SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 34, No. 6, pp. 1436–1442, 2015

Printed in the USA

Hazard/Risk Assessment

STEPWISE INFORMATION-FILTERING TOOL (SIFT): A METHOD FOR USING RISK ASSESSMENT METADATA IN A NONTRADITIONAL WAY

AMY BEASLEY,*† SCOTT E. BELANGER,‡ and RYAN R. OTTER†
†Middle Tennessee State University, Murfreesboro, Tennessee, USA
‡Environmental Stewardship Organization, Mason Business Center, The Procter & Gamble Company, Cincinnati, Ohio, USA

(Submitted 7 November 2014; Returned for Revision 28 January 2015; Accepted 18 February 2015)

Abstract: Tools exist to evaluate large ecotoxicity databases for risk assessment purposes, but these tools are less useful for alternative analytical purposes. In the present study, the authors developed the Stepwise Information-Filtering Tool (SIFT), a strategic method to select relevant, reliable data from a large ecotoxicity database; demonstrated utility in a case study of chronic toxicity data for statistical endpoint comparison purposes; and evaluated SIFT by comparison with 2 existing data evaluation methods. *Environ Toxicol Chem* 2015;34:1436–1442. © 2015 SETAC

Keywords: Hazard risk assessment Ecotoxicology Data quality Klimisch method REACH

INTRODUCTION

"Big data" just keeps getting bigger. The increasing accessibility of data and ease of collection, plus new regulatory requirements under international chemical management programs such as Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), the Organisation for Economic Co-operation and Development (OECD) High Production Volume Challenge, and Categorization of the Canadian Domestic Substances List have resulted in massive ecotoxicology metadatasets [1]. Chemical registration through the European Chemical Agency (ECHA) has produced a database of dossier information on the toxicity of more than 12 000 chemical substances manufactured at more than 10 ton/yr [2]. Other databases, such as the US Environmental Protection Agency's (USEPA) ECOTOX database of more than 10 000 individual chemicals, are collections of toxicity test information from government, peer-reviewed literature, and private sources [3]. These datasets can be thought of as massive data warehouses in which the data submitted or gathered is all directed toward eventual use in risk assessment methodologies. Similar datasets on a smaller scale are compiled by private industry and government toward the same risk assessment endpoints, but may be tailored to a subset of chemicals (e.g., pesticides, pharmaceuticals) [4,5] or housed by a particular organization (e.g., CAL-ECOTOX, Columbia ERC) [6,7].

In typical risk assessment methodology, database size can be an advantage. A traditional chemical risk assessment workflow begins by gathering all available toxicity data, because more toxicity data equals more certainty about potential risk. Ideally, the dataset would incorporate multiple trophic levels and multiple toxicity endpoints. After data evaluation for suitability, an assessment of acute and chronic toxicity endpoints would be used to derive a predicted no-effect concentration (PNEC). Uncertainty factors are applied to sensitive endpoints based on the depth and breadth of available data. Uncertainty factors vary by region or regulatory authority based on considerations of data requirements in each jurisdiction [2,8].

The reliability of the data used to derive the final PNEC is an important consideration; reliability can be determined by evaluating the confidence in test methodology used, adherence to that methodology, and accurate reporting of the test event. The REACH and ECOTOX databases rely on specific submission guidelines to ensure that data quality is represented accurately [2]. Tools exist that are useful for data evaluation in a traditional hazard/risk assessment context [9–11] and that use a set of predefined criteria to evaluate data reliability. Klimisch et al. [12] developed a simple, well-known scoring method for assessing reliability based on adherence to sound science, quality assurance, and good laboratory practice guidelines. The ToxRTool is a Klimisch-based tool intended to expand and clarify definitions of reliability for use in REACH registration compliance [13]. Similar tools are designed to evaluate specific chemical types (e.g., nanomaterials, pesticides) [4,14] and data types (e.g., nonstandard toxicity data) [15–17]. Ågerstrand et al. [5] created a set of comprehensive criteria that uses relevance as well as reliability criteria for use in evaluating pharmaceuticals [5]. Although these tools differ, each serves a similar purpose, which is to evaluate data toward a single hazard or risk endpoint target.

