized using the following equation: Response = $(V_{\text{chemical}}/V_{\text{vehicle control}}) * 100 - 100$, where V_{chemical} denotes the response of a chemical-treated viable embryo, and $V_{\text{vehicle control}}$ denotes the median value of responses of all the viable embryos in the vehicle control. Therefore, response <0 means chemical-treated embryos decrease similarity of movement compared with the vehicle control.

Figure 1. Overview of the steps to generate BMC and selectivity index for DT or NT effect using one compound (BDE-47) as an example. Each row represents a step with examples and each column shows the steps for a group of endpoints in one of the three categories (neurotox, devtox, and mortality). The descriptions of the steps are shown in the leftmost column, followed by two columns for NT data, and another two columns for DT data. The summary table for three endpoint categories (neurotox, devtox, and mortality) is shown at the last row. N/A means not applicable; some subplots were skipped on the figure to reduce the complexity.

BMC modeling

It is important to select an appropriate BMR because based on which, BMC is derived. Therefore, we created a procedure to select a BMR which is analogous to minimum activity threshold, based on intrinsic response variation in the endpoint. An R package was created to automate the processes (<https://github.com/moggces/Rcurvep>; last accessed October 23, 2018, v0.4) and a manuscript regarding the BMR selection procedure is in preparation. The steps (Curve simulation and Curve processing) were demonstrated using 1 chemical, BDE-47, on 2 endpoints, “LD_correlation(Spearman)” and “percent affected 4dpf” in Figure 1.

Curve simulation

For the two DT endpoints (ie, percentage of affected embryos and percentage of mortality), 1000 responses at each concentration were simulated using R boot package (Canty and Ripley, 2017) to bootstrap the vector of total number of embryos with number of incidences, and the percentage of incidence was calculated. For the NT endpoints (4 quantity type and similarity

type), only chemicals with at least 4 concentrations with viable embryos were processed (in this study, all of chemicals met the criterion and were all included in the analysis). The minimum number of 4 was chosen for allowing a concentration-response relationship formed. In total, 1000 responses at each concentration were simulated using the R *sample* function with replacement.

Curve processing

Curvep, a response noise filtering algorithm, was used to process the curves (Sedykh, 2016; Sedykh et al., 2011). Curvep relies on user-defined thresholds such as the baseline noise threshold (THR) and the maximum curve deviation (MXDV) to filter the response noise. Among thresholds, it is known that the THR (or minimum response threshold) has direct and significant impact on defining activity of testing chemicals. Responses smaller than the THR are considered as baseline noise and are adjusted to baseline (ie, 0). The activity of the chemical is defined as the concentration at the THR (point of departure, POD).

Table 1. The Definition of the Terms Used in the Study

Term	Definition
Endpoint	A specific directional (increased/decreased) effect (ie, response) with a tendency of monotonic, concentration-dependent relationship
Endpoint category	A group of endpoints designed to detect similar effects
Mortality endpoints	Three categories were created for this dataset: mortality, devtox, and neurotox Endpoints using percent of mortality as response
Developmental toxicity (devtox) endpoints	Two endpoints were created for this dataset: Percent mortality 4dpf and Percent mortality 2dpf Endpoints using percent of affected embryos (including dead embryos) as response
Quantity type neurotoxicity (neurotox) endpoints	Two endpoints were created for this dataset: Percent affected 4dpf and Percent affected 2dpf Endpoints based on total distance moved in either a light (L) or a dark (D) phase
Similarity type neurotoxicity (neurotox) endpoints	Four endpoints were created for this dataset: L1 distmoved 5dpf, D1 distmoved 5dpf, L2 distmoved 5dpf, and D2 distmoved 5dpf Endpoints based on movement pattern similarity between pairs of embryos only exposed to the vehicle control and embryos exposed to the chemical across the whole experiment time (LD) Three endpoints were created for this dataset: LD_correlation(Spearman), LD_correlation(Pearson), and LD_similarity(cosine)
Benchmark response (BMR)	The lowest response threshold at which the variance in potency estimation is sufficiently reduced
Benchmark concentration (BMC)	The concentration/potency at which the response is equivalent to the BMR
BMC _L /BMC _U	The lower bound or high bound of the 95% confidence interval of BMC
Active confidence score (score)	The fraction of simulated curves considered as active after Curvep
Selectivity index	The potency difference between neurotox and devtox endpoints (selectivity index for neurotoxic effect) or between devtox and mortality endpoints (selectivity index for developmental toxic effect)

Figure 2. An illustrative diagram to demonstrate the calculation of similarity index for NT data. a, The movement across time of chemical-treated embryos (BDE-47 at 10 μ M, each embryo with a different color). b, The movement across time of embryos in vehicle control wells (blues, each embryo with a different blue shade). c, Similarity (Spearman's rho in this case) of the movement profile was calculated between a single viable embryo treated by the chemical (C01, as an example) and all 16 single viable embryos in vehicle control only wells (V01 and V11 are shown as examples). d, The mean value (magenta dot) and the distribution of 16 calculated similarity values. The magenta dot is a single data point in the panel of similarity type NT endpoints (before normalization) in Figure 1. Other data points were generated accordingly by iteratively replacing C01 with other embryos treated by the chemical in this concentration.

