Environmental Sciences Europe
a SpringerOpen Journal

RESEARCH Open Access

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

Marlene Ågerstrand[1*], Magnus Breitholtz[2] and Christina Rudén[1]

Abstract

Background: Standard test data are still preferred and recommended for regulatory environmental risk assessments of pharmaceuticals even though data generated by non-standard tests could improve the scientific basis of risk assessments by providing relevant and more sensitive endpoints.
The aim of this study was to investigate if non-standard ecotoxicity data can be evaluated systematically in risk assessments of pharmaceuticals. This has been done by evaluating the usefulness of four reliability evaluation methods, and by investigating whether recently published non-standard ecotoxicity studies from the open scientific literature fulfill the criteria that these methods propose.

Results: The same test data were evaluated differently by the four methods in seven out of nine cases. The selected non-standard test data were considered reliable/acceptable in only 14 out of 36 cases.

Conclusions: The four evaluation methods differ in scope, user friendliness, and how criteria are weighted and summarized. This affected the outcome of the data evaluation.
The results suggest that there is room for improvements in how data are reported in the open scientific literature. Reliability evaluation criteria could be used as a checklist to ensure that all important aspects are reported and thereby increasing the possibility that the data could be used for regulatory risk assessment.

Background

Environmental risk assessment of pharmaceuticals

In 2006, the European Medicines Agency (EMA) decided that all new marketing authorisation applications for human pharmaceuticals should be accompanied by an environmental risk assessment [1]. The EMA risk assessment has a PEC/PNEC (predicted environmental concentration/predicted no effect concentration) approach and is divided into two phases. In phase II, data on the substance's physicochemical properties, persistence and bioaccumulation, and ecotoxicity are reviewed and the PNEC is estimated. All relevant data should be taken into account. Experimental studies should preferably follow standard test protocols but it is recognized that there are other acceptable methods. However, their use should be justified and studies should be conducted in compliance with good laboratory practices (GLP) [1].

Standard and non-standard ecotoxicity tests

Ecotoxicological testing can be done using a variety of methods and models. There are two general approaches: using standard or non-standard testing methodologies. Standard tests refer to tests performed and reported according to a method described and provided by an official international or national harmonization or standardization organization, such as the OECD (Organisation for

* Correspondence: maa2@kth.se
[1]Department of Philosophy and the History of Technology, Royal Institute of Technology/Kungliga Tekniska Högskolan, Teknikringen 78B, 100 44 Stockholm, Sweden
Full list of author information is available at the end of the article

SpringerOpen

© 2011 Marlene et al; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Economic Cooperation and Development), US EPA (United States Environmental Protection Agency), ASTM (American Society for Testing and Materials), AFNOR (Association Française de Normalisation), and ISO (International Organization for Standardization). The test standard establishes a uniform specification of the experimental setup and execution, methods for data analyses, and the reporting format for the test data. Non-standard tests, on the other hand, are tests performed according to any other test method.

Regardless of whether a test is performed according to a standard or not, they should meet some general scientific quality criteria to demonstrate the reliability and reproducibility of the test results. Examples of such general scientific quality criteria are a clear description of the endpoints, inclusion of appropriate controls, appropriate identification of test substance and test organism, stated exposure duration time and administration route, and transparent reporting of effect concentrations.

The major advantages of using standard tests are that the results are directly comparable across substances and that the data they generate will be readily accepted across jurisdictions. Test guidelines also contributes to promote the reliability of the data by making it easier to repeat the experiment if needed because of the detailed standard test procedures and extensive reporting of data that is required. The major disadvantage of standard test methods is that it does not always represent the most biologically relevant testing approach depending on the type of endpoint under investigation. Therefore, results from non-standardized tests may in some cases be more sensitive and thereby contribute additional and significant information to a risk assessment. Other disadvantages of standard tests are that they are inflexible and therefore there is no room for case-by-case adjustments and that it may take up to 10-15 years to develop a new standard test.

Given the characteristics and purposes of standard tests, it is not surprising that they are mostly performed by commercial laboratories while non-standardized methods are typically performed by research scientists and published in scientific journals. Standard tests are often performed according to GLP, whereas non-standard tests are seldom performed according to GLP. The primary objective of the OECD Principles of Good Laboratory Practice is to ensure the generation of high-quality and reliable test data related to the safety of chemical substances and preparations [2]. But concerns have also been raised regarding whether GLP is synonymous with good scientific practices, accurate reporting and valid data [3,4].

Testing for environmental effects caused by pharmaceutical substances

Pharmaceutical substances have a number of inherent properties that make them interesting from a regulatory perspective. First, pharmaceuticals are carefully designed to interact with biological processes. Second, this interaction should be as specific as possible, ideally influencing only one well-defined target molecule or cellular process, and have as few other side effects as possible. Third, this interaction should be achieved at low concentrations, meaning that the substance has to be relatively potent. Fourth, to achieve this, it is necessary that the active pharmaceutical ingredient is sufficiently persistent to remain un-metabolized long enough to reach the target organ in the human body.

It is fundamental for risk identification, and thus a crucial part of the risk assessment process, to have toxicity test methods that are adapted to their purpose. Currently available standard test methods for deriving regulatory toxicity data for the aquatic environment are in many cases not sufficiently sensitive to the types of very specific effects that can be expected from pharmaceutical substances [see *e.g.*, [5]]. The EMA guideline [1] recommends that standard tests measuring growth inhibition and reproduction failure are used in environmental risk assessment of pharmaceutical substances (OECD test number 201, 210 and 211). However, test data are for many pharmaceuticals still limited or not publically available. The sex hormone ethinylestradiol is one of few substances where a significant amount of both standard and non-standard test data is available. Table 1 presents the lowest reported standard and non-standard effect values (according to the Wikipharma database [6] and the environmental classification system at fass.se [7]), both no-observed effect concentration (NOEC) and EC_{50} values (lowest identified effect concentration where 50% of the tested population have been found to be affected), for ethinylestradiol.

When comparing toxicity values, the non-standard NOEC value is 32 times lower than the standard test NOEC value, and the non-standard EC_{50} value is over 95,000 times lower than the standard EC_{50} value. Ethinylestradiol can therefore be seen as an example where non-standard tests with more substance-specific endpoints are more sensitive compared to the standard tests.

There is a need to carefully evaluate the regulatory process of identifying pharmaceuticals that might pose a risk to non-target species in the aquatic environment, and make sure that relevant and sufficiently sensitive tests are used in the regulatory environmental risk assessment of pharmaceuticals. As we see it, there are at least three ways forward: (1) to develop new standard ecotoxicity tests better suited for pharmaceuticals, or (2) to adjust existing standard tests by supplementing them with additional endpoints relevant for different pharmacological modes-of-action, or (3) to increase the use of non-standard tests for risk assessment purposes.

Table 1 The lowest publically available standard and non-standard effect values for ethinylestradiol[a]

Test standard	Test species	Endpoint	Effect values	Reference
OECD 201[b]	*Desmodesmus* spp (algae)	Growth inhibition	EC_{50} 0.13 mg/L NOEC <0.1 mg/L	[34]
OECD 210[b] (modified)	*Danio rerio* (fish)	Condition factor (comprising effects on weight and length)	NOEC 0.00001 mg/L[c]	[35]
OECD 211[b]	*Daphnia* (crustacean)	Reproduction	EC_{50} 0.105 mg/L	[36]
Non-standard test	*Danio rerio* (fish)	Spawning, egg production, fertilization	NOEC 0.00000031 mg/L	[37]
Non-standard test	*Danio rerio* (fish)	Fertilization success	EC_{50} 0.0000011 mg/L	[38]

[a]According to the Wikipharma database [3] and the environmental classification system at http://www.fass.se[4]. [b]Recommended for use in environmental risk assessments of pharmaceuticals (EMA, 2006); [c]The highest tested concentration.