Given the size and complexity of risk assessment datasets, it seems obvious that such a collection could be mined many ways to serve many purposes. Size and complexity, however, can become a disadvantage without a method to categorize and select the best data for the purpose [18]. Data gaps and overlaps are likely, and data quality may be difficult to quantify [11,19]. The present study was designed first to develop a strategic and systematic method using user-defined criteria to evaluate large datasets, then to demonstrate the usefulness of this method in a case study of chronic toxicity test reports for the purpose of evaluating improvements to test methodology, and finally to further evaluate the unique

* Address correspondence to Beasleyamy1@gmail.com; As5b@mtmail.mtsu.edu
Published online 27 February 2015 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.2955

function of the methodology by comparing case study results with both the Klimisch et al. [12] and Ågerstrand et al. [5] methods.

METHODS

Development of the Stepwise Information-Filtering Tool

The Stepwise Information-Filtering Tool (SIFT) was developed to evaluate and refine large datasets, with an emphasis on data relevance and reliability. Initial user-defined data selection criteria used in developing SIFT were based on USEPA test guidelines, OECD test guidelines, and expert judgment (Figure 1) [20–24]. The steps for using SIFT are as follows: step 0: define the dataset (a purpose for the study is defined, and a broad master dataset is compiled that generally covers the defined purpose); step 1: relevance criteria (relevance criteria narrow the master dataset based on the defined purpose); step 2: validity criteria (validity criteria, including, but not limited to, adherence to test guidelines narrow the step 1 dataset); step 3: acceptability criteria (acceptability criteria described by the desired parameters of the test design and reporting narrow the step 2 dataset); step 4: additional criteria (additional user-defined criteria relevant to the defined purpose narrow the step 3 dataset to the final set of studies). It is important to note that although the order of the steps in the SIFT methodology remains the same for each user or purpose, the criteria within each step are completely user-defined. For example, the validity criteria in step 2 depend on the relevant test guidelines selected in step 1.

Case study

To demonstrate the functionality of SIFT, a case study was conducted based on a company's request for a comparative analysis of the statistical measurement endpoints in chronic invertebrate toxicity test reports. Criteria specific to this purpose were identified for each of the SIFT steps (Table 1). The master dataset was compiled and then narrowed to a final dataset using the identified SIFT criteria.

Methodology comparison

User-defined SIFT criteria as used in the endpoint comparison case study and criteria from methods described by Klimisch et al. [12] and Ågerstrand et al. [5] were used to evaluate the dataset for the defined case study purpose (Tables 2 and 3). Ågerstrand et al. [5] list criteria separated into relevance and reliability then notes each as mandatory or optional. The SIFT criteria used in the case study were compared with the Ågerstrand et al. [5] criteria for relevance and reliability then noted as mandatory, not useful (or necessary), or unclear.

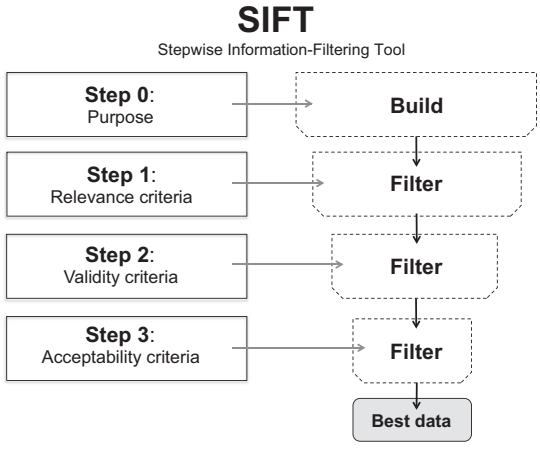

Figure 1. Process diagram of the Stepwise Information-Filtering Tool (SIFT) methodology.

