640 Received: 8 July 2016 | Returned for Revision: 1 September 2016 | Accepted: 17 November 2016

Health & Ecological Risk Assessment

Assessing the Reliability of Ecotoxicological Studies: An Overview of Current Needs and Approaches

Caroline Moermond,† Amy Beasley,‡ Roger Breton,§ Marion Junghans,‖ Ryszard Laskowski,# Keith Solomon,†† and Holly Zahner‡‡*

†Centre for Safety of Substances and Products, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
‡The Dow Chemical Company, Toxicology and Environmental Research and Consulting, Midland, Michigan, USA
§Intrinsik, Ottawa, Ontario, Canada
‖Swiss Centre for Applied Ecotoxicology Eawag-EPFL, Dübendorf, Switzerland
#Institute of Environmental Sciences, Jagiellonian University, Kraków, Poland
††Centre for Toxicology, School of Environmental Science, University of Guelph, Guelph, Ontario, Canada
‡‡US Food and Drug Administration, Center for Veterinary Medicine, Rockville, Maryland

EDITOR'S NOTE:

This is 1 of 4 companion articles resulting from a SETAC Pellston Workshop® on "Improving the Usability of Ecotoxicology in Regulatory Decision-Making," held August 2015 in Shepherdstown, West Virginia, USA. The main workshop objectives were to improve the reliability and reproducibility of ecotoxicity studies, improve the use of peer-reviewed studies in regulatory risk assessment of chemicals, and improve the methods used in risk assessments when evaluating single or multiple lines of evidence.

ABSTRACT

In general, reliable studies are well designed and well performed, and enough details on study design and performance are reported to assess the study. For hazard and risk assessment in various legal frameworks, many different types of ecotoxicity studies need to be evaluated for reliability. These studies vary in study design, methodology, quality, and level of detail reported (e.g., reviews, peer-reviewed research papers, or industry-sponsored studies documented under Good Laboratory Practice [GLP] guidelines). Regulators have the responsibility to make sound and verifiable decisions and should evaluate each study for reliability in accordance with scientific principles regardless of whether they were conducted in accordance with GLP and/or standardized methods. Thus, a systematic and transparent approach is needed to evaluate studies for reliability. In this paper, 8 different methods for reliability assessment were compared using a number of attributes: categorical versus numerical scoring methods, use of exclusion and critical criteria, weighting of criteria, whether methods are tested with case studies, domain of applicability, bias toward GLP studies, incorporation of standard guidelines in the evaluation method, number of criteria used, type of criteria considered, and availability of guidance material. Finally, some considerations are given on how to choose a suitable method for assessing reliability of ecotoxicity studies. *Integr Environ Assess Manag* 2017;13:640–651. © 2016 The Authors. *Integrated Environmental Assessment and Management* published by Wiley Periodicals, Inc. on behalf of Society of Environmental Toxicology & Chemistry (SETAC)

Keywords: Hazard assessment Risk assessment Quality evaluation Literature evaluation Reliability assessment

INTRODUCTION

Regulators have the responsibility to make sound and verifiable risk decisions to protect the environment. They often need to screen through a wealth of ecotoxicity studies to determine which studies to use as best available data for risk assessment. When doing so, they need to evaluate many different types of ecotoxicity studies, which vary in study design, methodologies, quality, and level of detail reported (e.g., reviews, peer-reviewed research papers, or industry-sponsored studies docume nted under Good Laboratory Practice [GLP] guidelines). These studies may be conducted using either standard methods or nonstandard methods, which adds to the complexity of the evaluations. Each study

* Address correspondence to caroline.moermond@rivm.nl

Published 21 November 2016 on wileyonlinelibrary.com/journal/ieam.

should be evaluated in accordance with scientific principles and allow for verification of results regardless of whether they were conducted in accordance with GLP and/or standardized methods. A systematic and transparent approach, as well as expert judgment, is needed to evaluate these studies (Ågerstrand, Küster et al. 2011; Bevan and Strother 2012; Beronius et al. 2014; Moermond et al. 2016). This will ensure consistency and reproducibility of the evaluations, which are key factors in public acceptance (Bevan and Strother 2012; Ågerstrand et al. 2014).

The aim of this study is to identify, describe, and compare methods for determining reliability of ecotoxicity data for regulatory decision making and risk assessment.

BACKGROUND

Both prospective risk assessment (marketing authorization applications) and retrospective risk assessment (derivation of environmental criteria or quality standards) are based on effects data generated by ecotoxicity studies. Because of the consequences of these assessments, these studies should be of sufficient quality on which to base a decision. Deficiencies in (nonecotoxicity) studies conducted in support of drug testing became a concern for the US Food and Drug Administration in 1975. These deficiencies ranged from poor recordkeeping to reports based on nonexistent tests (Seiler 2005). As a result, the US Food and Drug Administration developed GLP regulations, which stimulated similar activities in the US Environmental Protection Agency (USEPA), in agencies in other countries, and at the Organisation for Economic Co-operation and Development (OECD). GLP standards require that the protocol is fully documented, as are any deviations from the protocol, and that all raw data are available. Coupled with the development of GLP standards was the development of standardized test guidelines, for instance, by the USEPA and OECD. However, GLP and/or standard guidelines are not a guarantee that the correct hypothesis, experimental design, or most appropriate species is tested. In addition, they do not ensure that all relevant adverse responses (end points) for a given substance are tested (Beronius et al. 2014), and they may be inflexible (Ågerstrand, Breitholtz M, and Rudén 2011; Ågerstrand, Küster et al. 2011; Ågerstrand et al. 2014). However, results from nonstandardized studies reported in peer-reviewed research papers may, in some cases, contribute additional and important information to a risk assessment and should not necessarily be excluded from risk assessment simply because the study was not performed according to GLP and/or standardized guidelines. A peer-review study that followed nonstandard methods can be scientifically valid without GLP compliance; however, peer-review of these studies does not guarantee that the results are of sufficient quality (McCarty et al. 2012; Moermond et al. 2016).