OECD Organisation for Economic Cooperation and Development, NOEC no observed effect concentration. EC50 lowest identified effect concentration where 50% of the tested population have been found to be affected.

The development of new test standards is costly and may take up to 10 to 15 years [8] and since pharmaceuticals are a diverse group of substances when it comes to how they affect biological processes it is unlikely that new standards that would cover all relevant endpoints, could be developed in the near future.

Adjustments of current standard tests could increase the biological relevance for testing pharmaceutical substances. Therefore, a potential way forward could be that the standardization organizations initiate additional validation and expert commenting rounds, to standardize such adjustments. Still, such minor additions of existing standards would likely not be sufficient to ensure that the specific biological effects of most pharmaceuticals are covered by the tests.

Hence, in our view, an important and realistic way forward is to make increased use of non-standard test data to ensure a scientifically well-founded environmental risk assessment of pharmaceuticals. To enable the use of non-standard tests in risk assessments, two things are needed: that the legislation is designed so that non-standard tests can be included in a systematic and predictable way, and that non-standard tests are reported in a transparent and comprehensive way, much like required when using the standard test methods.

Reliability and relevance evaluation of (eco-)toxicity data

According to the TGD [9], an evaluation of data reliability should ensure "the inherent quality of a test relating to test methodology and the way that the performance and results of the test are described". Basically, this evaluation should answer the question: has the experiment generated and reported a true and correct result?

The assessment of the relevance of the data should describe "the extent to which a test is appropriate for a particular hazard or risk assessment" [9], *e.g.*, answer questions like: Is the measured endpoint a valid indicator of environmental risk? Is the experimental model sufficiently sensitive in relation to detecting the expected effects? Has the experimental model a sufficient statistical power? How representative is the experimental model to the environment that is aimed to be protected?

Evaluation of data can be done within different frameworks. It usually relies to a significant extent on case-by-case assessments based on expert judgment. However, there have also been attempts to make the evaluation process more structured. Such an approach can include checklists or even pre-defined evaluation criteria. A major advantage of using a structured way of evaluating data is increased transparency and predictability of the risk assessment process. For instance, both a checklist and pre-defined criteria will contribute to ensuring that at least a minimum and similar set of aspects are considered in each evaluation. Pre-defined evaluation criteria may also contribute to increased transparency of the evaluation process to the extent that these criteria are clearly reported to the relevant parties. Disadvantages of using pre-defined evaluation criteria or checklists are that they are obviously less flexible and need to focus on the general aspects of a study. In general, there is a need to strike a balance between flexibility and predictability in the data quality evaluation process; it will always include an element of expert judgment, but it is also, in our view, important to continuously seek to increase the predictability and transparency of this process.

Aim

The overall aim of this study was to investigate if the reliability of non-standard ecotoxicity data can be evaluated systematically in environmental risk assessments of pharmaceutical substances. Our hypothesis was that evaluation and reporting criteria can contribute to making the evaluation more systematic, predictable, and transparent, and facilitate the use of non-standard data for risk assessment purposes.

Method

This study is divided into two parts: (1) an evaluation of the usefulness of four methods for reliability evaluation of test data that have been proposed in the scientific

literature, and (2) an investigation of whether recently published non-standard ecotoxicity studies from the open scientific literature fulfill these reliability criteria.

Evaluation of existing reliability evaluation methods

The four evaluation methods used in this study are described by Klimisch *et al*. [10], Durda and Preziosi [11], Hobbs *et al*. [12], and Schneider *et al*. [13]. The reporting requirements from the OECD guidelines 201, 210, and 211 were used as a reference in the evaluation of the four methods. The reporting requirements were merged and generalized into 37 criteria so that they could be used on all types of endpoints and organisms.

Investigation whether published non-standard ecotoxicity studies fulfill proposed reliability criteria

The non-standard test data evaluated in this study (presented in Table 2[14-22]) have been selected for the current analyses since they were either (1) used in risk assessments of active pharmaceutical ingredients within the Swedish environmental classification and information system for pharmaceuticals [[23]; available at http://www.fass.se] or (2) used in a previous evaluation of this classification system [24]. These selection criteria resulted in a total of nine references that was then evaluated according to the four methods. Some references contain several ecotoxicological studies but only the part of the reference relevant for the chosen effect value was considered in this evaluation. Some of the evaluated studies could have been conducted according to a standardized method but since this is not reported in the reference, the study is treated as a non-standard test.

Results

The results section is divided into the following sections: hypothesis and endpoints (Table 3), protocol (Table 4), test substance (Table 5), test environment (Table 6), dosing system (Table 7), test species (Table 8), controls (Table 9), statistical design (Table 10), and biological effect (Table 11). Each section is reported and discussed in two parts; (1) an evaluation of the usefulness of existing proposed criteria for reliability evaluation of test data, and (2) an investigation of whether the evaluated non-standard study fulfill the proposed reliability criteria. The two parts are also summarized in the end (Summary of the evaluation of existing reliability evaluation methods and Summary of the reliability evaluation of the non-standard test data sections).

Hypothesis and endpoint

Usefulness of proposed criteria

All four evaluation methods consider that study endpoints should be stated and described. Durda and Preziosi [11] also considers hypothesis important for studies where NOEC or LOEC (lowest observed effect concentration) values are identified, and involves a

Table 2 Overview of the non-standard effect data evaluated in this study

Reference	Pharmaceutical substance	Endpoint	Test species and effect value
Andreozzi *et al.* [14]	Amoxicillin	Growth inhibition	*Synechococcus leopoliensis* NOEC (96 h) 0.78 μg/L
Ferrari *et al.* [15]	Ofloxacin	Growth inhibition	*Synechococcus leopolensis* EC_{50} (96 h) 16 μg/L
Ferrari *et al.* [15]	Sulfamethoxazole	Growth inhibition	*Synechococcus leopoliensis* EC_{50} (96 h) 26.8 μg/L
Ferrari *et al.* [15]	Propranolol	Growth inhibition	*Pseudokirchneriella subcapitata* NOEC (96 h) 5 mg/L Synechococcus leopolensis NOEC (96 h) 0.35 mg/L
Huggett *et al.* [16]	Propranolol	Number of eggs and percent hatch	*Oryzias latipes* LOEC (4 weeks) 0.5 μg/L
Robinson *et al.* [17]	Levofloxacin	Growth inhibition	*Pseudokirchneriella subcapitata* EC_{50} (72 h) 7.4 mg/L
Schmitt-Jansen *et al.* [18]	Diclofenac	Cell reproduction	*Scenedesmus vacuolatus* EC_{50} (24 h) 0.023 mg/L
Quinn *et al.* [19]	Ibuprofen	Morphology changes	*Hydra attenuata* EC_{50} (96 h) 1.65 mg/L
Metcalfe *et al.* [20]	Estradiol	Induced intersex (testis-ova)	*Oryzias latipes* NOEL (90 days) 0.0004 μg/L
Nentwig [21]	Fluoxetine	Reproduction inhibition	*Potamopyrgus antipodarum* NOEC (56 days) 0.47 μg/L
Halm *et al.* [22]	Estradiol	P450aromB mRNA expression in the brain	*Pinephales promelas* LOEC (4 days) 100 ng/L

NOEC/L no observed effect concentration/level, *LOEC* lowest observed effect concentration, EC_{50} lowest identified effect concentration where 50% of the tested population have been found to be affected.