RESULTS

Case study

Step 0: Define the dataset. The purpose of the case study was defined as "a comparative analysis of the statistical measurement endpoints in chronic invertebrate toxicity test reports." Database construction was initiated with a company-provided archive of reports spanning over 30 yr and included toxicity information on a wide range of chemicals. A targeted search of peer-reviewed literature was also performed to supplement the provided archive. The master dataset included 210 studies.

Step 1: Relevance criteria. Based on user-defined relevance criteria (Table 1) that fit the purpose identified in step 0, the master dataset was narrowed from 210 studies to 156 studies. Although studies from the literature rarely included effect data to the individual replicate level, 2 studies did pass the step 1 criteria. The no-observed-effect concentration (NOEC) was reported; or, when only a lowest-observed-effect concentration (LOEC) was reported, the NOEC was determined to be the next lowest test concentration below the LOEC.

Step 2: Validity criteria. Based on user-defined validity criteria (Table 1), the step 1 dataset was narrowed from 156 to 136 studies. Control mortality was defined per OECD test guideline 211 [21] for both species and minimum young applied to *Daphnia magna*. Valid study durations, in the context of an OECD test guideline 211 [21] for *Daphnia magna* and USEPA 1002.0 for *Ceriodaphnia* dubia [25], were defined as 21 d to 23 d and 7 d to 8 d, respectively.

Step 3: Acceptability criteria. Based on user-defined acceptability criteria (Table 1), the dataset from step 2 was narrowed from 136 to 95 studies. Effect concentrations for each study were calculated from reported raw effect data. Test type included semi-static and flow-through. Test strategy referred to configuration of replicates per concentration (e.g., 10 replicates of 1 daphnid, 4 replicates of 5 daphnid). At this step, chemical class was used to categorize the database, because some chemicals in the archive either were not in production or were proprietary formulations without Chemical Abstract Service (CAS) numbers. The CAS numbers were used to confirm class categorization when available. Dose–response and hormesis were analyzed from reported raw effect data.

Step 4: Additional criteria. Based on user-defined additional criteria (Table 1), the step 3 dataset was narrowed from 95 to 78 studies. Only tests of single-compound chemicals were included. Two of the following 4 parameters had to be reported for inclusion in the dataset: temperature, hardness, octanol–water partition coefficient (K_{OW}), or pH.

Methodology comparison

SIFT v. Ågerstrand. The Ågerstrand et al. method [5] included 75 total criteria. Of 12 relevance criteria, 11 were optional and only 1 was mandatory (reporting of references). Of 63 reliability criteria, 46 were mandatory and 23 optional, with many criteria represented in both mandatory and optional depending on the response given (Table 2).

Using criteria defined specifically for the case study, SIFT would have considered 4 of the Ågerstrand et al. [5] relevance criteria as mandatory per OECD test guideline 211 [10]. Three

Table 1. SIFT criteria developed specific to the purpose of the endpoint comparison case study

Step	Criteria
Step 1: Relevance criteria	Tests conducted under OECD *Daphnia* reproduction test protocol Test No. 211 [21] Species (*Daphnia magna* or *Ceriodaphnia dubia*) Reported NOEC value Effect data available to the level of individual replicate
Step 2: Validity criteria	$\leq$ 20% control mortality Minimum of 60 young per surviving adult by end of test Condition of organism (first instar, no ephippia, from the same culture) Test duration
Step 3: Acceptability criteria	Actual concentrations reported Minimum 5 concentrations plus control Parameters included or available: Test type, test strategy, chemical class, solvent used CAS number when available Dose–response present No hormetic effect
Step 4: Additional criteria	Single compound 2 of the following criteria reported: Hardness, pH, K_{OW}, temperature

SIFT = Stepwise Information-Filtering Tool; OECD = Organisation for Economic Co-operation and Development; K_{OW} = octanol–water partition coefficient.

additional relevance criteria would have been mandatory, whereas 3 were not useful and 2 were unclear. Of the 63 reliability criteria in the Ågerstrand et al. [5] method, 25 would have been mandatory with SIFT per OECD test guideline 211 [10]. Ten criteria would have been otherwise mandatory, whereas 24 would have been defined as not useful (or not necessary) and 4 as unclear. SIFT did not treat any criteria as optional.