Many methods for evaluating (eco)toxicological studies have been developed to assess the quality of studies used for risk assessment. The first of these was the approach suggested by Klimisch et al. (1997). Subsequently, several other methods for evaluating toxicological and ecotoxicological studies have been proposed by other government agencies, such as the USEPA (2011), the Dutch National Institute for Public Health and the Environment (RIVM) (Mensink et al. 2008), the European Food and Safety Authority (EFSA) (2015), and other researchers. A number of these methods (Klimisch et al. 1997; Durda and Preziosi 2000; Hobbs et al. 2005; Schneider et al. 2009; Breton et al. 2009; USEPA 2011; Van Der Kraak et al. 2014; Beasley et al. 2015; Moermond et al. 2016) are compared in this study.

Reliability, relevance and weight of evidence

Several terms are used in the literature and reports to describe how well a study is conducted (Table 1). For the purposes of consistency with existing uses, we have used the term *reliability* throughout this study. We used the following modification of Klimisch et al. (1997) to describe our definition of reliability and to explicitly include studies published in the open literature in the process: Reliability is the inherent quality of an effect value in a test report or publication relating to: 1) a clearly described experimental design to allow for the study to be repeated independently, 2) the way the experimental procedures were performed, and 3) the reporting of the results to provide evidence of the reproducibility and accuracy of the findings.

In this study, we assume that studies are assessed to be categorized as follows: 1) reliable, 2) not reliable, or 3) not assignable. These 3 categories were chosen to align with those reported by Klimisch et al. (1997) because they are commonly used by assessors. However, the definitions for each category have been redefined in this study. In general, reliable studies are well designed and well performed, and they report enough details on study design and performance to assess the study. A study categorized as "not reliable" has clear flaws in study design and/or how it was performed. A study may be categorized as "not assignable" when information on one or more vital parameters needed to make an assessment of the study is missing or insufficient (Klimisch et al. 1997; Moermond et al. 2016). In many frameworks, studies that are categorized as "not assignable" are not used for hazard and risk assessment, similar to studies that are categorized as "not reliable."

Every hazard and/or risk assessment of a substance starts with a problem formulation phase, in which the compound is characterized, protection goals are identified, and hypotheses to be tested are defined. After this phase, and irrespective of the purpose of the evaluation, a number of steps should be considered (Figure 1). First, the assessor must choose the most appropriate evaluation method for determining suitability of a study for use in hazard and/or risk assessment. This is critical because data used in hazard and/or risk assessment require a measure of quality. This requires consideration of how well a study was conducted (i.e., reliability) and how relevant the observations are to the question (i.e., relevance). The former is discussed in this study and the latter in a companion paper (Rudén et al. this issue). When used in hazard and/or risk assessment, reliability and relevance are often assessed separately,

 DOI: 10.1002/ieam.1870

Table 1. Definitions and synonyms of terms that are commonly used in assessment of reliability of data

Term	Definition	Synonyms
Categorize (~categories)	Approach in which effect values are sorted into distinct categories during assessment of reliability (see also Klimisch categories). No *numerical scoring* is applied to reliability categories; categories may be codes, letters, or numbers but have no numerical meaning. Therefore, when using reliability categories, no mathematical operations can be performed; i.e., reliability categories cannot be easily transformed into scores (*numerical scoring*).	Pigeonhole Classification Group Rank
Confidence	Confidence is a combination of *reliability* and *relevance* of the *effect values* in the study with regard to the problem formulation.	Adequacy Strength Weight of evidence
Critical criteria	This refers to criteria identified as critical or important to consider when conducting an assessment of reliability. Every method for reliability assessment has critical criteria identified (sometimes depending on the substance or organism tested), but they should not be confused with *exclusion criteria*. Critical criteria may have greater weight than other criteria because they are regarded as more important, e.g., criteria relating to analytical verification of exposure concentrations.	Red criteria Key criteria
Effect value	The effect value is the exposure concentration or dose that relates to the response in a toxicity test, for example, No observed effect concentration (NOEC) and median effect concentration (EC50) values. NOEC and EC50 values.	Observation Effect measure Test value Test end point Response
Endpoint	Biological process studied with regard to its susceptibility to the substance assessed in the bioassay, e.g., survival, growth, or reproduction.	Response
Exclusion criteria	Exclusion criteria include any criterion that by itself can lead to the exclusion of a given effect value or a study from further consideration. The decision regarding which exclusion criteria apply and whether they should be used depends on the problem formulation of the assessment and *expert judgment,* and is usually made before studies are assessed for reliability and relevance. Some assessment methods have no exclusion criteria; other methods exclude, for instance, all review articles, posters, and summaries.	Gatekeeper
Expert judgment	Expert judgment is a judgment and/or decision made by an expert (e.g., scientist or regulator) based upon expertise, experience, and knowledge of a specific area of interest.	Professional judgment
Klimisch categories	Klimisch categories have been identified by Klimisch et al. (1997) as follows: 1 = reliable, 2 = reliable with restrictions, 3 = unreliable, and 4 = unassignable. These categories have found widespread use in current guidance documents within the field of risk assessment. Therefore, some of the later developed reliability assessment methods have maintained this basic 4-category system. Hence the use of these 4 categories does not necessarily imply that the methods of the Klimisch procedure for assessment of reliability are being used.	Reliability categories K1-K4 R1-R4 Ri1-Ri4
Numerical scoring (~scores)	Approach in which the number of fulfilled reliability criteria is used to assign a reliability score to each assessed *effect value*. It results in a continuous canonical numerical value. The scoring approach allows for performing mathematical operations. The results of the scoring approach may be used to feed into a categorization approach. This transformation often explicitly considers *expert judgment* and the *weighting of criteria,* and hence turns the originally quantitative score into a qualitative assessment.	Scoring approach, system, or method

