Toxicology Research

VIEWPOINT

View Article Online
View Journal | View Issue

Cite this: *Toxicol. Res.*, 2017, **6**, 571

Making the most of expert judgment in hazard and risk assessment of chemicals

A. Beronius *[a] and M. Ågerstrand [b]

Evaluation of the reliability and relevance of toxicity and ecotoxicity studies is an integral step in the assessment of the hazards and risks of chemicals. This evaluation is inherently reliant on expert judgment, which often leads to differences between experts' conclusions regarding how individual studies can contribute to the body of evidence. The conclusions of regulatory assessment, such as establishing safe exposure levels for humans and the environment and calculations of margins of exposure, may have large consequences for which chemicals are permitted on the market and their allowed uses. It is therefore important that such assessments are based on all reliable and relevant scientific data, and that assessment principles and assumptions, such as expert judgment, are transparently applied. It is not possible nor desirable to completely eliminate expert judgment from the evaluation of (eco)toxicity studies. However, it is desirable to introduce measures that increase structure and transparency in the evaluation process so as to provide scientifically robust risk assessments that can be used for regulatory decision making. In this article we present results from workshop exercises with Nordic experts to illustrate how experts' evaluations regarding the reliability and relevance of (eco)toxicity studies for risk assessment may vary and discuss methods intended to promote structure and transparency in the evaluation process.

Received 19th April 2017,
Accepted 5th July 2017

DOI: 10.1039/c7tx00114b

rsc.li/toxicology-research

Introduction

Hazard and risk assessment of chemicals is conducted as a step in chemicals regulation and used as the scientific basis for approving or restricting the use of chemicals. The conclusions of regulatory assessment, such as establishing safe exposure levels for humans and the environment and calculations of margins of exposure, may have large consequences for which chemicals are allowed on the market and their allowed uses. It is therefore important that such assessments are based on all reliable and relevant scientific data, and that assessment principles and assumptions are transparently applied. Hazard and risk assessment is inherently reliant on expert judgment.[1–4] Expert judgment may introduce value-based assumptions and influences, for example, how scientific data are evaluated and incorporated in the body of evidence, as well as the choice of assessment factors and how uncertainties are handled. It is not possible, nor desirable, to eliminate the use of expert judgment in the assessment process; application of expert judgment enables hazard and risk assessments to be flexible enough to consider and handle aspects that are especially critical in a specific case. However, differences in expert judgment may lead to significant differences in conclusions regarding the hazards and risks of chemicals.[5,6] It is therefore critical that the aspects of regulatory assessment of chemicals that are especially influenced by expert judgment are carried out in a way that is transparent and structured. One such aspect is the process of evaluating the "quality" of individual toxicity and ecotoxicity studies used as evidence in the assessment. In the European regulatory context this often requires evaluating the reliability and relevance of studies, where reliability is defined as the inherent quality of a study and relevance is defined as the extent to which the data and tests are appropriate for a particular hazard or risk assessment.[7]

The need for detailed criteria and tools that can increase structure and transparency in the evaluation of study reliability and relevance prompted the Science in Risk Assessment and Policy (SciRAP) initiative.[8] SciRAP provides an online platform with tools for evaluating toxicity and ecotoxicity studies, as well as nanoecotoxicity studies. These currently incorporate the Criteria for Reporting and Evaluating ecotoxicity Data (CRED), including the nano version,[9,10] and specific criteria for evaluating *in vivo* toxicity studies. The *in vivo* toxicity criteria were first published by Beronius and co-workers[11] but have since been updated. Criteria for the evaluation of *in vitro* studies are in development and will be available shortly. In addition, the platform includes an online colour-coding tool intended to facilitate the application of the evaluation criteria.