BMR identification

For each simulated curve, the Curvep program with THR from 5 to 95 with an increment of 5 was applied. Then, a potency (ie,

POD) measurement at x threshold was reported (x is 5–95 with an increment of 5) for each simulated curve. For inactive curve (ie, all responses = 0 after Curvep), the POD was fixed to the

maximum tested concentration. The pooled variance of potency of all chemicals per *THR* was calculated. The BMR was considered as the *THR* at which the potency variance was sufficiently reduced and was the lowest threshold that potency variance was stabilized (Hsieh, 2016; Hsieh et al., 2015).

Post-BMC modeling

After BMC modeling, the activity results were summarized first by endpoint (Figure 1, activity report by endpoint), then by endpoint category (mortality/devtox/neurotox) (Figure 1, activity report by endpoint category). For each endpoint category, the activity call (active/inactive) and potency (BMC and its confidence interval) were reported for each chemical. Then, to quantify the “selectivity” of the DT or NT effect, the selectivity index was calculated based on the ratio of BMC of mortality to devtox or the ratio of BMC of devtox to neurotox (see the section of Selectivity index calculation).

Activity report by endpoint

Concentrations giving a response equivalent to the identified BMR were summarized to create a BMC value with a confidence interval. This BMC value was the median value (over all 1000 simulated curves) of the concentration at the BMR (an example can be found in Figure 1, “Processed 1000 curves (LD_correlation(Spearman))”, for the blue solid line) and its 95% confidence interval was calculated using the percentile method. The lower bound of the 95% confidence interval was designated as BMC_L and the upper bound was designated as BMC_U. A chemical was considered as active if over 50% of the 1000 simulated curves were not flat after Curvep based on the BMR and the percentage was assigned as active confidence score to the activity call.

Activity report by endpoint category

The toxicity endpoints derived from the zebrafish DT and NT data could be categorized into 3 groups: mortality (2 endpoints, 2dpf and 4dpf, from DT), devtox (2 endpoints, 2dpf and 4dpf, from DT), and neurotox (4 quantity type endpoints and 3 similarity type endpoints, from NT). A chemical was considered active in an endpoint category if it was active in any of the endpoints of the category. The lowest BMC and its associated confidence interval (BMC_L, BMC_U) of the endpoints in each group was used as the summarized activity.

Selectivity index calculation

For the active compound in the DT and/or NT assay, the selectivity index was calculated as follows: $\log_{10}(\text{BMC}_{\text{mortality}}/\text{BMC}_{\text{development}})$ for DT active compounds and $\log_{10}(\text{BMC}_{\text{development}}/\text{BMC}_{\text{behavior}})$ for NT active compounds. To be considered selective, the value of selectivity index should be larger than 0. For DT, it means that the BMC of DT is more potent than the BMC of mortality; for NT, the BMC of NT is more potent than the BMC of development; thus, a true NT effect can be distinguished from generalized toxicity that may occur at higher concentrations (especially important when interpreting hypo-activity).

Data availability

The activity results of chemicals (including the LOAEL used in the comparison) can be found in Supplementary File 1. The concentration response data of all the endpoints for all the chemicals were plotted and active compounds identified by the BMC

method along with the associated BMC value and confidence score were annotated on the plots (Supplementary Files 2 and 3). The distance moved per time interval data in NT assay and response data for endpoints are available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.rt7f0b0>.

RESULTS

Identification of BMR Values for Endpoints

We transformed the zebrafish DT and NT data into toxicity endpoints categorized into 3 groups: mortality, devtox, and neurotox. Selection of a proper BMR value for each endpoint is a critical step in BMC modeling as it is the basis for determining the BMC. The use of n standard deviation (SD) for BMR has been applied for BMC modeling both in vitro (Ryan et al., 2016; Sirenko et al., 2013) and in vivo (U.S. EPA, 2015) because SD is common metric to evaluate the variability of endpoint-specific background response. However, the choice of SD (either 1SD or 3SD) for BMR is not justifiable when the background response distribution is not normally distributed, which is the case in our dataset. Therefore, we applied a new approach (described in Materials and Methods section) to identify endpoint-specific BMRs along with the expected direction of the effect (eg, only increased effect is possible for DT endpoints). The BMRs together with the SD of responses in the vehicle control-only wells are provided in Table 2.