Table 3 Modified evaluation criteria/questions concerning endpoints and a summary of the evaluation results

Evaluation method	Evaluation criteria/question	Type of criteria/mark	Summary of evaluation results
Klimisch *et al.*	Data on the measured parameters (including definitions).	Recommended	Data is presented but endpoints are not clearly defined in at least two studies.
Durda and Preziosi True	Hypothesis clearly defined for studies where NOEC or LOEC values are identified.	Recommended	The hypothesis is not stated in any of these studies.
	Endpoints appropriate for hypothesis.	Recommended	Endpoints appropriate for dose-response analyses.
Hobbs *et al.*	Was the biological endpoint (*e.g.*, immobilization or population growth) stated and refined (10 marks)? Award 5 marks if the biological endpoint is only stated.	0, 5, or 10	Stated in all studies, but not always refined
Schneider *et al.*	Are the study endpoint(s) and their method(s) of determination clearly described?	Recommended, 0, or 1	Not clearly defined in at least two studies.

relevance criterion by asking whether the chosen endpoint is appropriate for the hypothesis. No OECD guideline criteria matched this category.

Result of the study evaluations

The nine selected studies report NOEC/LOEC values and EC_{50} values, but none of the studies selected for this evaluation clearly stated a hypothesis. Instead endpoints were described, in some cases very thoroughly and in other cases hardly at all. Describing why a specific endpoint was used will help clarify the importance of the conducted study. It should be noted that it has been argued that hypotheses should be replaced by dose-response analysis when deriving exotoxicological benchmarks [25-28].

Protocol

Usefulness of proposed criteria

Both Klimisch *et al*. [10] and Durda and Preziosi [11] have evaluation criteria that are wide and imprecise which opens up for a variety of different interpretations and opinions.

Schneider *et al*. [13] evaluation question is also wide and concerns relevance, rather than reliability. In the accompanying guidance material to the question, a variety of aspects are included: the chosen test system and its applicability domain, consideration of physiochemical properties and stability of test substance, number of replicates, number of concentration levels and their range and spread, suitability of administration method, inclusion of all relevant endpoints, and statistical evaluation. The method would benefit from separating these aspects into several questions.

Result of the study evaluations

For three of the selected studies, a clear description of the test procedure was lacking, the majority of the studies would benefit from improving their reporting of the study. The chosen study designs were in all cases relevant for the data aimed at.

Test substance

Usefulness of proposed criteria

Identification of test compound, source, physicochemical properties such as purity and stability, and other substances used are factors related to the "test compound" that the four evaluation methods together consider important, but none of the methods report all factors. Purity is the only factor reported by all four methods. Water solubility is not reported explicitly but can be

Table 4 OECDs reporting requirements, modified evaluation criteria/questions concerning protocol, and a summary of the evaluation results

Evaluation method	Evaluation criteria/question	Type of criteria/mark	Summary of evaluation results
Klimisch *et al.*	Clear description of the test procedure (complete documentation).	Recommended	Not stated in three studies.
Durda and Preziosi	Protocol described and followed (standardized protocol/validated protocol/published peer-reviewed protocol/scientifically accepted protocol). Tests are recommended to be conducted using good laboratory practices (GLP).	Mandatory/ recommended	All studies are peer reviewed and non-standard, but not according to GLP.
	Peer-reviewed study	Recommended	All studies are peer reviewed.
Schneider *et al.*	Is the study design chosen appropriate for obtaining the substance-specific data aimed at?	Mandatory, 0, or 1	Appropriate study design for dose-response analyses for all studies.
OECD guidelines	Clear reporting instructions for each test standard.	Mandatory	The non-standard studies were not evaluated according to the OECD reporting requirements.

Table 5 OECDs reporting requirements, modified evaluation criteria/questions concerning test substance, and summary of the evaluation results

Evaluation method	Evaluation criteria/question	Type of criteria/mark	Summary of evaluation results
Klimisch *et al.*	Specification of the test substance (purity, by-products).	Recommended	Purity is not stated in six studies, by-products are not stated in any of these studies.
	Data on neutralization of samples (for basic or acid substances).	Recommended	Not stated in any of these studies.
	Use of emulgators/solubilizers.	Recommended	Not stated in any of these studies.
Durda and Preziosi	Chemical species noted.	Recommended	Stated in all studies.
	Chemical source noted.	Recommended	Not stated in one study.
	Purity/stability noted.	Recommended	Purity is not stated in six studies, stability is not stated in any of these studies.
	Vehicle described (if used).	Recommended	Not stated in four studies.
Hobbs *et al.*	Were analytical reagent grade chemicals or the highest possible purity chemicals used for the experiment?	0 or 3	Not stated in six studies.
Schneider *et al.*	Is information on the source/origin of the substance given?	Recommended, 0, or 1	Not stated in one study.
	Was the test substance identified?	Mandatory, 0, or 1	Stated in all studies.
	Is the purity of the substance given?	Recommended, 0, or 1	Not stated in six studies.
	Is all information on the nature and/or physicochemical properties of the test item given, which you deem indispensable for judging the data?	Recommended, 0, or 1	Not stated in five studies.
OECD guidelines	Chemical identification data (*e.g.*, CAS Number); physical nature and relevant physical-chemical properties; water solubility; purity; stability; suspended solids.	Mandatory	The non-standard studies were not evaluated according to the OECD reporting requirements.

interpreted into Schneider *et al*. [13] method. The OECD guidelines do not report of any other factors in addition to the ones mentioned by the four methods.

Result of the study evaluations

The purity of the test substance was not stated in six of the nine studies. Neither stability nor by-products of the test substance was stated in any of the studies. Other factors such as vapor pressure, water solubility, octanol/water partitioning coefficient (logKow) and bioconcentration factor were missing in several or all of the studies.

Test environment

Usefulness of proposed criteria

Oxygen concentration, conductivity, temperature, pH, water hardness, salinity, light intensity, photoperiod, physical structure of the test chamber, test media, test organism density, food composition and food availability are abiotic and biotic factors that the four methods specify in their evaluation schemes. Several of the reported factors could modify the toxicity of chemicals [29]. None of the four evaluation methods include all factors. The OECD guidelines recommend that light quality, residual chlorine levels, total organic carbon (TOC), and chemical oxygen demand (COD) are reported in addition to the other factors.

Schneider *et al.* [13] did not specify which factors related to the test environment that should be evaluated and only consider it relevant for repeated dose toxicity studies. Klimisch *et al*. [10] has gathered the factors in only two criteria, while Durda and Preziosi [11] and Hobbs *et al*. [12] have divided the factors into several criteria/questions. Having one factor per criteria/question facilitates the evaluation since only one thing at a time has to be considered. Durda and Preziosi [11] has, contrary to Klimisch *et al*. [10] and Hobbs *et al*. [12], considered feeding protocols for long-term tests a "must" criterion.

Result of the study evaluations

None of the nine studies report all "test environment" factors. Temperature was reported more often than the other factors, whereas pH and dissolved oxygen (DO) were the factors that most of the studies failed to report. In some studies, information was given on the conditions for the cultivation/breeding stock but not for the experimental setup.

Dosing system

Usefulness of proposed criteria

Administrated concentrations, concentration control analysis, administration route, frequency and duration of exposure, and information of the dosing type are aspects related to the "dosing system" that the four evaluation methods all together consider as important. Durda and Preziosi [11] is the only method that covers all aspects. Both Durda and Preziosi [11] and Schneider *et al*. [13] consider administrated concentrations, administration route, frequency and duration of exposure as "must"

Table 6 OECDs reporting requirements, modified evaluation criteria/questions concerning test environment, and summary of the evaluation results