[a]*Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. E-mail: anna.beronius@ki.se*
[b]*Department of Environmental Science and Analytical Chemistry, Stockholm University, Sweden*

The tool generates a summary and colour profile of the evaluation results. In the case of *in vivo* toxicity studies the tool also calculates a numerical score. The evaluation results can be used as basis for conclusions concerning the reliability and relevance of studies used in hazard and risk assessment of chemicals. The different sets of evaluation criteria were initially developed based primarily on requirements and recommendations in relevant OECD test guidelines, and with consideration to other available methods for study evaluation. The SciRAP approach has undergone further refinement and development based on ring tests among risk assessors and application of the CRED and SciRAP *in vivo* criteria in different cases.[12–14] The criteria and tools, as well as guidance for researchers on how to report (eco)toxicity studies to fulfil regulatory requirements, are freely available online at http://www.scirap.org.

In November 2016, a Nordic workshop was organised for researchers and representatives of different authorities responsible for regulatory assessment of chemicals in Sweden, Norway, Finland, Denmark and Iceland. The purpose of the workshop was to discuss the use of the SciRAP approach for use in hazard and risk assessment in the Nordic countries and the EU. Workshop participants with expertise in toxicology and health risk assessment were asked to evaluate the reliability and relevance of a study on the developmental effects of bisphenol S in C57BL/6 mice,[15] as if it were considered for the setting of a European tolerable daily intake (TDI) for this compound. Evaluations were received from 12 participants. Workshop participants with expertise in ecotoxicology and environmental risk assessment were provided with a study that investigated histopathological alterations in fish for diclofenac,[16] and the task was to evaluate the study as if it were included in an assessment for an aquatic Environmental Quality Standard (EQS). Evaluations were received from 7 participants. The workshop and the results of these evaluations have been described in a workshop report published online as a Nordic Working Paper[17] and are further discussed in this paper. Specifically, the aims of this paper are to:

- Explore differences and commonalities in experts' evaluations of the studies,
- Illustrate how the SciRAP approach contributes to structured and transparent evaluation of studies for hazard and risk assessment of chemicals.

The purpose of this viewpoint article is not to provide a description of the SciRAP approach, beyond what is needed for the discussion on differences in experts' evaluations of study reliability and relevance. SciRAP has been previously described in several articles and reports.[8–11,13,17]

Differences and commonalities in experts' evaluations of the studies

The workshop exercises provide examples of how experts' evaluations of study reliability and relevance may vary, and show that variations remain even when tools intended to increase consistency and structure in the evaluation are used. These variations are likely due to differences in expertise and experience, but may in this case also have been partly due to evaluators' inexperience with the SciRAP approach, specifically. Several workshop participants expressed that they would have adjusted their evaluations after discussing it with their peers and gaining deeper understanding of the evaluation criteria and tool. It should however also be noted that the CRED evaluation method has been shown previously to contribute to a slight increase in consistency between evaluators when compared to the Klimisch evaluation method, *i.e.* the method currently recommended in several regulatory frameworks.[13,18] While the limited number of evaluations in these exercises do not allow for any deeper analyses of the performance of the SciRAP approach, some comments can be made regarding the extent of variations between evaluations and the specific points where evaluations were particularly varied and/or contradictory.

Evaluation of an *in vivo* toxicity study

The SciRAP method for evaluating reliability of *in vivo* toxicity studies consists of criteria for evaluating reporting quality and methodological quality of studies, separately. In the colour-coding tool, the weight of individual criteria may be increased or decreased, so that they have more or less impact on the outcome of the evaluation. This function has been introduced because different criteria may be more or less important in different cases, *e.g.* for different study types or different substances. The weights attributed to individual criteria are reflected in the output of the evaluation, *i.e.* criteria with increased weight contribute more to increasing or decreasing the score and to the % indicated as "fulfilled", "partially fulfilled" and "not fulfilled" in the colour profile.