Results showed that for DT-mortality/devtox endpoints, the SD values were similar across endpoints; for similarity type NT-neurotox endpoints, LD_similarity(cosine) was the lowest among three; for quantity type NT-neurotox endpoints, the SD values in the light phases were higher than the ones in the dark phases. When comparing SD values with BMR values, in general, when SD of vehicle control response was large, the BMR was also large. Take L1 for example, which had the highest SD in NT-neurotox (quantity) endpoints, the BMR value for increased effect direction was also highest (70) and the BMR identification algorithm failed to identify a BMR value for the decreased effect direction. Despite the relation between the SD value and BMR value, the SD value could not be used to estimate BMR values. For example, the BMR value was the same for both LD_similarity(cosine) and LD_similarity(Spearman) but their SD values were quite different (5.65% vs 17.91%). The results suggested that, for the endpoints in this dataset, we could not justify the selection of BMR using SD approach, probably due to nonnormality of the background response distribution, and our approach based on intrinsic response variation of data provided a justifiable option for BMR selection.

Comparison of Active Compounds Identified by NT Endpoints

For NT data, in addition to quantity type endpoints, we also added 3 similarity type endpoints to capture various behavioral patterns between vehicle controls and chemical-treated embryos. Our results showed that when comparing the 3 similarity type metrics, the endpoint based on cosine similarity did not appear to contribute any additional (or unique) information that was not already captured by endpoints based on Pearson's r or Spearman's ρ metric for this dataset (Figure 3a). Overall, the similarity type metrics added unique information (ie, captured additional active compounds) that the quantity type metrics alone missed as shown by the overlap in Figure 3b. The results suggested the need for use of a combination of both

Table 2. The Identified BMR Values for the Toxicity Endpoints of Zebrafish DT and NT Data

Endpoint	SD	BMR	Direction
DT-mortality			
Percent mortality 4dpf	2.42 (%)	25 (%)	1
Percent mortality 2dpf	2.27 (%)	25 (%)	1
DT-devtox			
Percent affected 4dpf	2.99 (%)	25 (%)	1
Percent affected 2dpf	2.82 (%)	25 (%)	1
NT-neurotox (similarity)			
LD_correlation(Spearman)	17.91 (%)	25 (%)	-1
LD_similarity(cosine)	5.65 (%)	25 (%)	-1
LD_correlation(Pearson)	19.27 (%)	30 (%)	-1
NT-neurotox (quantity)			
L1 distmoved 5dpf	55.78	NA	-1
		70	1
D1 distmoved 5dpf	13.04	20	-1
		25	1
L2 distmoved 5dpf	34.31	45	-1
		45	1
D2 distmoved 5dpf	16.90	20	-1
		25	1

Note: SD: standard deviation; NA: not able to identify; direction = 1 for increased effect and direction = -1 for decreased effect; L1: first light phase; D1: first dark phase; L2: second light phase; D2: second dark phase; LD: all light and dark phases; SDs are based on responses in the vehicle control wells.

Figure 3. Overlap comparison of the active compounds from zebrafish NT toxicity endpoints. a, Three similarity type NT endpoints (cosine similarity, Spearman's rho, and Pearson's *r*). The total number of active compounds is 19 (Spearman rho), 18 (Pearson *r*), and 5 (cosine similarity). The union of the active compounds from the 3 endpoints were used for the comparison in (b). b. Similarity type versus quantity type NT endpoints.

metrics to capture actives more comprehensively in this dataset. The details of comparison are explained below.

Individual NT Endpoints

By using BMR values, BMC modeling was conducted per endpoint, and the results were summarized. First, we investigated the overlap of active compounds for the 4 quantity type NT endpoints. We found that 77% (36/47) of the active calls (a chemical can be active in multiple phases) were identified from endpoints related to the dark phases with hypo-activity (Supplementary Table 1). The chemicals with hypo-active effect identified in both the dark phases were generally concordant (64%, Supplementary Figure 2) but both phases also identified some unique hypo-active compounds. Only 5 unique active compounds were solely identified in the light phases (vs 28 unique chemicals in all phases) and only heptachlor

consistently showed hyper-activity in both light phases. Three active compounds (3, 3'-iminodipropionitrile, carbamic acid, butyl-, 3-iodo-2-propynyl ester, and amoxicillin) were only identified in the first light phase (L1), which has highest BMR (70) in all endpoints. For the 3 similarity type NT endpoints, the overlap of active compounds between 3 endpoints was plotted using the Venn diagram (Hulsen *et al.*, 2008, Figure 3a). The endpoint based on cosine similarity metric produced the fewest active compounds. The endpoints based on the Pearson's *r* metric and Spearman's rho metric produced almost identical active compounds except one additional active compound (lead (II) acetate trihydrate) from the endpoint based on Spearman's rho metric. In summary, results were similar when using either Pearson's *r* metric or Spearman's rho metric in this dataset and the total number of active compounds identified by the similarity type metrics were 19 (Spearman's rho), 18 (Pearson's *r*), and 5 (cosine similarity).