Evaluation method	Evaluation criteria/question	Type of criteria/mark	Summary of evaluation results
Klimisch *et al.*	Data on physical and chemical test conditions (pH, conductivity, light intensity, temperature, hardness of water).	Recommended	Temperature is not stated in two studies, pH is not stated in seven studies, conductivity is not stated in any of these studies, hardness of water is not stated in any of these studies, light intensity is not stated in two studies.
	Data on the feeding of the test animals (chronic studies).	Recommended	Stated in all studies.
Durda and Preziosi	Water characteristics recorded (*e.g.*, pH, DO, temperature).	Recommended	Temperature is not stated in two studies, pH is not stated in seven studies, DO is not stated in three studies.
	Light intensity and/or photoperiod.	Recommended	Not stated in two studies.
	Physical structure of test environment recorded.	Recommended	Not stated in one study.
	Number of animals per test apparatus (*e.g.*, aquaria) noted.	Recommended	Not stated in one study.
	Feeding protocols noted (long-term tests).	Mandatory	Not stated in one study.
	Food composition noted/known.	Recommended	Not stated in one study.
Hobbs *et al.*	Was the temperature measured and stated (3 marks)? Award 1 mark if only the temperature settings of the room or chamber are stated?	0, 1, or 3	Not stated in two studies and only specified in two studies.
	For tests not using aquatic macrophytes and alga (*i.e.*, non-plant), was the dissolved oxygen content of the test water measured during the test?	0 or 3	Not stated in three out of five studies.
	Was the pH measured and values stated? Award 1 mark if it is measured but not stated or if the pH of the dilution water only is measured and stated.	0, 1, or 3	Not stated in seven studies.
	For marine and estuarine water, was the salinity/ conductivity measured and stated?	0 or 3	No marine studies were evaluated.
	Was the type of test media used stated?	0 or 5	Not stated in three studies.
Schneider *et al.*	For repeated dose toxicity studies only: Is information given on the housing or feeding conditions?	Recommended, 0, or 1	No repeated dose toxicity studies were evaluated.
OECD guidelines	Photoperiod, light intensity and quality (source, homogeneity); pH values at the beginning and at the end of the test at all treatments; hardness; temperature; dissolved oxygen concentration; test vessel description including volume; detailed information on feeding (*e.g.*, type of food(s), source, amount given and frequency); residual chlorine levels (if measured); TOC and COD; Salinity of the test medium.	Mandatory	The non-standard studies were not evaluated according to the OECD reporting requirements.

criteria. The OECD guidelines also recommend that date of start of the test; method of preparation of stock solutions, the recovery efficiency of the method, and the limit of quantification in the test matrix should be reported.

Result of the study evaluations

The tested concentrations were not stated in two studies and in two other studies only the concentration range was reported. Concentration control analyses were made in five of the nine studies. The administration routes were rarely stated explicitly. A possible reason for this is that it is considered self-evident when it comes to aquatic toxicity testing.

Test species

Usefulness of proposed criteria

Identification of test species, number of individuals, investigated period of the lifecycle, reproductive condition, sex, strain, source, body weight, length or mass are aspects related to the "test species" that the four evaluation methods all together as consider important. None of the methods cover all aspects. Only Schneider *et al.* [13] has a "must" criterion: identification of the test species. The OECD guidelines also recommend that culture conditions and methods of collecting the test species are described.

Result of the study evaluations

Complete information about the test organism was missing in all nine studies.

Controls

Usefulness of proposed criteria

The reported aspects connected to "controls" are: use of control (positive, negative and/or solvent), acceptability criteria, control media identical to test media in all respect except the treatment variable, and origin of the control and test organisms. Only Durda and

Table 7 OECDs reporting requirements, modified evaluation criteria/questions concerning dosing system, and summary of the evaluation results

Evaluation method	Evaluation criteria/question	Type of criteria/mark	Summary of evaluation results
Klimisch *et al.*	Data on concentration control analysis.	Recommended	Concentrations are not measured in four studies.
	Data on dosing the test substance (static, semi-static, flow through system).	Recommended	Not clearly stated, but could be understood for all but one study.
	Data on the exposure period.	Recommended	Stated in all studies.
Durda and Preziosi	Dose (measured preferred).	Mandatory	Concentration ranges are presented for two studies, tested concentrations are lacking for two studies, concentrations are not measured in four studies.
	Administration route (environmentally relevant preferred).	Mandatory	Not clearly stated, but could be understood for all but one study.
	Exposure schedule (intermittent, continuous or *ad libitum*).	Recommended	Not stated in one study.
	Exposure duration.	Mandatory	Stated in all studies.
Hobbs *et al.*	Were the chemical concentrations measured?	0 or 4	Concentrations are not measured in four studies.
	Was the type of exposure (*e.g.*, static, flow through) stated?	0 or 4	Not clearly stated, but could be understood for all but one study.
	Was the duration of the exposure stated (*e.g.*, 48 or 96 h)?	0 or 10	Stated in all studies.
Schneider *et al.*	Are doses administered or concentrations in application media given?	Mandatory, 0, or 1	Concentration ranges are presented for two studies, tested concentrations are lacking for two studies. Concentrations are not measured in four studies.
	For inhalation studies and repeated dose toxicity studies only (give point for other study types): Were achieved concentrations analytically verified or was stability of the test substance otherwise ensured or made plausible?	Recommended, 0, or 1	No inhalation and/or repeated dose toxicity study was evaluated.
	Is the administration route given?	Mandatory, 0, or 1	Not clearly stated, but could be understood for all but one study.
	Are sufficient details of the administration scheme given to judge the study?	Recommended, 0, or 1	Not stated in three studies.
	Are frequency and duration of exposure as well as time-points of observations explained?	Mandatory, 0, or 1	Stated in all studies.
OECD guidelines	Date of start of the test; test duration; test procedure used (*e.g.*, semi-static or flow through); method of preparation of stock solutions; test concentrations, nominal and measured; the recovery efficiency of the method; the limit of quantification in the test matrix.	Mandatory	The non-standard studies were not evaluated according to the OECD reporting requirements.

Preziosi [11] cover all these aspects. Klimisch *et al.* [10] does not mention any of the aspects. Both Durda and Preziosi [11] and Schneider *et al.* [13] have "must" criteria/questions. The four methods all together covered the same aspects as the OECD guidelines.

When it comes to acceptability criteria for the test, *e.g.*, control mortality, two different approaches can be seen. Hobbs *et al.* [12] only requires that the acceptability criteria are stated while Durda and Preziosi [11] and Schneider *et al.* [13] ask whether the results connected to the acceptability criteria are reliable or acceptable. Durda and Preziosi [11] provides a percent limit for the control mortality while Schneider *et al.* [13], in the guidance material, asks whether the variability of the results and control was acceptable and if control values were within reasonable range.

Result of the study evaluations

In two of the evaluated studies, the authors did not report whether controls were used or not. Acceptability criteria were not stated in any of the studies, but since Schneider *et al.* [13] asks whether the variability of the results and control was acceptable, one of the studies were considered to fulfill this evaluation question. The origin of the control and test organisms and whether the control media is identical to test media in all respect except the treatment variable was not specifically stated in any of the studies. A possible reason for this could be that this is considered to be self-evident.

Statistical design

Usefulness of proposed criteria

The reported aspects connected to "statistical design" are: statistical method and results including significance levels and estimates of variability, (sufficient) sample size and replicates, randomized treatments and independent observations. None of the methods cover all aspects. The four methods all together covered the same aspects as the OECD guidelines. Schneider *et al.* [13]

Table 8 OECDs reporting requirements, modified evaluation criteria/questions concerning test species, and summary of the evaluation results

Evaluation method	Evaluation criteria/question	Type of criteria/mark	Summary of evaluation results[a]
Klimisch *et al.*	Data on the test species and the number of individuals tested.	Recommended	Not stated in one study.
	Information about the investigated period of the life cycle of the test animals (chronic studies).	Recommended	Not stated in three studies.
Durda and Preziosi	Body weight or length.	Recommended	Not stated in two studies.
	Age/life stage.	Recommended	Not stated in three studies.
	Reproductive condition.	Recommended	Not stated in four studies.
	Gender.	Recommended	Not stated in one study.
	Strain.	Recommended	Not stated in four studies.
	Source.	Recommended	Not stated in one study.
	No previous exposure (*e.g.*, for field-collected specimens).	Recommended	No field-collected test species.
	No concomitant exposure (*e.g.*, during field studies).	Recommended	No field studies.
Hobbs *et al.*	Were the characteristics of the test organism (*e.g.*, length, mass, age) stated?	0 or 5	Not stated in two studies.
Schneider *et al.*	Is the species given?	Mandatory, 0, or 1	Stated in all of the studies.
	Is age or body weight of the test organisms at the start of the study given?	Recommended, 0, or 1	Body weight is not stated in two studies, age is not stated in three studies.
	Is the sex of the test organism given?	Recommended, 0, or 1	Not stated in one study.
	Is information given on the strain of test animals plus, if considered necessary to judge the study, other specifications (see explanation for examples)?	Recommended, 0, or 1	Not stated in four studies.
OECD guidelines	Scientific name; strain/clone; source; culture conditions including medium and volume; methods of collection.	Mandatory	The non-standard studies were not evaluated according to the OECD reporting requirements.