(*Continued*)

Table 1. (*Continued*)

Term	Definition	Synonyms
Problem formulation	Problem formulation, as defined by the USEPA (1998), is a planning and scoping process that establishes the goals, breadth, and focus of the risk assessment.	
Relevance (~relevant)	Relevance is "covering the extent to which data and/or tests are appropriate for a particular hazard identification or risk characterization" (Klimisch et al. 1997, p 2). Thus, where *reliability* of an *effect value* is the same for every assessment, the relevance of that effect value depends on the purpose of the assessment. Thus, the same effect value may be relevant for one assessment, but not for another.	Significance Importance Applicability Germaneness Weight Appropriateness
Reliability (~reliable)	Reliability is the inherent quality of an *effect value* in a test report or publication relating to: 1) a clearly described experimental design to allow for the study to be repeated independently, 2) the way the experimental procedures were performed, and 3) the reporting of the results to provide evidence of the reproducibility and accuracy of the findings (modified from Klimisch et al. 1997).	Quality Strength Validity Consistency
Response	Response refers to the attributes of the organisms that are quantified in the test or bioassay, e.g., survival, number of progeny produced, or growth rate.	Test endpoint
Tiered approach	This method consists of a series of steps in which data are progressively excluded from further consideration, leaving only the data with highest *reliability* and *relevance* for further assessment. This approach is usually intended to minimize the effort while focusing on the more relevant information.	
Not assignable	This refers to *effect values* for which not enough information is reported to be classified into *categories* or fed into a *numerical scoring* system. This categorization may change if more data become available, e.g., through author communication. Hence, it should not be confused with *not reliable*.	Unassignable
Not reliable	*Effect values* assessed as "not reliable" have obvious flaws in study design and/or how it was performed. Missing information on *key criteria* is not sufficient to assign a study or effect value as unreliable. These *effect values* would then qualify as "not assignable."	Unreliable
Weighting of criteria	Weighting of criteria assigns different importance to certain criteria. Under *numerical scoring* approaches this might mean that the most important criteria get the highest scores. The weighting should be consistent with the *problem formulation* and may be based on *expert judgment*.	

The definitions given are a result of discussions among the authors. The synonyms are intended to relate to other existing terminology.

but for the final assessment they must be considered together, because they are mutually inclusive (Figure 1). Reliability concerns the intrinsic quality of the study. Because of this, the evaluation of reliability should not depend on the goal for which the study is assessed. This is in contrast to the evaluation of relevance, where the goal of the assessment determines whether or not a study may be used for hazard and/or risk assessment.

In the next step, a search for studies and reports is performed, which are screened for their general applicability to the problem formulation and/or goal of the assessment. This first screening is usually based on the title and/or abstract. After this, a more extensive study evaluation is performed, regarding both reliability and relevance of the individual effect values. In some cases, different responses (e.g., mortality and behavior) from the same study may be evaluated separately regarding their reliability and relevance. Depending on the purpose of the assessment and the amount of data available, it may be appropriate to use the results of the evaluation of reliability and relevance in a weight-of-evidence (WoE) approach. Finally, the results (i.e., the data found to be reliable and relevant) are used for hazard and/or risk assessment. This study compares a number of evaluation methods and provides considerations when choosing a method for reliability assessment.

 DOI: 10.1002/ieam.1870

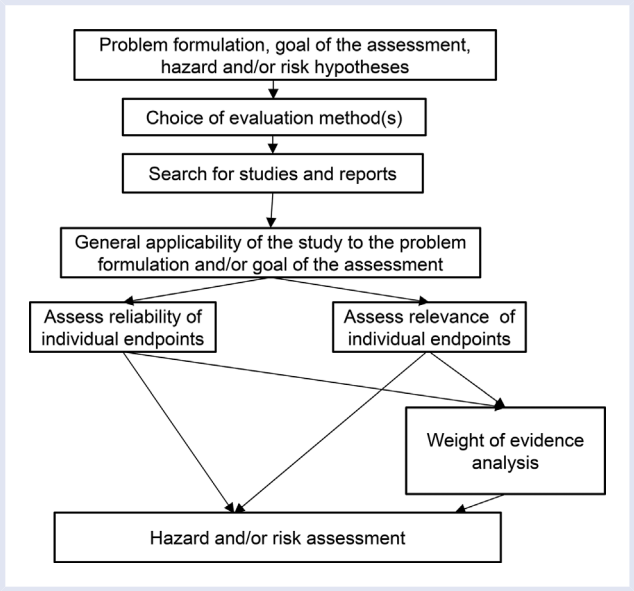

Figure 1. General illustration of the overall framework for assessment of reliability of data for hazard and risk assessment. This includes a weight-of-evidence analysis when appropriate. This study mainly addresses reliability assessments.

Types of problems encountered when evaluating published ecotoxicity studies

When a peer-reviewed research paper is used in a regulatory framework, a company applying for marketing authorization or approval of a substance and the risk assessor from a regulatory office often have to conduct a reliability assessment. In the European Union and other joint evaluation efforts, different assessors from many European Union Member States or countries might have to assess the same study or at least agree on the outcome of the assessment of the available data. This may lead to very different evaluation results, particularly when the selection of the evaluation method is not clear or if there is no recommended evaluation method for a specific regulatory framework. A lack of transparency regarding the results of these assessments and a lack of harmonization in the methods used can lead to different assessments for the same compound, resulting in inconsistent assessments across frameworks (Schenk 2010; Ågerstrand, Breitholtz, and Rudén 2011; Bevan and Strother 2012; Beronius et al. 2014; Moermond et al. 2016).