We also compared the other quantitative activity information from the NT endpoints. The distribution of the BMC, $\log_{10}(\text{BMC}/\text{BMC}_L)$, $\log_{10}(\text{BMC}_U/\text{BMC}_L)$, and active confidence scores of the active compounds were plotted (Supplementary Figure 3). The endpoint based on cosine similarity metric captured fewer active compounds and they tended to have lower active confidence score. Hence, based on this dataset, the cosine metric appeared to have comparatively inferior sensitivity. Except the cosine similarity endpoint, the BMC distribution was similar between the remaining endpoints, but active compounds in the similarity type NT endpoints provided smaller $\log_{10}(\text{BMC}/\text{BMC}_L)$ ranges and higher active confidence scores compared with the quantity type NT endpoints. The results indicated that responses in the similarity type NT endpoints presented a clearer biological signal of effect, thus higher confidence score, and similarity type NT endpoints may be potentially a preferred method to use as an endpoint.

Grouped NT Endpoints

The activity information for the four quantity type NT endpoints and three similarity type NT endpoints were summarized separately. The overlap of the active compounds between endpoints was compared (Figure 3b). Overall, more unique active compounds were identified by the quantity type NT endpoints in comparison with similarity type NT endpoints, where two unique active compounds, lead (II) acetate trihydrate and parathion, were identified. Compared with lead (II) acetate trihydrate, which was borderline active, parathion was a potent active compound missed by the quantity type NT endpoints, probably due to its severe non-monotonicity in the dark phases but its activity could be captured by the similarity type NT endpoints because of the clear concentration-response relationship when considering movement pattern similarity to the vehicle control. Because each endpoint type captured different active compounds, for this dataset, both quantity type and similarity type NT endpoints provided important activity information.

Activity Comparison Between BMC Approach Versus LOAEL Approach

There are known differences, such as method concepts, required inputs, and types of activity output between the LOAEL method (used by the data generator) and the BMC method which was employed here. For comparison, we listed the notable differences between these two methods in Table 3. For DT data, a regression method used by the data generator to

Table 3. Method Difference Between BMC Approach and Other Approaches for DT and NT Data

	LOAEL	BMC	EC ₅₀
Developmental toxicity (DT)			
Method	Threshold (20% of embryos showing malformations)	Curvep, a noise filtering algorithm, model-free	Sigmoidal nonlinear regression, fit to Hill equation
Data per calculation	Percent of malformation at a concentration	Endpoint responses at all concentrations	Endpoint responses at all concentrations
Vehicle control	NA	NA	As one data point in fitting
Potency report	LOAEL	BMC + confidence interval	EC ₅₀
Activity call report	Active	Active + active confidence score	Active
Prerequisites for calculation	NA	At least four concentrations	At least four concentrations
Neurotoxicity (NT)			
Method	Unpaired Student's t test adjusted for multiple comparison testing	Curvep, a noise filtering algorithm	NA
Data per calculation	Distance moved per time bin at a concentration	Endpoint responses at all concentrations	
Vehicle control	Used in calculation	Used in response normalization	
Potency report	LOAEL	BMC + confidence interval	
Activity call report	Neuroactive (hyper-active) or toxic (hypo-active)	Hyper-active ^a or hypo-active ^a or active (with confidence)	
Prerequisites for calculation	NA	At least 4 concentrations with viable embryos	

^aOnly for the quantity type NT endpoints; NA: not available.

generate EC₅₀ value was also listed for comparison. This regression method was similar to the BMC method. The only two differences were (1) the algorithm used for data fitting/processing and (2) the inclusion of the responses from the vehicle control wells in the curve fitting. The difference between the BMC method and the LOAEL method was larger in the attribute of "data per calculation." For example, for a chemical with 5 tested concentrations, in total, there would be 100 calculations conducted for 20-time bins in the LOAEL method versus only 1 in the BMC method per endpoint. Also, the BMC approach required at least 4 tested concentrations; without this, an activity call could not be derived. However, in this study, all of chemicals met the criterion and were all included in the analysis.

The BMC-LOAEL comparison was done at the unique chemical level. For LOAEL approach, the lowest (most potent) LOAEL value was used for chemicals with duplicate testing. For BMC approach, a chemical was considered active overall for DT or NT if it was active in any of the DT or NT endpoints respectively and the lowest BMC value was reported.