[a]Some aspects are not relevant for all studies, depending on choice of test species. This has been considered in the evaluation.

Table 9 OECDs reporting requirements, modified evaluation criteria/questions concerning controls, and a summary of the evaluation results

Evaluation method	Evaluation criteria/question	Type of criteria/mark	Summary of evaluation results
Durda and Preziosi	Control media identical to test media in all respect except the treatment variable.	Mandatory	Not stated in any of these studies, controls not stated at all in two studies.
	Control and test organism drawn from same population.	Mandatory	Not stated in any of these studies, controls not stated at all in two studies.
	Acceptable control mortality/morbidity (approx. 10%).	Mandatory	Not stated in any of these studies.
	Vehicle control (as appropriate).	Recommended	Not stated in six studies.
	Positive and/or negative control (optional).	Recommended	Controls not stated in two studies.
Hobbs *et al.*	Were test acceptability criteria stated (*e.g.*, mortality in controls must not exceed a certain percentage) OR	0 or 5	Not stated in any of these studies.
	Were the acceptability criteria inferred (*e.g.*, test method used [USEPA, OECD, ASTM *etc.*] uses validation criteria) (award 2 marks).	0, 2, or 5	No standard tests were evaluated.
	Were appropriate controls (*e.g.*, no-toxicant control and/or solvent control) used?	0 or 5	Not stated in two studies.
	Were parallel reference toxicant toxicity test conducted?	0 or 4	Not stated in any of these studies.
Schneider *et al.*	Were negative (where required) and positive controls (where required) included (give point also, when absent but not required)?	Mandatory, 0 or 1	Not stated in two studies.
	Are the quantitative study results reliable (variability of controls and treatments)?	Recommended, 0, or 1	Not stated in eight studies.
OECD guidelines	Evidence that controls met the overall survival acceptability standard of the test species.	Mandatory	The non-standard studies were not evaluated according to the OECD reporting requirements.

Table 10 OECDs reporting requirements, modified evaluation criteria/questions concerning statistical design, and summary of the evaluation results

Evaluation method	Evaluation criteria/question	Type of criteria/mark	Summary of evaluation results
Klimisch *et al.*	Data on the statistical evaluations (including method).	Recommended	Not stated in two studies, results missing for one study.
Durda and Preziosi	Sufficient sample size/replicates.	Recommended	Not stated in one study, sufficient in all other studies.
	Randomized treatments.	Recommended	Not stated in any of these studies.
	Independence of observations.	Recommended	Not stated in any of these studies.
	Appropriate statistical model.	Recommended	Not stated in two studies.
	Statistically significant responses noted.	Recommended	Not stated in three studies.
Hobbs *et al.*	Was each control and chemical concentration at least duplicated?	0 or 5	Not stated in one study, controls not stated in two studies.
	Was an appropriate statistical method or model used to determine the toxicity?	0 or 4	Not stated in two studies.
	For NOEC and LOEC data was the significance level 0.05 or less? or FOR LC and EC data was an estimate of variability provided?	0 or 4	Not stated in five studies.
Schneider *et al.*	Is the number of animals per group given?	Mandatory, 0, or 1	Not stated in one study.
	Are the statistical methods applied for data analysis given and applied in a transparent manner?	Recommended, 0, or 1	Not stated in three studies.
OECD guidelines	Statistical analysis and treatment of data; number of test chambers and replicates; number of embryos per replicate; if ANOVA has been used, the size of the effect which can be detected (*e.g.*, the least significant difference).	Mandatory	The non-standard studies were not evaluated according to the OECD reporting requirements.

has a "must" criterion: the number of animals has to be stated. Hobbs *et al*. [12] and Durda and Preziosi [11] instead ask for at least two replicates or sufficient sample size/replicates.

Result of the study evaluations

The statistical model used was not stated by two studies, results from the statistical calculations were missing for another one. The significance level or the estimation of

Table 11 OECDs reporting requirements, modified evaluation criteria/questions concerning biological effect, and summary of the evaluation results

Evaluation method	Evaluation criteria/question	Type of criteria/mark	Summary of evaluation results
Klimisch *et al.*	Determined effect concentrations (EC/LC/NOEC/LOEC).	Recommended	Not stated in one study but could be understood.
Durda and Preziosi	Quantitative measurement of response.	Recommended	Stated in all studies.
	Results reproduced by others.	Recommended	Not considered.
	Consistent with other findings.	Recommended	Not considered.
	Dose-response observed.	Recommended	Not stated in three studies, no relationship in one study.
Hobbs *et al.*	Was the biological effect stated (*e.g.*, LC or NOEC)?	0 or 5	Not stated in one study but could be understood.
	Was the biological effect quantified (*e.g.*, 50% effect, 25% effect)? The effect for NOEC and LOEC data must be quantified.	0 or 5	Info about tested concentrations is missing for one LOEC/NOEC value.
	Was there a concentration-response relationship either observable or stated?	0 or 4	Not stated in three studies, no relationship for one study.
Schneider *et al.*	Is the description of the study results for all endpoints investigated transparent and complete?	Recommended, 0, or 1	Not stated in four studies.
OECD guidelines	Concentration-response data, the slope of the dose-response curve and its standard error; EC_{50} or EC_{10} and associated confidence intervals; LOEC and NOEC and the statistical methods used for their determination; calculated response variables for each treatment replicate, with mean values and coefficient of variation for replicates.	Mandatory	The non-standard studies were not evaluated according to the OECD reporting requirements.

variability was missing for five studies. Eight of the nine studies used a sufficient sample size but none stated whether treatments were randomized or how observations were made. A possible reason for not reporting this could be that it is considered being self-evident, or that the researchers use traditional and widely accepted statistical methods.

Biological effect

Usefulness of proposed criteria

The reported aspects connected to "biological effect" are: stated and quantified effects for all endpoints, concentration-response relationship, and whether the results have been reproduced by others or are consistent with other findings. Durda and Preziosi [11] have covered all aspects. The OECD guidelines also recommend that calculated response variables for each treatment replicate, with mean values and coefficient of variation for replicates is reported.

Result of the study evaluations

All but one study had a determined EC_{50}, LOEC, NOEC or NOEL value (no observed effect level). There was no concentration-response relationship reported in one study and unclear in three other studies.

Summary of the evaluation of existing reliability evaluation methods

The four reliability evaluation methods are described and compared in Table 12. The results from the evaluation are discussed in three parts: Scope; How evaluation criteria are weighted and summarized; User friendliness. Conclusions from this evaluation are presented in the last section.

Scope

The four methods differ in their scope. Durda and Preziosi [11] have twice as many criteria as the other methods. Still, the four methods all include criteria from the same categories with the exception of Klimisch *et al*. [10] and Hobbs *et al*. [12] that lack criteria concerning controls and protocol, respectively. The Schneider *et al*. [13] method also includes aspects that are related to relevance.

The criteria vary in extent and specification, *e.g.*, Hobbs *et al*. [12] have one criterion in the test species category while Durda and Preziosi [11] use eight different criteria for the same issue (see Table 8). A disadvantage of using wide or unspecified criteria is that aspects could be forgotten about. A disadvantage of using too precise criteria is decreased flexibility.