The evaluations conducted by the workshop participants show variations both in how individual criteria were weighed and how they were judged (Fig. 1 and 2). The different evaluators' scores for reporting quality and methodological quality varied between 47.9–89.1 (average 61.5) and 40.0–84.4 (average 68.1), respectively. Notably, more criteria were judged as "partially fulfilled" in the evaluation of methodological quality than for reporting quality. This may have several explanations. For example, it may be an indication that the evaluation of methodological quality is more complex or difficult (requiring a higher level of expertise) than the evaluation of reporting quality. However, it may also be that the criteria are worded in a way that makes it difficult to judge strictly if it is "fulfilled" or "not fulfilled".

In terms of the reporting quality of the study, evaluation of criteria relating to the description of the test compounds and controls (criteria 1, 2, 4 and 5) showed larger variation, *i.e.* weights varied between "not applicable" and "3", and individual criteria were judged both as "fulfilled" and "not fulfilled" by more than one evaluator. In addition, the criteria about the method for allocating animals to different treatments (criterion 19) and to different tests and measurements (criterion 24), as well as criterion 27 regarding the reporting of the statistical unit, showed large variation, in terms of being fulfilled

	ID 1	ID 2	ID 3	ID 4	ID 5	ID 6	ID 7	ID 8	ID 9	ID 11
SciRAP score (reporting quality)	53.2	63.3	61.6	53.3	65.6	60.0	89.1	67.2	53.5	47.9
Evaluation Criteria:										
1. The chemical name, ID or CAS-number of the test compound was given.	3	3	1	2	2	2	2	2	3	3
2. The purity of the test compound was stated or is traceable[...]	3	2	2	2	3	2	2	2	3	3
3. The vehicle was described.	2	2	3	2	2	2	2	2	3	2
4. It was stated that a negative control group was included.	2	1	3	2	2	2	2	2	3	2
5. Any positive control, if used, was described.	3	2	1	2	2	2	N/A	N/A	3	N/A
6. The animal model (species, strain, age or life stage and sex) was described.	3	2	3	2	2	2	2	2	3	2
7. The method for individual identification of animals was described.	2	2	1	2	2	2	2	2	2	2
8. The housing temperature was stated.	2	2	2	2	2	2	N/A	2	2	2
9. The relative humidity was stated.	2	2	1	2	2	2	N/A	2	2	2
10. The light-dark cycle was described.	2	2	1	2	2	2	1	2	2	2
11. The number of animals per sex in each cage was stated.	3	2	2	2	2	2	2	2	3	2
12. The cage materials were described.	2	2	3	2	3	2	2	2	2	2
13. Any materials used for physical enrichment were described.	2	2	3	2	3	2	N/A	2	2	2
14. Water bottle materials were described.	3	2	3	2	3	2	2	2	2	2
15. The bedding material used was described.	2	2	3	2	2	2	N/A	2	2	2
16. The type and source of feed were reported.	3	2	3	2	2	2	2	2	2	2
17. The source of drinking water was reported.	3	2	3	2	3	2	1	2	2	2
18. The administered dose levels or concentrations were stated.	3	2	3	2	2	2	2	2	3	3
19. The method for allocating animals to different treatments was stated.	3	2	1	2	2	2	2	2	2	3
20. The total number of animals per dose group was stated.	3	2	3	2	2	2	2	2	3	3
21. The route of administration was stated.	3	2	3	2	2	2	2	2	3	3
22. The sex and age (or life stage) of the animals at start of dosing was stated[...]	3	2	3	2	2	2	2	2	2	3
23. The test and/or analytical methods used were sufficiently described to allow for evaluation of the reliability of results.	3	2	3	2	2	2	2	2	3	3
24. The method for allocating animals to different tests and measurements was stated.	2	2	3	2	2	2	2	2	2	2
25. The sex, age and number of animals per dose group subjected to separate tests and measurements was stated.	3	2	3	2	2	2	2	2	3	2
26. The statistical methods and software used were described.	3	2	3	2	2	2	2	2	3	3
27. The statistical unit, e.g. the individual or the litter, was stated.	3	2	3	2	2	2	2	2	2	3
28. All results for the investigated endpoints were reported[...]	3	2	3	2	2	2	2	2	3	3
29. The funding sources for the study were stated.	3	2	3	2	2	2	2	2	2	3
30. Any competing interests were disclosed[...]	2	2	1	2	2	2	N/A	2	2	3