Bookkeeping systems are sometimes used in some regulatory agencies. In these systems, reliability is categorized depending on predetermined levels of compliance with reliability criteria (e.g., if 60% of all criteria are met, a study is assessed to be reliable). These systems may give more weight to certain criteria, but they are often very general and do not consider the particular test requirements of a compound or test organism. This increases the chance of miscategorization. When an evaluation method is used that relies on expert (professional) judgment and experience, the chance of a study being miscategorized might decrease, although the opportunity for disparities based on differences of opinions between experts may be introduced. Systems that are based on expert judgment need clear guidance, especially when they are also used by less experienced assessors. When these systems are used, transparency on the approach needs to be provided.

In addition, a lack of clear guidance on how to evaluate studies and which characteristics of studies to evaluate often leads to a tendency toward accepting standard guideline and/or GLP studies as best available data for risk assessment by some assessors (Moermond et al. 2016). Several researchers have concluded that this may arise because peer-reviewed research papers, as currently published, often lack details to allow a full reliability assessment to be completed (McCarty et al. 2012; Moermond et al. 2016). Also, studies that are categorized as "not assignable" by one assessor because of a lack of data could be categorized as "not reliable" by another assessor, thereby increasing inconsistency (Moermond et al. 2016). For example, review articles and handbooks often provide insufficient experimental details to evaluate the reliability of the effect values provided by their sources. Moermond et al. (2016) proposed to categorize effect values from such studies as "not assignable." These studies may benefit from reassessment when additional information has been obtained from the original study or authors, leading to a new categorization as "reliable" or "not reliable."

If peer-reviewed research papers do not provide details on the design, methods, and results, then such studies will be categorized as "not assignable" more often than GLP and/or guideline studies, which may result in the exclusion of useful peer-reviewed research papers from use in risk assessments. To enhance the usability of peer-reviewed research papers for risk assessment, reporting recommendations should be followed, and more details need to be reported using supplemental information (Hanson et al. 2016; Moermond et al. 2016). In addition, directed efforts to obtain needed information from the authors could be considered as an integral part of the risk-assessment protocol.

EVALUATION METHODS

To compare a number of currently available methods for assessing reliability of studies (Klimisch et al. 1997; Durda and Preziosi 2000; Hobbs et al. 2005; Breton et al. 2009; Schneider et al. 2009; USEPA 2011; Van Der Kraak et al. 2014; Beasley et al. 2015; Moermond et al. 2016), we compiled a summary table (Table 2) of their attributes. The methods chosen for this comparison were developed by different researchers. When more than 1 method was developed by the same group or researcher, the most recent method is used for this comparison. The attributes in Table 2 are discussed later in relation to how they can be used for hazard and/or risk assessment. Where applicable, special attention was directed to the applicability of these attributes to WoE. Some of these methods include a tiered approach that consists of first identifying publications or studies for review to be put into a data set (USEPA 2011; Beasley et al. 2015) and then performing relevance and reliability assessments (Klimisch et al. 1997; USEPA 2011; Van Der Kraak et al.

Table 2. Summary of the methods used to categorize reliability of papers and reports on toxicity studies

		Toxicity studies								
Attributes of the method	Type	Klimisch et al. (1997)	Durda and Preziosi (2000)	Hobbs et al. (2005)	Schneider et al. (2009)	Breton et al. (2009)	USEPA (2011)	Van Der Kraak et al. (2014)	Beasley et al. (2015)	Moermond et al. (2016)
Categories or numerical scoring	General	Categories	Categories	Numerical	Yes, but classification is based on numerical scale (i.e., K1 $\geq$ 80%), automatic classification can be overruled by expert judgment	Categories based on scoring	Categories	Scoring	Not applicable	Categories
Tiered method	General	No	Can be used as a 2-tiered approach	No	Yes, 2 steps	No	Yes	No	Yes, 4 tiers	No
Evaluation categories	General	4 categories: reliable without restrictions, reliable with restrictions, unreliable, and unassignable	5 descriptive categories	No, list of criteria	The first 3 Klimisch categories	Yes, 4 categories	Yes, 3 categories	Continuous value from 0 to 4	Yes, 4 tiers	Yes, 20
Use of exclusion or critical criteria	General	No	Yes, critical criteria determined using "must" terminology	No	Yes, "red criteria"	No	Yes	Few exclusion criteria used because aim was a nonbiased weight of evidence	Criteria are user defined	Yes, depending on problem formulation
Weighting of criteria	General	No	No	Yes	No	Yes	No	Some (critical) criteria have more weight than others but can be modified by expert judgment	No, equally weighted	No
Tested on case studies	General	No	Yes, multiple examples	Yes, used to evaluate 2 articles	Yes	Yes, 614 studies with 20% of studies rescored by another assessor	No	Not via a ring test, but large number of papers assessed	Case study provided	Yes, ring test with 8 articles and 75 participants
Applicable to tests in soils and sediments	General	Not specifically	No	No	No	No	Yes	Yes, with modification of general criteria	Yes, with appropriate criteria choice	No, but can relatively easily be rewritten to do so
Applicable to studies using quantitative structure activity relationship, genotoxicity, or in vitro procedures	General	No	No	No	Yes (in vitro)	No	No	Yes, with modification of general criteria	Yes, with appropriate criteria choice	No

(*Continued*)

Integr Environ Assess Manag 2017:640–651 DOI: 10.1002/ieam.1870

Table 2. (*Continued*)