Another example deals with criteria concerning dose/concentration. Hobbs *et al*. [12] and Klimisch *et al*. [10]

Table 12 Description and comparison of the four reliability evaluation methods

	Klimisch *et al*.	Durda and Preziosi	Hobbs *et al*.	Schneider *et al*.
Data types	Toxicity (in vivo and in vitro) and ecotoxicity (acute and chronic) data.	Ecotoxicity data.	Ecotoxicity (both acute and chronic) data.	Toxicity data (both in vivo and in vitro).
Coverage	Reliability.	Reliability.	Reliability.	Reliability and also a few aspects of relevance.
Evaluation categories	Reliable without restrictions, reliable with restrictions, not reliable and not assignable.	High, moderate and low quality, not reliable and not assignable.	High, acceptable and unacceptable quality.	Reliable without restrictions, reliable with restrictions, not reliable and not assignable.
No. of evaluation criteria/questions	12 (acute ecotoxicity), 14 (chronic ecotoxicity).	40	20	21
No. of aspects per criteria/questions	Several.	1	1	Several.
Type of criteria/questions	Recommended.	Recommended and mandatory.	Recommended, mark between 0 and 10.	Recommended and mandatory, mark between 0 and 1.
Additional guidance to evaluator	No.	Yes.	No.	Yes.
Information on how to summarize the evaluation	Not stated.	Stated.	Stated.	Stated and calculated automatically.
Additional information	Recommended in the REACH guidance document for industrial chemicals [39].	Based on standards from US EPA, OECD and ASTM.	Based on a method developed for the Australasian ecotoxicity database.	The method is called ToxRTool (Toxicological data reliability assessment Tool).
No. of OECD criteria that the method matched	14/37	22/37	15/37	14/37

US EPA United States Environmental Protection Agency, *OECD* Organisation for Economic Co-operation and Development, *ASTM* American Society for Testing and Materials.

do not explicitly state that the doses should be reported. The Schneider *et al.* [13] criteria could be interpreted in different ways; it is not clear whether nominal or measured concentrations are required.

There are also some examples of criteria where a minimum level is presented, *e.g.*, Hobbs *et al.* [12] criteria ask whether each control and chemical concentration is at least duplicated. Others leave more to the evaluator by asking if the sample size and replicates is sufficient (see [11]).

When it comes to acceptability criteria, *e.g.*, 10% mortality in the control, two different approaches can be seen. Either that the acceptability criteria are stated [12] or the requirement that the results connect to the acceptability criteria are reliable or acceptable [11,13]. The different approaches put different demands on the evaluator.

Durda and Preziosi [11], Hobbs *et al.* [12] and Schneider *et al.* [13] all have unique evaluation criteria that the other methods are missing.

The method described by Durda and Preziosi [11] has been developed from guidelines and it is therefore not surprising that this method has the highest resemblance with the OECD reporting requirements, 22 out of 37 criteria. The other three methods have each included less than half of the OECD reporting requirements in their list of criteria.

Some of the reporting requirements in the OECD guidelines were not reported by any of the four evaluation methods: date of start of the test; method of preparation of stock solutions; the recovery efficiency of the method; the limit of quantification in the test matrix; culture conditions; methods of collection; light quality; residual chlorine levels; TOC and COD; and the coefficient of variation. Several of these reporting requirements could be important for a general reliability evaluation.

How evaluation criteria are weighted and summarized

The four methods differ in how criteria are weighted and summarized.

Klimisch *et al.* [10] have not weighted their criteria and have not stated how to summarize the evaluation; this could result in a wide range of reliability evaluation results for the same test data. Klimisch *et al.* [10] has reserved the highest reliability category for studies carried out according to accepted guidelines and preferably performed according to good laboratory practice (GLP), or for methods that are very similar to a guideline method. This means that a study that has a design that is significantly different from standard test methods can never be put in the highest reliability category.

Durda and Preziosi [11] distinguish between mandatory and optional criteria. All mandatory criteria have to be fulfilled to receive the lowest acceptable reliability criteria and the highest reliability category is reserved for studies that fulfill all 40 evaluation criteria. GLP is preferred and the highest reliability category applies to standard tests or test closely related to standard tests. However, in practice, it is possible that non-standard tests fulfill all criteria, except the one regarding GLP, regardless of whether the test is closely related to a standard test or not.

Hobbs *et al.* [12] has weighted the criteria by assigning scores between 0 and 10. The total score for each study is then divided with the total possible score, which varies depending on what type of chemical, test organism and test media, and this results in a quality score and a quality class.

Schneider *et al.* [13] have divided their evaluation questions into mandatory and optional. All mandatory criteria have to be fulfilled to receive the lowest acceptable reliability category. The highest reliability category is reserved for studies which fulfill all mandatory evaluation questions and at least 18 of the 21 evaluation questions in total. This method differed from the other methods by, in our evaluation, assigning studies the highest reliability category (see the Summary of the reliability evaluation of the non-standard test data section).

User friendliness

User friendliness is defined in this study as a method that has clear instructions and an uncomplicated procedure.

All four methods include evaluation criteria that are wide and imprecise which opens up for a variety of different interpretations. Klimisch *et al.* [10] and Schneider *et al.* [13] compound criteria, *i.e.*, criteria that include several aspects. Having more delimited criteria/questions facilitates the evaluation since only one aspect at a time has to be considered.

Additional information that complements the evaluation criteria/questions makes an evaluation method easy to use. Both Durda and Preziosi [11] and Schneider *et al.* [13] have useful additional guidance.

The Klimisch method [10] lacks information how to summarize the evaluation. This complicates the work of the evaluator. Schneider [13] summarizes the results automatically in a pre-formatted excel-sheet. Both Durda and Preziosi [9] and Hobbs *et al.* [12] apply manual summarization of the evaluations.

Conclusions from the evaluation of existing reliability evaluation methods

The evaluation methods differ in their scope, user friendliness, and how criteria are weighted and summarized. Depending on the evaluators' previous experience

and knowledge, the outcome of the different methods can therefore differ. For the evaluators, it is important to be aware of the different methods strengths and limitations.

Durda and Preziosi [11] provide the method with the broadest scope and it also had the highest resemblance with the OECD guidelines. Durda and Preziosi [9], Hobbs *et al*. [12] and Schneider *et al*. [13] differ in how evaluation criteria are weighted and summarized but all three methods are functional and understandable. Durda and Preziosi [9] and Schneider *et al*. [13] both provide useful guidance information to the risk assessors which enhance the user friendliness.

Summary of the reliability evaluation of the non-standard test data

All four methods require some degree of expert judgment. They are developed to help risk assessors evaluate data, not to replace the risk assessor. Therefore it is likely that two experts evaluating the same study end up with slightly different results depending on their expertise and previous experiences. We have in our evaluation strived to make a uniform treatment of the evaluation methods and the selected studies. The evaluation method described by Klimisch *et al*. [10] requires more expert judgment when the evaluation is merged since instructions for this is lacking.

A striking result of this exercise is that many of the aspects considered important in the different evaluation methods are not reported by the authors of the selected studies. Examples of aspects often omitted are information about the controls, results from statistical evaluations, whether there is a dose-response relationship or not, tested concentrations, and clear description of the test environment. To safeguard against under-reporting, we recommend that a checklist containing all applicable reliability criteria should be used.

Overall the evaluation of the nine selected non-standard tests resulted in a low number of studies with acceptable reliability. The nine selected studies were evaluated by four different methods which resulted in 36 evaluations. Only 14 (39%) of these resulted in acceptable quality, reliable with restrictions, or reliable without restrictions (Table 13).