SIFT v. Klimisch. The Klimisch et al. method [12] included 4 reliability criteria. In the present case study, all user-defined criteria were based largely on the OECD test guideline 211 [21] as specified under step 1, which would have resulted in a Klimisch score of 1 or 2 depending on expert judgment of whether test guidelines were followed appropriately. Klimisch et al. [12] do not include relevance criteria (Table 3).

DISCUSSION

Complex ecotoxicology datasets are becoming more common and more accessible. With such so-called "big data," evaluating data reliability is important. Although there are tools to evaluate data in a risk assessment context, gaps still exist for tools to evaluate the same data for other approaches. The Klimisch et al. [12] method was intended to simplify and clarify the data evaluation process and was considered highly useful for uses associated with impending REACH legislation [12]. With the Klimisch et al. method [12], data are generally categorized into 1 of 4 reliability classes to include the maximum reliable data to build a weight of evidence approach. The flexibility of the category definitions allows the method to be tailored to numerous risk assessment scenarios (e.g., in vitro and in vivo testing). Critics claim that the broad categories are too open to interpretation, thus complicating standardization and transparency; another concern is the expertise needed to interpret and apply evaluation criteria correctly [26–28]. Nevertheless, the Klimisch et al. method [12] is the most frequently cited method to date (Web of Science Core Collection) and is now part of the ECHA guidance on REACH registration [29]. Ågerstrand et al. [5] created a more comprehensive method for evaluating pharmaceutical ecotoxicology data in risk assessment [5]. To counter the typical focus on reliability, this method defines relevance criteria and emphasizes clear, comprehensive criteria definitions for increased transparency and ease of use; however, such explicit definitions can result in a narrow range of application [17]. The Klimisch et al. [12] and Ågerstrand et al. [5] methods were chosen to compare with SIFT because they were representative of a variety of existing tools [5,12].

Relevance

When the purpose of a study is a comparison within a dataset as opposed to the traditional comparison to an external reference value, the notion of relevance becomes an issue of initial data selection. This shift in approach means that the relevance of data is completely divorced from data reliability. The SIFT method uses upfront data-collection decisions that apply directly to each particular purpose. Data that do not include parameters key to the purpose are not useful, regardless of reliability. In an example from the case study, the presence of raw effect data to the individual replicate was essential to final endpoint calculations and comparisons. Therefore, "effect data present" was chosen as a step 1 criterion in the case study; Klimisch et al. [12] did not specify criteria that would address raw data, whereas Ågerstrand et al. [5] did but as an optional, late-stage criterion.

Both Klimisch et al. [12] and Ågerstrand et al. [5] consider relevance to be dependent on reliability. The Klimisch et al. [12] method does not predefine criteria to establish relevance. Instead, the method provides general guidance on how to use relevance to compare equally reliable tests. The Ågerstrand et al. [5] method elaborates on this approach, citing the influence of REACH guidance to assess relevance as "appropriateness of the test when it comes to a particular risk, e.g. whether the experimental model is representative to the environment that is aimed to be protected" [5]. Unlike Klimisch et al. [12], the method of Ågerstrand et al. [5] includes predefined relevance criteria, and relevance is considered equally as important as reliability to a complete evaluation of data quality. In the present case study, the selected step 1 relevance criteria netted 210 studies from proprietary and public literature that were suitable for further reliability screening. The Klimisch et al. [12] method would have required that all invertebrate studies available in the

Table 2. Comparison of Agerstrand versus SIFT criteria as applied to the case study dataset[a]