DT Data

Figure 4a presents the overlap of active compounds by BMC approach and the active compounds from the LOAEL approach (including chemicals with solubility issue at high concentrations and embryotoxic chemicals). The active compound concordance between BMC approach and the LOAEL approach was 100% ($n = 50$) and none of the negative control compounds was active. The value of BMC and LOAEL for the endpoint of "percent of affected embryo at 4 dpf" was compared. Generally, the BMC value was more potent than the LOAEL value (Figure 4b). Only 6 out of 50 active compounds had the LOAEL value more potent than the BMC value and their average potency difference between LOAEL and BMC was 0.136 (~1.37-fold), which was comparable with the average potency difference of the remaining active compounds (0.151, ~1.42-fold). The average (median) confidence interval

range was 0.266 (0.265), similar to the mean (median) tested concentration interval 0.26 (0.30). The results suggested that the amount of variation of potency estimation from the analysis/endpoints approximated the sensitivity of DT assay design.

NT Data

In order to conduct a fair comparison, some active compounds from the LOAEL approach were excluded from the comparison (eg, active compounds with all embryos being dead at the highest tested concentrations were excluded because there were no viable embryos for responses for BMC modeling). Excluded active compounds from LOAEL analysis can be found in Supplementary File.

The overlap in active compounds between the LOAEL and BMC approaches was plotted in Figure 5a. There were 26 overlapping active compounds and the overall active compound concordance was 74%. The LOAEL approach identified 5 additional unique active compounds (1-methyl-4-phenylpyridinium iodide, acenaphthene, anthracene, benzo(k)fluoranthene, and valproic acid sodium salt), whereas the BMC approach identified 4 additional active compounds, 3 of which were only hyper-active in the first light phase (3, 3'-iminodipropionitrile, carbamic acid, butyl-, 3-iodo-2-propynyl ester, and amoxicillin) and the other one is 4-H-cyclopenta(d, e, f)phenanthrene, which was only hyper-active in the second dark phase. None of negative control compounds was considered active in either BMC or LOAEL approach. The lowest BMC value among all the neurotoxic endpoints and the lowest LOAEL value from replicates were compared (Figure 5b). In summary, 16 of 26 active compounds had a more potent (ie, smaller) BMC versus a LOAEL and the average BMC-LOAEL potency difference was 0.286 (~1.93 fold). For the remaining active compounds (10 out of 26), although the BMC was less potent, the LOAEL-BMC difference was small (0.107, ~1.28 fold). The deltamethrin was an exception, which had the highest LOAEL-BMC difference (1.11). However, its

Figure 4. Comparison of BMC results versus LOAEL results based on zebrafish DT data. a. Active compound overlap. b. Potency comparison. In 43 out of 50 cases, the BMC was lower than the LOAEL and the reverse cases were highlighted. The 95% confidence interval of BMC was shown as the black error bar.

LOAEL was also less reproducible (the difference in LOAELs between the duplicates of deltamethrin was 0.7). The average (median) range of confidence interval was 0.75 (0.69), larger than the DT data, 0.30 (0.3). The results indicated that there was a greater amount of variation in potency estimation from the analysis/endpoints in the NT assay compared with the DT assay.

Chemical Prioritization

In addition to the potency, selectivity is also important in prioritizing the developmental toxicants and neurotoxicants. The Teratogenic Index (TI) from development (EC_{50}) and mortality (LC_{50}) endpoints can be adopted for quantifying selectivity for the DT data; there is no equivalent index for NT data. A $\text{TI} > 2$ (~ 0.3 in $\log_{10}$ unit) threshold was suggested based on the sensitivity/specificity using 27 chemicals including known DT chemicals, active compounds in *in vitro* embryotoxicity assays, and known inactive compounds (Selderslaghs et al., 2012). We found that there was a high ($R^2 = 0.96$) coefficient of determination (percent of variance explained) between TI and the selectivity index using BMC values for DT, indicating a direct

transferability between these two indices (Supplementary Figure 4). In the calculation, the 7 active compounds that had precipitation issue (thus, EC_{50} estimation stopped at the concentration where precipitation occurred but currently this information was not used in the BMC approach) were not considered. Based on the selectivity index, BMD_L , and active confidence score, we can prioritize the active compounds identified in DT and NT (Figure 6). A similar plot with a focus on potency is provided in Supplementary Figure 5. In the NT assay, most active pesticides were considered as selective and only 2 flame retardants (BDE-47 and 2-ethylhexyl diphenyl phosphate, EHDP) had comparable selectivity with high active confidence score. In the DT assay, the compounds with highest selectivity include pesticides, Polycyclic Aromatic Hydrocarbons (PAHs), and 2 flame retardants (isopropylated phenyl phosphate, IPP and Firemaster 550). This plot helps readers to focus on compounds with high potency and selectivity to prioritize compounds for further testing in *in vivo* rodent studies. One example of such a prioritization is a class of flame retardants (shown by the yellow dots), many of which are novel organophosphates with unknown DT and NT potential, currently