Also, the result from the four evaluation methods differed at a surprisingly high rate. Using the four methods lead to the same evaluation result for two studies only [14,15], both were summarized as studies with unacceptable quality/not reliable. The evaluation result differed by one quality data level, from unacceptable quality/not reliable to acceptable quality/reliable with restrictions, for five studies [16-20] and by two quality data levels, from unacceptable quality/not reliable to high quality/reliable without restrictions, for two studies [21,22] (Table 13).

Table 13 Summary of the reliability evaluation of non-standard test data

Reference	Evaluation method			
	Klimisch	Durda	Hobbs	Schneider
Andreozzi *et al.* [14]	-	-	-	-
Ferrari *et al.* [15]	-	-	-	-
Huggett *et al.* [16]	-	-	+	-
Robinson *et al.* [17]	+	-	+	-
Schmitt-Jansen *et al.* [18]	+	-	+	-
Quinn *et al.* [19]	-	-	+	-
Metcalfe *et al.* [20]	-	-	+	+
Nentwig [21]	+	-	+	++
Halm *et al.* [22]	+	-	+	++

- Unacceptable quality/not reliable, + acceptable quality/reliable with restrictions, ++ high quality/reliable without restrictions.

Durda and Preziosi [11] did not accept any of the studies since the mandatory criteria acceptable control mortality/morbidity was not reported. Hobbs *et al*. [12] has a similar criterion but since it is not mandatory it does not have the same effect on the summarized evaluations.

Other reasons why the reliability was considered unsatisfactory according to one or more evaluation method were lack of information about: chemical concentration control analysis, physical and chemical test conditions, specification of the test substance, a clear description of the test procedure, and the investigated period of the life cycle of the test organisms.

Discussion and conclusions

Standard test data are still preferred for regulatory environmental risk assessments of pharmaceutical substances. Accepting non-standard test data is likely to increase the regulatory agencies' work load since it is more complicated to evaluate these data compared to relying on standards only. More structured evaluation methods can help risk assessors and evaluators to use non-standard test data. But, as we have shown in this study, the design of the evaluation method is crucial since it can affect the outcome of the evaluation significantly.

The evaluation methods scrutinized in this study all require expert judgement. In our view, it is neither possible nor desirable to develop a method that completely leaves out expert judgement, but we can strive towards a method that reduces vagueness and elements of case-by-case interpretations. Both Hobbs *et al*. [12] and Schneider *et al*. [13] have tested and modified their respective method during the development process in order to increase the likelihood that evaluators arrive at similar conclusions.

The actual use of the four evaluation methods is unclear but Klimisch *et al*. [10] has been cited 62 times,

Schneider *et al*. [13] seven times, Durda and Preziosi [11] twice, and Hobbs *et al*. [12] once (ISI Web of Knowledge, 2010-11-02). The method described by Hobbs *et al*. [12] has, according to the authors themselves, been used by several Australian and New Zealand authorities.

The problem with non-standard data found to have low reliability could be either that the studies are poorly performed or that they are under-reported. Under-reporting could be a consequence from journals' desire to publish concise papers. However, an increasing number of journals provide possibilities to include additional data as supplementary electronic information which means that this should not be a major obstacle for making such information publicly available. It is also well known that environmental factors, such as oxygen saturation, salinity, pH, hardness, and temperature, can have drastic impact on uptake and effects of chemical substances [*e.g.*, [30-32]] and therefore, as proposed by most standard protocols, these aspects need to be monitored and reported in order to ensure the reproducibility of the test data, *i.e.*, a key issue in the scientific process.

For the nine studies investigated in the present paper, under-reporting could very well be a significant reason for the evaluation outcome. Aspects like use of controls, results from statistical evaluations, whether there are a concentration-response relationships or not, tested concentrations and clear description of the test subject and test environment are important and should be included in all publications presenting ecotoxicity data. Reliability evaluation methods can be used as checklists for authors and reviewers to ensure that all important aspects of the test method are included in their reports. A more structured reporting format could ensure the reliability of the test data without limiting the researcher's creativity in the design of a non-standard study.

As it is today, data with low reliability will not be included in regulatory environmental risk assessment. We believe that none of the nine selected non-standard studies could have been used in an environmental risk assessment of pharmaceuticals according to the EMA guideline. However, we still see that the studies could contribute to the risk assessment by acting as supporting information. We are currently developing a new suggestion on how to report and evaluate ecotoxicity data from the open scientific literature in regulatory risk assessment of pharmaceuticals. The new set of criteria is developed in collaboration between regulators at the German Federal Environment Agency (UBA) and researchers within the Swedish research program MistraPharma http://www.mistrapharma.se[33]. The criteria are based on the four methods evaluated in this study and the OECD reporting requirements, and have been further developed to include both reliability and relevance of test data. Intended users are risk assessors and researchers performing ecotoxicological experiments, but the criteria can also be used for education purposes and in the peer-review process for scientific papers. This approach intends to bridge the gap between the regulator and the scientist's needs and way of work.

It is important to remember that much of the research done within the field of ecotoxicology and risk assessment is financed through tax payer's money and to not find a way for use of this data in risk assessments would be an inefficient and irresponsible handling of resources.

Acknowledgements

The study was financed by Formas (the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning) and Mistra (the Foundation for Strategic Environmental Research).

Author details

[1]Department of Philosophy and the History of Technology, Royal Institute of Technology/Kungliga Tekniska Högskolan, Teknikringen 78B, 100 44 Stockholm, Sweden [2]Department of Applied Environmental Science, Stockholm University, Svante Arrhenius väg 8c, 106 91 Stockholm, Sweden

Authors' contributions

MÅ was the main contributor to this paper. MB and CR have participated in the design of the study, commented on the analysis, and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 14 February 2011 Accepted: 9 May 2011
Published: 9 May 2011

References

1. European Medicines Agency (EMA), Committee for medicinal products for human use (CHMP): **Guideline on the environmental risk assessment of medicinal products for human use.** 2006, Ref EMEA/CRMP/SWP/4447/00.
2. OECD: **OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring. No. 1 OECD Principles of Good Laboratory Practice.** Paris: Organization for Economic and Co-operative Development; 1998.
3. vom Saal F, Myers JP: **Good Laboratory Practices Are Not Synonymous with Good Scientific Practices, Accurate Reporting, or Valid Data.** *Environmental Health Perspective, Perspectives Correspondence* 2010, **118(2)**:A60.
4. Myers JP, vom Saal FS, Akingbemi BT, Arizono K, Belcher S, Colborn T, Chahoud I, Crain DA, Farabollini F, Guillette LJ, Hassold T, Ho S, Hunt PA, Iguchi T, Jobling S, Kanno J, Laufer H, Marcus M, McLachlan JA, Nadal A, Oehlmann J, Olea N, Palanza P, Parmigiani S, Rubin BS, Schoenfelder G, Sonnenschein C, Soto AM, Talsness CE, Taylor JA, Vandenberg LN, Vandenbergh JG, Vogel S, Watson CS, Welshons WV, Zoeller RT: **Why public health agencies cannot depend on good laboratory practices as a criterion for selecting data: the case of bisphenol A.** *Environmental Health Perspective* 2009, **117**:309-315.
5. Gunnarsson L, Jauhiainen A, Kristiansson E, Nerman O, Larsson DG: **Evolutionary conservation of human drug targets in organisms used for environmental risk assessments.** *Environmental Science and Technology* 2008, **42(15)**:5807-5813.
6. Molander L, Ågerstrand M, Rudén C: **WikiPharma–a freely available, easily accessible, interactive and comprehensive database for environmental effect data for pharmaceuticals.** *Regulatory Toxicology and Pharmacology* 2009, **55**:367-371.
7. Mattson B: **A voluntary environmental classification system for pharmaceutical substances.** *Drug Information Journal* 2007, **41(2)**:187-91.
8. Breitholtz M, Lundström E, Dahl U, Forbes V: **Improving the Value of Standard Toxicity Test Data in REACH.** In *Regulating Chemical Risks:*