Fig. 1 Evaluation of reporting quality of the provided *in vivo* toxicity study from 10 different evaluators. Green, yellow, red and grey indicate that the criterion was reported, partially reported, not reported or not determined, respectively. The numbers indicate the weight attributed to each criterion where 2 is default, and 1 or 3 means the criterion was considered less important or specifically important, respectively, in this case. N/A = the criterion was considered not applicable in this case and has been removed from the evaluation. Two evaluators, ID# 10 and 12, did not complete the evaluation of reporting quality due to technical difficulties.

or not, between evaluators. While differences in the weight attributed to criteria may reflect the evaluators' differing opinions about the importance of reporting such information, the contradictions in judging whether this information is in fact reported or not is rather intriguing. The explanation may be that the wording of the criteria is not clear and therefore leading to different interpretations of the criteria themselves. Some of the variation resulting from differences in the interpretation of criteria may be reduced by providing guidance also for these reporting quality criteria.

Methodological quality criteria showing large variation and contradictory evaluations were those relating to impurities of the test compound (criterion 1), contamination of the test system (criterion 9), as well as whether allocation of animals to treatments (criterion 10) and to tests and measurements (criterion 14) were randomized. These were criteria that were judged both as "fulfilled" and "not fulfilled" by more than one evaluator. Contradictions in the evaluations of these criteria are likely due to differences in evaluators' expertise regarding the compound being tested, as well as the study type and design. Such differences could possibly be solved to a certain degree by discussions between the evaluators. Differences in how criteria were interpreted is also a possible explanation. Guidance for interpreting and evaluating these criteria are provided in the SciRAP tool, with reference to relevant test guidelines and guidance documents from *e.g.* the OECD. However, such guidance may need to be improved. Especially for criteria where large variations and contradictions are consistently observed.

In this exercise, the participants were not asked to conclude on a level of reliability for the evaluated study. The intention of SciRAP is that the outcome could be used in different contexts and for different types of hazard or risk assessment questions. Thus, the score and colour profile can be used together *e.g.* to rank studies within a body of evidence or to categorise into different reliability categories. However, SciRAP does not provide any specific directions or cut-off scores for concluding on a level of reliability.

Currently, in the SciRAP approach, evaluation of relevance of *in vivo* toxicity studies is not conducted in the colour-coding tool. The evaluation is restricted to judging whether the study is relevant or not for the case at hand by considering a set of items intended as guidance in this process. All workshop par-