Figure 5. Comparison of BMC results versus LOAEL results based on zebrafish NT data. a. Active compound overlap. b. Potency comparison. In 16 out of 26 cases, the BMC was lower than the LOAEL and the reverse cases were highlighted. The 95% confidence interval of BMC was shown as the black error bar.

being used as replacements for some of the phased-out brominated compounds as indicated by increasing trends in their exposure (Mizouchi *et al.*, 2015; Sugeng *et al.*, 2017). As seen in the figure, many of these compounds showed selective DT effects in this assay; and 2 of them (EHDP and BDE-47) appear to be selective for NT. This information may be used to help design further assays in mammals for either individual chemicals or flame retardants as a group or compare this pattern with existing literature from *in vivo* studies in mammals (wherever available) to evaluate the validity of this method of prioritization.

DISCUSSION

Over the past decade, the zebrafish has rapidly been gaining popularity as a screening tool in drug development and toxicology due to its simplicity, yet relevant features as a *in vivo* model primarily due to conservation of underlying molecular pathways relevant to those in mammalian models (Planchart *et al.*, 2016). However, this is accompanied by some challenges

including understanding differences in strains, toxicokinetics in zebrafish, and data analysis hurdles, several of which have been identified previously (Planchart *et al.*, 2016; Sipes *et al.*, 2011; Truong *et al.*, 2016). Herein, we focus on developing a unified data analysis method that can potentially be applied across laboratories using different strains and protocols, to better evaluate the potential for zebrafish as a prioritization tool for further hazard characterization. This is a valuable contribution to the literature because to our knowledge, this is the first time that a BMC approach has been applied to zebrafish NT data. Our BMC approach centers around identifying an appropriate BMR value, which has the meaning of minimum activity threshold, based on intrinsic variation of the data in a toxicity endpoint. The more common $n \times \text{SD}$ approach for BMR is not justifiable in our case because the background response distribution tends to be nonnormal. Despite the assumption of monotonicity of concentration-response data in BMC modeling, the curve processing program (Curvep) applied in this study allows certain amount of nonmonotonicity (Sedykh, 2016), which can happen at the highest tested concentration due to unhealthy embryos.

Figure 6. Chemical prioritization based on BMC_L (y-axis), selectivity index (x-axis), and active confidence score (point size). Red dashed line: selectivity index threshold. Higher selectivity index value is equivalent to more selective. The link (https://hsiehjh.shinyapps.io/interactive_selectivity_plot/; last accessed October 23, 2018) provides the interactive version of the plot.

For the toxicity endpoints used in the BMC analysis, we applied 2 types endpoints for the NT data: quantity type and similarity type. The quantity type endpoints focus on the increased/decreased effect in each of the light switching phases and similarity type endpoints focus on the decreased similarity effect in the whole testing period. For this dataset, we found that similarity type endpoints were not sensitive enough when dividing data into light/dark phases probably due to limited number of data points in each phase for this dataset. When comparing the active compounds, both types (quantity and similarity) provide unique and complementary information by identifying some overlapping and some distinct active compounds. In terms of the potency estimation, active compounds in the similarity type NT endpoints (Pearson's r and Spearman's ρ) had smaller ranges (BMC/BMC_L) and higher active confidence scores compared with the quantity type NT endpoints, which may indicate a clearer biological signal of effect and potentially a preferred method to use as an endpoint.

When comparing results between the BMC approach and the LOAEL approach, we found that although in general, the active compound concordance was high for both DT and NT data (100% for DT vs 74% for NT), there was less concordance for NT versus DT. A possible explanation is that the toxicity endpoints used in the LOAEL and the BMC approaches differ more for the NT data than for the DT data. For the NT assay although the LOAEL approach used endpoints based on individual tests within each time bin, the BMC approach considered an overall effect either in light/dark phases, or in overall testing period. Thus, inherently the LOAEL approach could be more sensitive and needs further tests to adjust for multiple testing problem. Also, for the BMC approach, although a certain degree of

nonmonotonicity can be tolerated (eg, the drop of response due to unhealthy embryos), a concentration-response trend is expected. Comparatively, the LOAEL approach does not require this assumption. Considering the differences between two approaches, it is reassuring that most of the active compounds were still concordant in the NT assay even though overall, there were more active compounds identified by the LOAEL approach. These discordant active compounds tend to have weaker responses. In addition, although the BMC approach provided an overall summary of concentration-response effect that may be easier to digest, the movement pattern information (eg, constant moving regardless the light switches) was already condensed by the similarity metric and was not obtainable in the BMC results. Hence, it will also be important to inspect raw data for the movement pattern information when considering the BMC method.