European and Global Challenges. Edited by: Eriksson J, Gilek M, Rudén C. Dordrecht: Springer; 2010:85-98.
9. European Commission: **European Commission Technical Guidance Document in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances and Commission Regulation (EC). No 1488/94 on Risk Assessment for Existing Substances, Part II.** 2003 [http://ecb.jrc.ec.europa.eu/documents/TECHNICAL_GUIDANCE_DOCUMENT/tgdpart2_2ed.pdf].
10. Klimisch HJ, Andreae M, Tillmann U: **A systematic approach for evaluating the quality of experimental toxicological and exotoxicological data.** *Regulatory toxicology and pharmacology* 1997, **25**:1-5.
11. Durda JL, Preziosi DV: **Data quality evaluation of toxicological studies used to derive ecotoxicological benchmarks.** *Human and ecological risk assessment* 2000, **6(5)**:747-765.
12. Hobbs DA, Warne MStJ, Markich SJ: **Evaluation of criteria used to assess the quality of aquatic toxicity data.** *Integrated environmental assessment and management* 2005, **1(3)**:174-180.
13. Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, Hartung T, Hoffmann S: **"ToxRTool", a new tool to assess the reliability of toxicological data.** *Toxicology Letters* 2009, **189(2)**:138-144.
14. Andreozzi R, Caprio V, Ciniglia C, de Champdor M, Giudice RL, Marotta R, Zuccato E: **Antibiotics in the Environment: Occurrence in Italian STPs, Fate, and Preliminary Assessment on Algal Toxicity of Amoxicillin.** *Environmental Science and Technology* 2004, **38(24)**:6832-6838.
15. Ferrari B, Mons R, Vollat B, Fraysse B, Paxéus N, Giudice RL, Pollio A, Garric J: **Environmental risk assessment of six human pharmaceuticals: Are the current environmental risk assessment procedures sufficient for the protection of the aquatic environment?** *Environmental Toxicology and Chemistry* 2004, **23(5)**:1344-1354.
16. Huggett DB, Brooks BW, Peterson B, Foran CM, Schlenk D: **Toxicity of Select Beta Adrenergic Receptor-Blocking Pharmaceuticals (B-Blockers) on Aquatic Organisms.** *Archives of Environmental Contamination and Toxicology* 2002, **43**:229-235.
17. Robinson AA, Belden JB, Lydy MJ: **Toxicity of fluoroquinolone antibiotics to aquatic organisms.** *Environmental toxicology and Chemistry* 2005, **24(2)**:423-430.
18. Schmitt-Jansen M, Bartels P, Adler N, Altenburger R: **Phytotoxicity assessment of diclofenac and its phototransformation products.** *Analytical and Bioanalytical Chemistry* 2007, **387**:1389-1396.
19. Quinn B, Gagné F, Blaise C: **An investigation into the acute and chronic toxicity of eleven pharmaceuticals (and their solvents) found in wastewater effluent on the cnidarian, Hydra attenuata.** *Science of the Total Environment* 2008, **389**:306-314.
20. Metcalfe CD, Metcalfe TL, Kiparissis Y, Koenig BG, Khan C, Hughes RJ, Croley TR, March RE, Potter T: **Estrogenic potency of chemicals detected in sewage treatment plant effluents as determined by in vivo assays with Japanese medaka (Oryzias latipes).** *Environmental Toxicology and Chemistry* 2001, **20(2)**:297-308.
21. Nentwig G: **Effects of pharmaceuticals on aquatic invertebrates. Part II: The antidepressant drug fluoxetine.** *Archives of Environmental Contamination and Toxicology* 2007, **52**:163-170.
22. Halm S, Pounds N, Maddix S, Rand-Weaver M, Sumpter JP, Hutchinson TH, Tyler CR: **Exposure to exogenous 17ß-oestradiol disrupts P450aromB mRNA expression in the brain and gonad of adult fathead minnows (Pimephales promelas).** *Aquatic Toxicology* 2002, **60**:285-299.
23. Mattson B: **A voluntary environmental classification system for pharmaceutical substances.** *Drug Information Journal* 2007, **41(2)**:187-191.
24. Ågerstrand M, Rudén C: **Evaluation of the accuracy and consistency of the Swedish Environmental Classification and Information System for pharmaceuticals.** *Science of the Total Environment* 2010, **408**:2327-2339.
25. van der Hoven N: **How to measure no effect. Part III. Statistical aspects of NOEC, ECx and NEC estimates.** *Environmetrics* 1997, **8**:225-261.
26. Suter GW: **Abuse of hypothesis testing statistics in ecological risk assessment.** *Human and Ecological Risk Assessment* 1996, **2(2)**:331-347.
27. Newman MC: **"What exactly are you inferring?" A closer look at hypothesis testing.** *Environmental Toxicology and Chemistry* 2008, **27**:1013-1019.
28. Warne MStJ, Van Dam R: **NOEC and LOEC data should no longer be generated or used.** *Australasian Journal of Ecotoxicology* 2008, **14(1)**:1.
29. van Leeuwen CJ, Vermeire TG, Eitors: **Risk Assessment of Chemicals: An Introduction.** Berlin: Springer; 2007, ISBN1402061013.
30. Chapman PM, Wang FY, Janssen C, Persoone G, Allen HE: **Ecotoxicology of metals in aquatic sediments: binding and release, bioavailability, risk assessment, and remediation.** *Canadian Journal of Fisheries and Aquatic Sciences* 1998, **55(10)**:2221-2243.
31. Heugens EHW, Jager T, Creyghton R, Kraak MHS, Hendriks AJ, Van Straalen NM, Admiraal W: **Temperature-dependent effects of cadmium on Daphnia magna: accumulation versus sensitivity.** *Environmental Science and Technology* 2003, **37(10)**:2145-2151.
32. Gardeström J, Elfwing T, Lof M, Tedengren M, Davenport JL, Davenport J: **The effect of thermal stress on protein composition in dogwhelks (Nucella lapillus) under non-normoxic and hyperoxic conditions.** *Comparative Biochemistry and Physiology A–Molecular and Integrative Physiology* 2007, **148(4)**:869-875.
33. Ågerstrand M, Küster A, Bachmann J, Breitholtz M, Ebert I, Rechenberg B, Rudén C: **Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals.**.
34. Schering's report A12518: **Environmental risk assessment of ethinyletsradiol.** Environmental Pollution; 2011 [http://www.fass.se], Available through the Swedish environmental classification and information system for pharmaceuticals website:.
35. Bogers R, Mutsaerds E, Druke J, Roode DF, de Murk AJ, Burg B, van der Legler J: **Estrogenic endpoints in fish early life-stage tests: luciferase and vitellogenin induction in estrogen-responsive transgenic zebrafish.** *Environmental Toxicology and Chemistry* 2006, **25(1)**:241-247.
36. Schering-Plough: **Environmental risk assessment of ethinyletsradiol.** [http://www.fass.se], Available through the Swedish environmental classification and information system for pharmaceuticals website:.
37. Schäfers C, Teigeler M, Wenzel A, Maack G, Fenske M, Segner H: **Concentration- and time-dependent effects of the synthetic estrogen, 17a-ethinylestradiol, on reproductive capabilities of the szebrafish, Danio rerio.** *Journal of Toxicology and Environemntal Health Part A* 2007, **70**:768-779.
38. Segner H, Navas JM, Schäfers C, Wenzel A: **Potencies of estrogenic compounds in in vitro screening assays and in life cycle tests with zebrafish in vivo.** *Ecotoxicology and Environmental Safety* 2003, **54**:315-322.
39. ECHA (European Chemicals Agency): **Guidance information for the implementation of REACH. Guidance on information requirements and chemical safety assessment.** 2008, Chapter R.4: Evaluation of available information.

doi:10.1186/2190-4715-23-17
Cite this article as: Ågerstrand *et al.*: 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* 2011 **23**:17.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com