Attributes of the method	Type	Toxicity studies								
		Klimisch et al. (1997)	Durda and Preziosi (2000)	Hobbs et al. (2005)	Schneider et al. (2009)	Breton et al. (2009)	USEPA (2011)	Van Der Kraak et al. (2014)	Beasley et al. (2015)	Moermond et al. (2016)
Bias toward GLP studies	Data	Yes	No	No	No	GLP and QA is a greater-weight criterion, but not a critical criterion	No	GLP and QA is a greater-weight criterion, but not a critical criterion	None	No
Applicable to guideline and/or nonguideline studies	Data	Yes	Yes	Yes	Yes	All studies can be evaluated, but evaluation scheme is designed for OECD test guidelines	Yes	Yes	Yes, all studies can be evaluated	Yes, any type
Applicable to data studies in other areas	Data	Human toxicology and ecotoxicology	Ecotoxicology	Aquatic ecotoxicology only	Human health	Ecotoxicology OECD 201-203	Ecotoxicology, aquatic and terrestrial	Ecotoxicology	Criteria can be chosen for any type of data	Aquatic ecotoxicology
Usable for large data sets (high throughput)	Data	Yes	Yes	Yes, but may be complex	Yes	Yes	Yes	Usable for all data sets, but better for large numbers of papers	Usable for single-study or high-throughput scenarios	Yes
Amount of guidance provided	Guidance	In general	Detailed	Detailed	Detailed	Available through supplemental information	Good considerations, poor guidance for classification	Guidance provided for criteria and scoring	General guidance for criteria selection	Detailed guidance for each criterion
Incorporation of standard guidelines in the scheme	Criteria	Yes, in general; no specific validity criteria included in method	No	No	Yes	Yes	No	Not specific, but generally similar	OECD or standard guideline criteria can be incorporated	Yes, OECD validity criteria are used for the criteria or in the guidance with the criteria
Number of criteria used	Criteria	Depending on type of study	30	20	19 mandatory	Depending on OECD guidelines	Criteria listed, primarily based on expert judgment	3 key criteria and infinite general comments with a focus on weaknesses	User defined	20
Consideration of bias	Criteria	No	No	No	May be considered on a case-by-case basis	No	Yes, skewed toward expert judgment	Transparent reporting of evaluation reduces bias, no exclusion bias	Transparent reporting of evaluation reduces bias, no exclusion bias	Yes
Criteria used for evaluation of reliability										
Design	Criteria	Only in relation to the closest	Yes	Yes	Yes	Yes	No	Yes	User defined	Yes

		guideline								
Test organism	Criteria	Yes	Yes	Yes	Yes	Yes	Yes	Yes	User defined	Yes
Chemical characterization	Criteria	Yes	Yes	Yes	Yes	Yes	Yes	Yes	User defined	Yes
Exposure conditions	Criteria	Yes	Yes	Yes	Yes	Yes	Yes	Yes	User defined	Yes
Statistics	Criteria	Yes, but defined only in the guideline of reference	Yes	Yes	Yes, but only basic with regard to documentation	Yes	Yes	Yes	User defined	Yes
Reporting of results	Criteria	Yes	Yes	Yes	Yes	Yes	No	Yes	User defined	Not as a criterion, but discussed in text
Reporting of evaluation results										
Guidance given on how to summarize the evaluation	Reporting	In IUCLID datasheets	Yes	Yes	Yes	Expert judgment	Yes	Yes	Only studies passing all 4 steps are selected	Yes, extensive guidance in paper
Distinct form/tool?	Reporting	No	Yes	Yes	Yes, ToxRTool	Yes	Yes	Yes	Yes, 4-tier diagram can serve as tool	Yes, an Excel spreadsheet is provided
Transparency and/or reporting of evaluation results	Reporting	No	Yes	Yes	Yes	Yes, but Environment Canada evaluations are confidential	Yes	Yes	Yes	Yes

OECD = Organisation for Economic Co-operation and Development; QA = Quality Assurance.

DOI: 10.1002/ieam.1870

2014; Beasley et al. 2015; Moermond et al. 2016), whereas other methods focus only on reliability assessment (e.g., Durda and Preziosi 2000; Hobbs et al. 2005; Breton et al. 2009; Schneider et al. 2009). For the purpose of this comparison, we focused only on the differences regarding assessment of reliability.

Domain of applicability

Some methods have been designed to be applicable to all types of ecotoxicity tests, with broad criteria that fit both aquatic and terrestrial tests (Klimisch et al. 1997; Durda and Preziosi 2000; Schneider et al. 2009; USEPA 2011; Beasley et al. 2015). For example, Durda and Preziosi (2000) explicitly stated that criteria specific for particular environmental compartments were not included in their method to avoid building a system with an impractically large number of criteria. Other methods focus on tests relating to one environmental compartment only, for example, aquatic tests (Hobbs et al. 2005; Breton et al. 2009; Van Der Kraak et al. 2014; Moermond et al. 2016). Some methods have been designed specifically for evaluation studies related to human health effects (Schneider et al. 2009), whereas others are applicable to both human and ecological data (Klimisch et al. 1997). Although some methods were specifically designed to be quite general (Klimisch et al. 1997; Beasley et al. 2015), even the more focused methods considered in this review were flexible enough to adapt to other environmental compartments because many criteria for toxicological and ecotoxicological studies are common.

All of the methods can be applied to both guideline and nonguideline studies, although the method of Breton et al. (2009) is the only one specifically designed to assess guideline studies. For evaluating in vitro studies, the method proposed by Schneider et al. (2009) is applicable directly without modifications; other methods may be adapted for those types of studies.

Tested on case studies

Several methods have been tested on case studies, either internally (Durda and Preziosi 2000; Breton et al. 2009; Van Der Kraak et al. 2014; Beasley et al. 2015) or using an external "round-robin" or ring test assessment. For example, the method presented by Hobbs et al. (2005) was tested using 2 studies and 23 participants, whereas a round-robin test of the CRED method by Moermond et al. (2016) used 8 studies and 75 participants (Kase et al., 2016). A method that has been validated and tested for clarity of the guidance using ring tests with multiple users will likely provide more consistent results.