	Application to case study[b]				
Ågerstrand Criteria	Ågerstrand mandatory	Ågerstrand optional	SIFT mandatory	SIFT not useful	SIFT criteria unclear
Relevance Criteria					
Is the substance tested representative for the substance being risk assessed?		x		x	
Is the appropriate test species studied?		x	O		
Are the appropriate life-stage(s) studied?		x	O		
Are the appropriate endpoint(s) studied?		x	x		
Is the route of exposure relevant for the species?		x	O		
Does the test exposure scenario exist for the tested substance?		x			x
Are the stated tested doses/concentrations appropriate?		x			x
How do the tested doses relate to measured or predicted environmental concentrations (if available)?		x		x	
Is the time of exposure relevant and appropriate for the studied endpoints?		x	O		
Have other critical parameters influencing the endpoints than exposure time been considered adequately?		x	x		
Should the measured endpoint be considered to be an adverse effect or not?		x		x	
Are the references reported?	x		x		
Reliability criteria					
Purpose of study		x		x	
Description of endpoints	x		O		
Protocol standard/modified standard (if used)	x		O		
Test compound identification (e.g., name, CAS number)	x		x[c]		
Physicochemical data (e.g., volatility, stability, solubility, degradability, adsorption)		x		x	
Source		x		x	
Purity	x			x	
Vehicle (if used)	x		O		
Radiolabelled (if used)		x		x	
Tested doses or concentrations	x		O		
Measured doses or concentrations	x		x		
Exposure duration	x		O		
Exposure route	x		x[c]		
Exposure schedule (static, semistatic, flow-through system, other)	x		x		
Method of preparation of stock solutions		x		x	
Time-points of observations	x		O		
Analytical method	x			x	
Scientific name	x		O		
Body weight or length		x	x[c]		
Age/life-stage	x		O		
Growth/reproductive condition		x	O		
Gender	x[c]	x[c]		x	
Strain, clone	x[c]	x[c]	O		
Source	x			x	
Culture handling	x			x	
History of contamination for field-collected species	x			x	
Control described	x		O		
Control media identical to test media in all respects except treatment	x		O		
Control(s) identical to treatments in physical and chemical test conditions aspects: light, location, temperature in the room/climate chamber	x		O		
Control and test organism drawn from same population	x		O		
Control mortality/morbidity	x		O		

(*continued*)

Table 2. (*Continued*)

Ågerstrand Criteria	Application to case study[b]				
	Ågerstrand mandatory	Ågerstrand optional	SIFT mandatory	SIFT not useful	SIFT criteria unclear
Positive/negative control (if used)	x		O		
Vehicle control (if used)	x		O		
Known concentrations of vehicle (if used) in treatments and control	x		O		
Control mortality/morbidity reported for vehicle/positive control (if used)	x		O		
Historical control data		x		x	
Test environment pH	x		x		
Temperature	x		x		
Conductivity	x[c]	x[c]		x	
Light intensity and quality (source and homogeneity)	x[c]	x[c]		x	
Photo period	x			x	
Hardness of water		x	x		
Dissolved oxygen content	x[c]			x	
Ammonium/nitrite content in water	x[c]	x[c]		x	
Material and volume on aquarium/container	x		O		
Test medium	x		O		
Feeding protocols (for long-term tests)	x		O		
Food composition		x		x	
Sample size per replicates, number of organisms per replicates	x		O		
No. of organisms from each replicate used for statistical analysis (if not all used)	x				x
Randomized treatments		x	O		
Independence of observations		x	O		
Statistical method used	x				x
Significance level for NOEC and LOEC data (0.05 or less)	x		x		
Estimate of variability for LC and EC data	x			x	
Results reproduced by others		x		x	
Consistent with other findings		x		x	
Statistically significant responses noted (e.g., ECx)	x			x	
Dose–response reported in figure/text/table	x			x	
Each effect concentrations explicitly related to a specific endpoint		x			x
References to support the reliability of the study should be reported		x			x
Produced according to good laboratory practice guidelines		x		x	
Availability of raw data		x	x		

[a]Adapted from Ågerstrand et al. [5].
[b]O denotes criteria standard to OECD Test No. 211 [21]; x denotes criteria used with definitions as stated by Ågerstrand et al. [5].
[c]Criteria definitions with caveats (e.g., Ågerstrand's strain and clone are mandatory for algal tests and *Daphnia* only).
NOEC = no-observed-effect concentration; LOEC = lowest-observed-effect concentration; LC = lethal concentration; EC = effective concentration; ECx = concentration having an effect on x% of the population.