Selectivity is an important aspect in both DT and NT data. We calculated a selectivity index based on the difference between BMC values for the respective endpoints. The BMC value in our setting represents the most potent activity which has the lowest point estimation variance in this dataset. The variance was estimated based on curves generated from sampling (with replacement) responses by concentrations. The value of the selectivity index larger than 0 represents that there is a difference of the BMC values between the respective endpoints. A biologically significant threshold needs to be derived using chemicals with known toxicity. For DT data, we found that the selectivity index was concordant with the more commonly used TI threshold (Selderslaghs *et al.*, 2012). In addition to the BMC, sampling procedures also provide additional benefits such as the confidence interval of BMC (BMC_L and BMC_U) and active confidence

score. This information can be integrated for prioritizing chemicals and will be useful in a chemical risk assessment framework.

In summary, toxicology shifts from an observational science to a more predictive field, there is a rapid switch from using more traditional rodent-based models to incorporation of high-throughput and high-content cell-based and alternate animal models such as zebrafish (Tice et al., 2013; Truong et al., 2016; Behl et al., forthcoming). With the adoption of new assays and models, one of the key challenges is the ability to develop a flexible data analysis approach that can be applied across different models for a systematic comparison of data. Although we have shown how a BMC approach can be applied to a battery of in vitro assays previously (Behl et al., 2015; Ryan et al., 2016), herein, we show for the first time, how it can be applied to more complex datasets such as zebrafish behavior. Ongoing studies involve applying this model to additional datasets to further evaluate its performance with an ultimate goal to compare and harmonize zebrafish data analysis across laboratories to strengthen its utility as an effective screening tool.

SUPPLEMENTARY DATA

Supplementary data are available at Toxicological Sciences online.

ACKNOWLEDGMENTS

We thank Dr Nisha S. Sipes for providing valuable comments when reviewing the manuscript and Dr Ainhoa Alzualde (Biobide, San Sebastián, Guipuzcoa, Spain) for helping to address Reviewers' comments.

FUNDING

Financial support was provided by the NIEHS contracts including HHSN273201400015C and HHSN273201700005C.

REFERENCES

- 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.
- 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., Hsieh, J.-H., Shafer, T. J., Mundy, W. R., Rice, J. R., Boyd, W. A., Freedman, J. H., Hunter, E. S., Jarema, K. A., Padilla, S., et al. (2015). Use of alternative assays to identify and prioritize organophosphorus flame retardants for potential developmental and neurotoxicity. *Neurotoxicol. Teratol.* **52**, 181–193.
- Behl, M., Ryan, K., Hsieh, J.-H., Parham, F., Shapiro, A., Collins, B. J., Birnbaum, L. S., Bucher, J. R., Walker, N. J., Foster, P. M., et al. (2018). Screening for developmental neurotoxicity (DNT) at the National Toxicology Program: The future is now! *Toxicol. Sci.* Forthcoming.
- Benchmark dose modeling – Introduction. https://clu-in.org/conf/tio/BMDS1/slides/BMDS_Introduction2.pdf. Accessed May 18, 2018.
- Bruni, G., Rennekamp, A. J., Velenich, A., McCarroll, M., Gendele, L., Fertsch, E., Taylor, J., Lakhani, P., Lenssen, D., Evron, T., et al. (2016). Zebrafish behavioral profiling identifies multitarget antipsychotic-like compounds. *Nat. Chem. Biol.* **12**, 559–566.
- Canty, A., and Ripley, B. (2017). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3–19.
- Crump, K. S. (1984). A new method for determining allowable daily intakes. *Fundam. Appl. Toxicol. Off. J. Soc. Toxicol.* **4**, 854–871.
- 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.
- 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.
- Hill, A. J., Teraoka, H., Heideman, W., and Peterson, R. E. (2005). Zebrafish as a model vertebrate for investigating chemical toxicity. *Toxicol. Sci.* **86**, 6–19.
- Horzmann, K. A., and Freeman, J. L. (2018). Making waves: New developments in toxicology with the zebrafish. *Toxicol. Sci.* **163**, 5–12.
- Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503.
- Hsieh, J.-H., Sedykh, A., Huang, R., Xia, M., and Tice, R. R. (2015). A data analysis pipeline accounting for artifacts in Tox21 quantitative high-throughput screening assays. *J. Biomol. Screen* **20**, 887–897.
- Hsieh, J.-H. (2016) Accounting artifacts in high-throughput toxicity assays. In *High-Throughput Screening Assays in Toxicology, Methods in Molecular Biology* (H. Zhu and M. Xia, Eds.), pp. 143–152. Springer, New York.
- Hulsen, T., de Vlieg, J., and Alkema, W. (2008). BioVenn – A web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**, 488.
- Jarema, K. A., Hunter, D. L., Shaffer, R. M., Behl, M., and Padilla, S. (2015). Acute and developmental behavioral effects of flame retardants and related chemicals in zebrafish. *Neurotoxicol. Teratol.* **52**, 194–209.
- Letamendia, A., Quevedo, C., Ibarbia, I., Virto, J. M., Holgado, O., Diez, M., Belmonte, J. C. I., and Callol-Massot, C. (2012). Development and validation of an automated high-throughput system for zebrafish in vivo screenings. *PLoS One* **7**, e36690.
- McCullum, C. W., Ducharme, N. A., Bondesson, M., and Gustafsson, J.-A. (2011). Developmental toxicity screening in zebrafish. *Birth Defects Res. C Embryo Today* **93**, 67–114.
- Mizouchi, S., Ichiba, M., Takigami, H., Kajiwar, N., Takamuku, T., Miyajima, T., Kodama, H., Someya, T., and Ueno, D. (2015). Exposure assessment of organophosphorus and organobromine flame retardants via indoor dust from elementary schools and domestic houses. *Chemosphere* **123**, 17–25.
- NTP. SEAZIT: Systematic evaluation of the application of zebrafish in toxicology. National Toxicology Program. <https://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/dev-tox/seazit/index.html>. Accessed May 18, 2018.
- Otte, J. C., Schultz, B., Fruth, D., Fabian, E., van Ravenzwaay, B., Hidding, B., and Salinas, E. R. (2017). Intrinsic xenobiotic metabolizing enzyme activities in early life stages of zebrafish (*Danio rerio*). *Toxicol. Sci.* **159**, 86–93.