Exclusion criteria and critical criteria

Both exclusion and critical criteria can be used to categorize a study to be reliable or not reliable, regardless of whether other criteria are met (see Table 1 for definitions). Criteria to exclude a study from further assessment are usually applied before a study is assessed in detail (Durda and Preziosi 2000; Schneider et al. 2009; USEPA 2011; Van Der Kraak et al. 2014). The purpose of using exclusion criteria is mostly to reduce the number of studies to be evaluated in detail to maximize the use of resources. Critical criteria are used during the assessment to identify study characteristics that are the most important or must be met to determine reliability (Durda and Preziosi 2000; Schneider et al. 2009; USEPA 2011; Van Der Kraak et al. 2014). Some methods do not have specific critical criteria but explicitly state that expert judgment is needed to decide which criteria are critical, because this may be different for each substance or organism tested (Beasley et al. 2015; Moermond et al. 2016).

Weighting of criteria

Applying different weights to evaluation criteria is used in some methods, especially those using numerical scoring, to recognize the greater importance of certain aspects of the methods (Hobbs et al. 2005; Breton et al. 2009; Van Der Kraak et al. 2014). These weights should be consistent with the problem formulation and the hypotheses being tested, and may be different for different compounds or organisms. For example, frequent measures of exposure concentrations in a test are more important for compounds that degrade rapidly than for those that do not.

Categories versus numerical scoring

Generally, reliability assigned to a study may be reported in 2 ways: either as a score or as a category. Scoring is the process where a number of reliability criteria are quantified, resulting in a numerical value. Often, these criteria take the form of questions with yes or no answers. Some criteria might be given greater weight than others, if they are regarded as key or critical criteria, based on expert judgment. Expert judgment may be used for this, based on, for example, the properties of the chemical. Care should be taken when reporting the results of the scoring and weighting approach, because the approach provides a numerical result, but precision and/or accuracy is not quantified. Scores with too much implied precision should be avoided. For WoE, the reliable data might be used in combination with assessment of the relevance of the response, and a scoring system may be preferred (Breton et al. 2009; Van Der Kraak et al. 2014).

When categories are used to summarize reliability, these are also assigned using a set of criteria, but the number of criteria met is not quantitatively evaluated, the criteria are not weighted, and the final assignment is made by reporting the category without a numerical meaning—for example, reliable and not reliable or A, B, C, or D. Sometimes numerals are used for category names—for example, 1, 2, 3, or 4—but they have no numerical meaning. The assignment of categories relies heavily on expert judgment. This categorical classification of data is mostly used when analyzing data for regulatory risk and/or hazard assessment and derivation of environmental quality criteria.

If required, scores can be translated into categories. For example, the scores can be divided into quartiles, from the worst (least trusted) to the best (most trustworthy) (Breton et al. 2009). Once the score of a study is translated into a category, it is important that one or more experts confirm whether the assigned

category is appropriate, and this should be adequately documented. When developing the threshold value along the continuous scoring scale where the study is reliable or not reliable, the best threshold may be different for each assessment and the context for which the assessment has been conducted. For some studies, 95% of all criteria may be met, but if 1 critical criterion is not met, the study may still be categorized as "not reliable." If this is done with automated scoring systems, there is greater likelihood of false-positive or false-negative results; however, this can be addressed by the application of a score for expert judgment (Van Der Kraak et al. 2014).

Bias toward GLP studies

Since the publication of the Klimisch method (Klimisch et al. 1997), concerns have been voiced about its perceived bias toward GLP studies (vom Saal and Myers 2010). From the methods evaluated in this study, only Durda and Preziosi (2000) state that GLP studies alone can be reliable, whereas several methods are biased toward studies conducted under GLP because higher weight is assigned to GLP studies with quality assurance and quality control (Breton et al. 2009; USEPA 2011; Van Der Kraak et al. 2014). Some methods are not biased toward GLP (Hobbs et al. 2005; Schneider et al. 2009; Beasley et al. 2015; Moermond et al. 2016), sometimes specifically stating so (Schneider et al. 2009; Moermond et al. 2016).

GLP studies (regardless if they follow a standard protocol) are usually more thoroughly documented than non-GLP studies. This often makes GLP studies easier to evaluate. However, GLP is not, by definition, a guarantee of the quality of the study (McCarty et al. 2012). GLP does not address the experimental design per se but rather establishes requirements for testing facilities, maintenance and calibration of equipment, training of personnel, independent quality-assurance inspections, recording of data, and recordkeeping. Studies following GLP might have flaws in the setup, performance, or treatment of data, whereas peer-reviewed research studies without GLP compliance may be scientifically sound and well performed.

Incorporation of standard guidelines in the evaluation method

Most of the methods incorporated some aspects of standard guidelines, such as the OECD toxicity testing methods (Breton et al. 2009; Schneider et al. 2009; Moermond et al. 2016). However, it will introduce a potential bias if studies from the literature that do not strictly adhere to guideline protocols are excluded. It is expected that many valuable peer-reviewed research studies would be conducted on nonstandard organisms, use different routes of exposure, and/or characterize responses or endpoints not specified in standard guidelines. Reliability assessment methods should be appropriate to evaluate these kinds of studies.

Number of criteria used

The number of criteria and/or subcriteria differ between methods, varying from 4 (Van Der Kraak et al. 2014) to as many as 40 (Breton et al. 2009). Sometimes the number of criteria is user defined (Beasley et al. 2015) or depends specifically on the type of study assessed (Breton et al. 2009). In the end, the number of criteria is not important, as long as it is possible to accurately assess the criteria and to understand the effect that individual criteria have on the final categorization, particularly critical criteria that should play a greater role in determining the reliability of the entire study. It is also essential that the criteria and method are clearly described to allow for transparent reporting of the assessment, overall and for the individual criteria, and consistency in reliability evaluations.