Table 3. Criteria developed by Klimisch et al. [12] to evaluate reliability[a]

Category	Definition (from Klimisch et al. [12])
1. Reliable without restrictions	"Studies or data from the literature or reports which were carried out or generated according to generally valid and/or internationally accepted testing guidelines (preferably performed according to GLP) or in which the test parameters documented are based on a specific (national) testing guideline (preferably performed according to GLP) or in which all parameters described are closely related/comparable to a guideline method."
2. Reliable with restrictions	"Studies or data from the literature, reports (mostly not performed according to GLP), in which the test parameters documented do not totally comply with the specific testing guideline, but are sufficient to accept the data or in which investigations are described which cannot be subsumed under a testing guideline, but which are nevertheless well documented and scientifically acceptable."
3. Not reliable	"Studies or data from the literature/reports in which there were interferences between the measuring system and the test substance or in which organisms/test systems were used which are not relevant in relation to the exposure (e.g., unphysiologic pathways of application) or which were carried out or generated according to a method which is not acceptable, the documentation of which is not sufficient for assessment and which is not convincing for an expert judgment."
4. Not assignable	"Studies or data from the literature, which do not give sufficient experimental details and which are only listed in short abstracts or secondary literature (books, reviews, etc.)."

[a]Table adapted from Schneider et al. [13].
GLP = good laboratory practice.

proprietary database be screened for validity, because the original purpose did not specify a particular test guideline or species. The Ågerstrand et al. [5] method would not have allowed 3 of the 4 case study step 1 criteria as relevance criteria and would have introduced 11 additional criteria to determine relevance. SIFT's focus on relevance in the initial data selection means that subsequent evaluations for reliability and acceptability are targeted only to useful data [5].

Balance between flexibility and utility

The development of SIFT stemmed from the difficulty in interpreting and adapting existing methods to the case study purpose. Because the initial question was not a traditional hazard/risk assessment and the case study was conducted by a research scientist rather than a risk assessment expert, both ease of use and flexibility in the method were important. Similar difficulties, even within the risk assessment community, have led to refining or creating new methods [9,11,30]. It seems that criteria definitions are the crux of the balance between the flexibility to tailor the method and the ability to use the method correctly. Klimisch et al. [12] and Ågerstrand et al. [5] represent opposing ends of this spectrum; the Klimisch et al. [12] method uses a few broad categories that require expert interpretation, whereas the Ågerstrand et al. [5] method uses many specific criteria, partly to alleviate the need for expert interpretation. If criteria are defined thoroughly, the transparency of the data and the possibility of standardization are increased [5]. Specific definitions also minimize the need for expert judgment; if criteria are spelled out, then less experienced users can assess data quality confidently, and the likelihood of consistent data evaluation increases [9]. Conversely, a method with many narrow criteria may appear so complex and cumbersome that it is not used [18,31]. Perhaps more importantly, data that might otherwise be useful could be discarded, therefore diminishing the sample size of data necessary for decision-making.

The SIFT method preserves flexibility through broadly defined steps and upfront selection of criteria specific to the end-goal analysis. This method will still benefit from expert judgment; however, a basic understanding of the testing framework used should allow the systematic breakdown of a dataset using SIFT criteria.