- 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™ Phase I chemical library. *Reprod. Toxicol.* **33**, 174–187.
- 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.
- Ryan, K. R., Sirenko, O., Parham, F., Hsieh, J.-H., Cromwell, E. F., Tice, R. R., and Behl, M. (2016). Neurite outgrowth in human induced pluripotent stem cell-derived neurons as a high-throughput screen for developmental neurotoxicity or neurotoxicity. *NeuroToxicology* **53**, 271–281.
- Sedykh, A. (2016) CurveP method for rendering high-throughput screening dose-response data into digital fingerprints. In *High-Throughput Screening Assays in Toxicology, Methods in Molecular Biology* (H. Zhu and M. Xia, Eds.), pp. 135–141. Springer, New York.
- Sedykh, A., Zhu, H., Tang, H., Zhang, L., Richard, A., Rusyn, I., and Tropsha, A. (2011). Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environ. Health Perspect.* **119**, 364–370.
- Selderslaghs, I. W. T., Blust, R., and Witters, H. E. (2012). Feasibility study of the zebrafish assay as an alternative method to screen for developmental toxicity and embryotoxicity using a training set of 27 compounds. *Reprod. Toxicol.* **33**, 142–154.
- 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.
- Sirenko, O., Cromwell, E. F., Crittenden, C., Wignall, J. A., Wright, F. A., and Rusyn, I. (2013). Assessment of beating parameters in human induced pluripotent stem cells enables quantitative in vitro screening for cardiotoxicity. *Toxicol. Appl. Pharmacol.* **273**, 500–507.
- Sugeng, E. J., Leonards, P. E. G., and van de Bor, M. (2017). Brominated and organophosphorus flame retardants in body wipes and house dust, and an estimation of house dust hand-loadings in Dutch toddlers. *Environ. Res.* **158**, 789–797.
- Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: A Tox21 update. *Environ. Health Perspect.* **121**, 756–765.
- Truong, L., Simonich, M. T., and Tanguay, R. L. (2016) Better, faster, cheaper: Getting the most out of high-throughput screening with zebrafish. In *High-Throughput Screening Assays in Toxicology, Methods in Molecular Biology* (H. Zhu and M. Xia, Eds), pp. 89–98. Springer, New York.
- 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.
- Tzima, E., Serifi, I., Tsikari, I., Alzualde, A., Leonardos, I., and Papamarcaki, T. (2017). Transcriptional and behavioral responses of zebrafish larvae to microcystin-LR exposure. *Int. J. Mol. Sci.* **18**, 365.
- U.S. EPA. (2015). Benchmark dose software (BMDS). <https://www.epa.gov/bmds/download-benchmark-dose-software-bmds>. Accessed May 21, 2018.
- Van Voorhis, W. C., Adams, J. H., Adelfio, R., Ah Yong, V., Akabas, M. H., Alano, P., Alday, A., Alemán Resto, Y., Alsibae, A., Alzualde, A., et al. (2016). Open source drug discovery with the malaria box compound collection for neglected diseases and beyond. *PLoS Pathog.* **12**, e1005763.
- 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.