Type of criteria considered

Most methods include criteria regarding chemical characterization, experimental design, exposure conditions, test organism, and statistics. Only in the method of Beasley et al. (2015), where the criteria are user defined, might some of these aspects not be fully taken into account. Some methods include the quality of the reporting as a criterion (Klimisch et al. 1997; Durda and Preziosi 2000; Schneider et al. 2009; Van Der Kraak et al. 2014). However, the way results are reported could influence the categorization of reliability, and assessors should be aware of this (Moermond et al. 2016).

Guidance on how to evaluate criteria provided

Some methods provide detailed guidance on how individual criteria should be evaluated (Klimisch et al. 1997; Hobbs et al. 2005; Breton et al. 2009; Schneider et al. 2009; European Chemicals Agency 2010; USEPA 2011; Van Der Kraak et al. 2014; Beasley et al. 2015; Moermond et al. 2016). The availability of clear guidance can help to increase consistency when methods are used by different agencies and groups.

Guidance and/or tool provided on how to report results of the evaluation

A transparent, structured method of reporting the results of the evaluation increases the sense of trust in the evaluation. Some methods do not provide guidance or tools on how to report the results of the reliability assessment (Durda and Preziosi 2000; Hobbs et al. 2005; Beasley et al. 2015), others provide only very general statements on reporting the assessment (Klimisch et al. 1997), whereas yet others provide distinct tools, usually in the form of spreadsheets or scoring sheets (Breton et al. 2009; Schneider et al. 2009; Van Der Kraak et al. 2014; Moermond et al. 2016). It would enhance the transparency of the evaluation results if, in addition to stating whether a criterion is met, the rationale behind the choice made is also documented (Moermond et al. 2016). A well-documented, transparent, and reproducible evaluation increases confidence and consistency. Besides this, it will save time in the risk-assessment procedure when results need to be discussed among assessors or with other parties and will aid in harmonization of assessment among different regulatory frameworks.

 DOI: 10.1002/ieam.1870

OTHER CONSIDERATIONS WHEN EVALUATING RELIABILITY OF DATA

It is evident that a one-size-fits-all approach for evaluating reliability of ecotoxicity studies does not exist. Before starting the evaluation, an assessor should determine whether existing evaluation methods are applicable to the task. If the available evaluation methods are not suitable, assessors might consider developing their own evaluation methods. However, this may make it more difficult to compare their results with assessments that used more standard methods and also poses the burden of documenting the validity of their new method (Figure 1). Table 2, along with the points highlighted in this study, address major characteristics of the different approaches that are helpful when choosing or developing a method for reliability assessment, such as the determination of appropriate evaluation criteria (including exclusion and critical criteria), provision of clear guidance on how to evaluate these criteria, and guidance and/or tools on how to transparently report results of the evaluation. In addition, a number of other considerations exist when choosing or developing a method for reliability assessment:

- Selection of peer-reviewed research papers and the choice of reliability evaluation method will depend on the predefined problem formulation or purpose of the assessment (see Figure 1). This should be properly documented in the risk assessment.
- The number of studies to be evaluated may influence the choice for an assessment method. All methods discussed in this study were designed for reliability assessment of a single study or a small number of studies, except for the method of Van Der Kraak et al. (2014), which was specifically designed to evaluate a large number of studies (but may also be used when a smaller number of studies needs to be assessed). Depending on the objectives of the assessment, all methods can be adapted to assess a large number of studies. The time needed for the evaluation of a study usually depends more on the study itself (and the clarity of the reporting) than on the evaluation method. The choice of the method may also depend on how much transparency is needed in reporting the results.
- A tiered approach may be useful to efficiently screen publications for relevance and reliability, and prioritize studies for a more thorough evaluation, thereby increasing efficiency in the process (Figure 1). The method of Beasley et al. (2015) is an example of a tiered approach, where a larger data set is reduced by an evaluation of relevance in the first tier and a reliability assessment is performed only for the remaining studies. This is particularly useful in situations where a large number of studies need to be categorized or screened for reliability, and studies that are unsuitable for the end-goal assessment need to be excluded from further evaluation. A tiered assessment can be user defined. That is, criteria may depend on the purpose of the assessment or be determined by the parameters of the study guideline or protocol (e.g., validity criteria specified in a standardized test guideline).
- Expert judgment is always needed when assessing reliability because it is virtually impossible to devise a system that includes criteria that always apply to all studies evaluated and still be capable of capturing all the particulars of every study being assessed. However, expert judgment does not exclude the use of objective criteria, which provide uniformity of assessment. Some methods explicitly combine expert judgment with objective criteria (Van Der Kraak et al. 2014; Moermond et al. 2016) by providing detailed guidance. It is essential to make the use of expert judgment transparent to avoid the perception of bias. Study summaries or study evaluations are useful to document and communicate the rationale for expert judgment, which may aid readers in determining whether they agree with the expert's opinion.
- To ensure consistent assessments, it may be important to review an assessment again after evaluation of all studies has been completed, to account for information obtained during the process. Having independent quality assurance is helpful (Van Der Kraak et al. 2014), and a third-party review for reliability assessments, such as an expert group or an agency from another country, is also useful (Moermond et al. 2016).

CONCLUSIONS

When evaluating the reliability of an ecotoxicity study or individual effect value from this study, a systematic and transparent assessment method that utilizes expert judgment (based on clear guidance) is critical to ensure an unbiased and consistent assessment of reliability. In addition, transparent reporting of the methods used and the outcome of the assessment is important for consistency within and between regulatory agencies and other scientists and for informing the public.