Application

The SIFT method is fundamentally about versatility in application. Both the Klimisch et al. [12] and Ågerstrand et al. [5] methods were created to serve very specific purposes: Klimisch's to characterize data quality broadly for inclusion in chemical registration databases and Ågerstrand's to improve reliability and harmonization of data reporting in scientific literature with an emphasis on pharmaceuticals. The goal of SIFT is to make risk assessment metadatasets more accessible to more applications (e.g., expert vs nonexpert, basic vs applied science, industry vs government vs academia). As part of the toxicity dataset analysis process, SIFT would be useful to inform business decisions such as the evaluation of contract lab performance or cost basis of in-house testing. Other potential applications include studies similar to Dowse et al. [16] or Euling et al. [32]. Dowse et al. [16] used toxicity data to compare standard testing with rapid testing in the construction of species sensitivity distributions. Euling et al. [32] performed a case study of toxicity data for dibutyl phthalate exposure to determine whether toxicogenomic data would be useful to elucidating modes of action [32].

CONCLUSION

The SIFT method is a useful addition to existing methods for risk assessment data evaluation. Current methods aim to evaluate large toxicity datasets toward traditional risk assessment purposes, leaving a gap for a method to evaluate these datasets in nontraditional ways. This tool applies user-defined evaluation criteria that gauge relevance and reliability in a stepwise manner, keeping a balance between the flexibility and utility of the method. Thus, SIFT is applicable across disciplines and levels of expertise.

Data availability—Metadata, associated metadata, and calculation tools are all disclosed within the manuscript.

REFERENCES

1. European Commission. 2006. Regulation (EC) 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH), establishing a European Chemicals Agency,

amending Directive 1999/45/EC and repealing Council Regulation (EEC) 793/93 and Commission Regulation (EC) No. 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official J Eur Union* L396:374–375.

2. European Chemicals Agency. 2008. Guidance on information requirements and chemical safety assessment—Chapter R.7b: Endpoint specific guidance. Helsinki, Findland.
3. US Environmental Protection Agency. 2012. *ECOTOX User Guide*: *ECOTOXicology Database System*, Ver 4.0. Washington, DC.
4. Roncaglioni A, Benfenati E, Boriani E, Clook M. 2004. A protocol to select high quality datasets of ecotoxicity values for pesticides. *J Environ Sci Health B* 39:641–652.
5. Ågerstrand M, Küster A, Bachmann J, Breitholtz M, Ebert I, Rechenberg B, Ruden C. 2011. Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals. *Environ Pollut* 159:2487–2492.
6. Office of Environmental Health Hazard Assessment, University of California at Davis. 2003. *Cal/Ecotox: California Wildlife Biology, Exposure Factor, and Toxicity Database*. State of California, Davis, CA, USA.
7. US Department of the Interior, US Geological Survey. 2001. *Columbia Environmental Research Center Acute Toxicity Database*. Columbia, MO.
8. Zeeman M, Fairbrother A, Gorsuch JW. 1995. Chapter 7—Environmental toxicology: Testing and screening. In Screening and testing chemicals in commerce: Background paper. OTA-BP-ENV-166. Congress of the United States, Office of Technology Assessment, Washington, DC, pp 59–68.
9. Hobbs DA, Warne MSJ, Markich SJ. 2005. Evaluation of criteria used to assess the quality of aquatic toxicity data. *Integr Environ Assess Manag* 1:174–180.
10. Breton RL, Gilron G, Thompson R, Rodney S, Teed S. 2009. A new quality assurance system for the evaluation of ecotoxicity studies submitted under the new substances notification regulations in Canada. *Integr Environ Assess Manag* 5:127–137.
11. Durda JL, Preziosi DV. 2000. Data quality evaluation of toxicological studies used to derive ecotoxicological benchmarks. *Hum Ecol Risk Assess* 6:747–765.
12. Klimisch H-J, Andreae M, Tillmann U. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul Toxicol Pharmacol* 25:1–5.
13. Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, Hartung T, Hoffmann S. 2009. "ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicol Lett* 189:138–144.
14. Card JW, Magnuson BA. 2010. A method to assess the quality of studies that examine the toxicity of engineered nanomaterials. *Int J Toxicol* 29:402–410.
15. Parkerton TF, Arnot JA, Weisbrod AV, Russom C, Hoke RA, Woodburn K, Traas T, Bonnell M, Burkhard LP, Lampi MA. 2008. Guidance for evaluating in vivo fish bioaccumulation data. *Integr Environ Assess Manag* 4:139–155.
16. Dowse R, Tang D, Palmer CG, Kefford BJ. 2013. Risk assessment using the species sensitivity distribution method: Data quality versus data quantity. *Environ Toxicol Chem* 32:1360–1369.
17. Ågerstrand M, Breitholtz M, Rudén C. 2011. Comparison of four different methods for reliability evaluation of ecotoxicity data: A case study of non-standard test data used in environmental risk assessments of pharmaceutical substances. *Environmental Sciences Europe* 23:1–15.
18. Przybylak K, Madden J, Cronin M, Hewitt M. 2012. Assessing toxicological data quality: Basic principles, existing schemes and current limitations. *SAR QSAR Environ Res* 23:435–459.
19. Wheeler J, Grist E, Leung K, Morritt D, Crane M. 2002. Species sensitivity distributions: Data and model choice. *Mar Pollut Bull* 45:192–202.
20. Organisation for Economic Co-operation and Development. 2004. Test No. 202: *Daphnia* sp. acute immobilisation test. *OECD Guidelines for the Testing of Chemicals*. Paris France.
21. Organisation for Economic Co-operation and Development. 2012. Test No. 211: *Daphnia magna* reproduction test. *OECD Guidelines for the Testing of Chemicals*. Paris France.
22. Organisation for Economic Co-operation and Development. 2013. Test No. 236: Fish embryo acute toxicity (fet) test. *OECD Guidelines for the Testing of Chemicals*. Paris France.
23. Organisation for Economic Co-operation and Development. 2011. Test No. 201: Freshwater alga and cyanobacteria, growth inhibition test. *OECD Guidelines for the Testing of Chemicals*. Paris France.
24. Organisation for Economic Co-operation and Development. 1992. Test No. 210: Fish, early-life stage toxicity test. *OECD Guidelines for the Testing of Chemicals*. Paris France.
25. US Environmental Protection Agency. 2002. Daphnid, *Ceriodaphnia dubia*, survival and reproduction test methd 1002.0. In *Short-term Methods for Estimating the Chronic Toxicity of Efluents and Receiving Waters to Freshwater Organisms*, 4th ed. EPA 821/R-02/013. Washington, DC, pp 141–196.
26. Maxim L, Van der Sluijs JP. 2014. Qualichem In Vivo: A tool for assessing the quality of in vivo studies and its application for bisphenol A. *PloS ONE* 9: e87738.
27. Petersen K, Lindeman B. 2014. *Road to Regulation of Endocrine Disruptors and Combination Effects*. Nordic Council of Ministers, Copenhagen, Denmark.
28. de Vries P, Murk AJ. 2013. Compliance of LC50 and NOEC data with Benford's Law: An indication of reliability? *Ecotoxicol Environ Saf* 98:171–178.
29. European Chemicals Agency. 2012. *Guidance on Data Sharing,* Ver 2.0, Helsinki, Finland.
30. Küster A, Bachmann J, Brandt U, Ebert I, Hickmann S, Klein-Goedicke J, Maack G, Schmitz S, Thumm E, Rechenberg B. 2009. Regulatory demands on data quality for the environmental risk assessment of pharmaceuticals. *Regul Toxicol Pharmacol* 55:276– 280.
31. Money CD, Tomenson JA, Penman MG, Boogaard PJ, Jeffrey Lewis R. 2013. A systematic approach for evaluating and scoring human data. *Regul Toxicol Pharmacol* 66:241–247.
32. Euling SY, Thompson CM, Chiu WA, Benson R. 2013. An approach for integrating toxicogenomic data in risk assessment: The dibutyl phthalate case study. *Toxicol Appl Pharmacol* 271:324–335.