This study provides a starting point for the ecological risk assessment process because it presents a comparison of different reliability assessment methods and discusses important attributes to consider when determining an appropriate method. Several methods for reliability assessment are described in the literature, each of which has attributes that may or may not be useful for the problem formulation or purpose of the assessment. It is important for assessors to keep in mind that studies not conducted using standard methods and/or in compliance with GLP may still provide useful information that is applicable to the problem formulation. Therefore, they should not be excluded from use in a hazard or risk assessment simply because they do not meet the criteria necessary for GLP or standard methods. In addition, it would be helpful if authors of published studies utilize supplemental information to provide specific details on the methods used and results generated in their studies. The additional information may allow for a study and/or effect value, especially for nonstandardized studies and endpoints, to be categorized as "reliable" rather than "not assignable" or "not reliable."

Once reliability of a study or effect value is determined, the assessor should also consider relevance of the study and/or effect value prior to using the data in a WoE approach and hazard and/or risk assessment. A companion paper discusses the assessment of relevance (Rudén et al. this issue).

Acknowledgment—We thank the Society of Environmental Toxicology and Chemistry for organizing the Pellston workshop, "Improving the Usability of Ecotoxicology in Regulatory Decision-Making."

Disclaimer—This paper is the work of the authors and does not necessarily reflect the view or opinions of their institutions.

Data availability—No data, metadata, or calculation tools were used in this study.

REFERENCES

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. *Environ Sci Eur* 23:17.

Ågerstrand M, Edvardsson L, Rudén C. 2014. Bad reporting or bad science? Systematic data evaluation as a means to improve the use of peer-reviewed studies in risk assessments of chemicals. *Human Ecol Risk Assess* 20:1427–1445.

Ågerstrand M, Küster A, Bachmann J, Breitholtz M, Ebert I, Rechenberg B, Rudén 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.

Beasley A, Belanger SE, Otter RR. 2015. Stepwise Information-Filtering Tool (SIFT): A method for using risk assessment metadata in a nontraditional way. *Environ Toxicol Chem* 34:1436–1442.

Beronius A, Molander L, Rudén C, Hanberg A. 2014. Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: A proposal to improve evaluation criteria and reporting. *J Appl Toxicol* 34:607–617.

Bevan C, Strother D. 2012. Best practices for evaluating method validity, data quality and study reliability of toxicity studies for chemical hazard risk assessments. Washington (DC): American Chemical Council, Centre for Advancing Risk Assessment Science and Policy. 26 p.

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.

Durda JL, Preziosi DV. 2000. Data quality evaluation of toxicological studies used to derive ecotoxicological benchmarks. *Human Ecol Risk Assess* 6:747–765.

European Chemicals Agency. 2010. Practical guide 2: How to report weight of evidence. Helsinki (FI): European Chemicals Agency. 26 p. http://echa.europa.eu/documents/10162/13655/pg_report_weight_of_evidence_en.pdf

European Food and Safety Authority. 2015. Principles and process for dealing with data and evidence in scientific assessments. *EFSA J* 13(4121):1–36.

Hanson ML, Wolff BA, Green JW, Kivi M, Panter GH, Warne MS, Ågerstrand M, Sumpter JP. 2016. How we can make ecotoxicology more valuable to environmental protection? *Sci Tot Environ* DOI: 10.1016/j.scitotenv.2016.07.160

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.

Kase R, Korkaric M, Werner I, Ågerstrand M. 2016. Criteria for Reporting and Evaluating ecotoxicity Data (CRED): Comparison and perception of the Klimisch and CRED methods for evaluating reliability and relevance of ecotoxicity studies. *Environ Sci Eur* 28:7.

Klimisch H-J., Andreae M, Tillmann U. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Reg Toxicol Pharmacol* 25:1–5.

McCarty LS, Borgert CJ, Mihaich EM. 2012. Information quality in regulatory decision-making: Peer review versus good laboratory practice. *Environ Health Perspect* 120:927–934.

Mensink BJWG, Smit CE, Montforts MHMM. 2008. Manual for summarising and evaluating environmental aspects of plant protection products. Bilthoven (NL): RIVM. 78 p. RIVM Report 601712004/2008.

Moermond C, Kase R, Korkaric M, Agerstrand M. 2016. CRED: Criteria for reporting and evaluating ecotoxicity data. *Environ Toxicol Chem* 35:1297–1309.

Rudén C, Adams J, Ågerstrand M, Brock TCM, Buonsante V, Poulsen V, Schlekat CE, Wheeler JR, Henry TR. 2017. Assessing the relevance of ecotoxicological studies for regulatory decision-making. *Integr Environ Assess Manag* 13:652–663.

Schenk L. 2010. Comparison of data used for setting exposure limits. *Int J Occup Environ Health* 16:249–262.

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.

Seiler JP. 2005. Good Laboratory Practice: The why and the how. New York (NY): Springer. 468 p.

[USEPA] US Environmental Protection Agency. 1998. Guidelines for ecological risk assessment. Washington (DC): USEPA, Risk Assessment Forum. 191 p. EPA/630/R-95/002F.

[USEPA] US Environmental Protection Agency. 2011. Evaluation guidelines for ecological toxicity data in the open literature. Procedures for screening, viewing, and using published open literature toxicity data in ecological risk assessments. Washington (DC): USEPA, Office of Pesticide Programs. 74 p.

Van Der Kraak GJ, Hosmer AJ, Hanson ML, Kloas W, Solomon KR. 2014. Effects of atrazine in fish, amphibians, and reptiles: An analysis based on quantitative weight of evidence. *Crit Rev Toxicol* 44(Suppl5):1–66.

vom Saal FS, Myers JP. 2010. Good laboratory practices are not synonymous with good scientific practices, accurate reporting, or valid data. *Environ Health Perspect* 118:A60.