	ID 1	ID 2	ID 3	ID 4	ID 5	ID 6	ID 7	ID 8	ID 9	ID 10	ID 11	ID 12
SciRAP score (methodological quality)	61.0	72.4	62.5	71.1	76.9	75.0	84.4	80.6	73.3	40.0	44.0	75.7
Evaluation Criteria:												
1. The test compound or mixture was unlikely to contain any impurities that may significantly have affected its toxicity.	3	3	2	2	2	2	2	2	3	3	3	3
2. An appropriate vehicle was used that is not expected to interfere with the absorption, distribution, metabolism, excretion or toxicity of the test compound.	2	2	3	2	2	2	2	2	3	2	3	2
3. A concurrent negative control group was included.	1	2	3	2	2	2	2	2	3	3	3	3
4. An appropriate positive control group was included, and the expected result was observed from this treatment	3	1	1	2	2	N/A	N/A	N/A	2	2	N/A	N/A
5. A reliable and sensitive animal model was used for investigating the test compound and selected endpoints.	3	2	3	2	2	2	2	2	2	2	3	2
6. Animals were individually identified.	3	2	2	2	2	2	N/A	2	2	2	2	2
7. Housing conditions (temperature, relative humidity, light-dark cycle) were appropriate for the study type and animal model.	2	2	2	2	2	2	N/A	2	2	2	2	1
8. The number of animals per sex in each cage were appropriate for the study type and animal model.	3	2	3	2	2	2	2	2	2	2	2	2
9. The test system was unlikely to contain contaminants that could affect study results, such as organic pollutants, pesticide residues, heavy metals, and mycotoxins, as well as phytoestrogens.	3	2	3	2	3	2	2	2	2	2	3	2
10. The allocation of animals to different treatments was randomized.	3	2	1	2	2	2	2	2	2	2	3	2
11. The route of administration was appropriate and not likely to interfere with the study results.	2	2	3	2	2	2	2	2	3	2	3	3
12. The timing and duration of administration were appropriate for investigating the included endpoints.	3	2	3	2	2	2	2	2	3	2	3	2
13. The frequency of administration is appropriate for investigating the included endpoints.	2	2	3	2	2	2	2	2	2	2	3	2
14. The allocation of animals to different tests and measurements was randomized.	3	2	1	2	2	2	2	2	3	2	3	2
15. Reliable, and sensitive test methods were used for investigating the selected endpoints.	3	2	3	2	2	2	2	2	2	2	3	2
16. Measurements were collected at suitable time points in order to generate sensitive, valid and reliable data.	2	2	3	2	2	2	2	2	2	2	3	2
17. A sufficient number of animals per dose group were subjected to separate tests/data collection/measurements to generate reliable and valid results.	3	2	3	2	2	2	2	2	2	2	3	2
18. The statistical methods have been clearly described and do not seem inappropriate, unusual or unfamiliar.	3	2	3	2	2	2	2	2	3	2	3	2

Fig. 2 Evaluation of methodological quality of the provided *in vivo* toxicity study from 12 different evaluators. Green, yellow, red and grey indicate that the criterion was fulfilled, partially fulfilled, not fulfilled or not determined, respectively. The numbers indicate the weight attributed to each criterion where 2 is default, and 1 or 3 means the criterion is considered less important or specifically important in this case, respectively. N/A = the criterion is considered not applicable in this case and has been removed from the evaluation.

ticipants considered the study relevant for the purpose of setting a TDI for the tested compound, but several participants emphasized that it would be used together with other evidence (in a weight of evidence approach) and not on its own.

Evaluation of an ecotoxicity study

Based on the evaluation result and guidance for the different reliability and relevance categories each participant provided a conclusion on the ecotoxicity study. One participant categorised the study as "reliable without restrictions" while the others found it to be "reliable with restrictions". Three participants categorised the study as "relevant without restrictions" while the others chose "relevant with restrictions". There was an overlap between the percentage fulfilled/partially fulfilled criteria and the assigned category for both reliability or relevance. For reliability, the percentage fulfilled/partially fulfilled criteria ranged from 47.5–87.5 (average 69.8) for the category "reliable with restrictions", compared to 70 for "reliable without restrictions". For relevance, the percentage fulfilled/partially fulfilled criteria ranged from 62–77.3 (average 71.3) for the category "relevant with restrictions", compared to 69.2–86.2 (average 80) for "relevant without restrictions". This indicates that an individual criterion can play a crucial role when reliability and relevance categories are assigned, meaning that one specific criteria can alter the outcome considerably. This result is supported by a previous study.[13] In the evaluation process, expert judgement plays a crucial role and an evaluation method that promotes transparency and clear justifications is therefore desirable.

Seemingly contradictive, less than half of the participants choose to use the option of weighing the criteria prior to the evaluation. In discussions with the participants some expressed a need for additional guidance on how to weigh criteria since this was new to many. In general, the perception towards the possibility of weighing criteria was positive since it increases the transparency of the evaluations and gives an

indication of what was considered important for the particular study under evaluation.

For the relevance evaluation of the ecotoxicity study participants disagreed on whether the endpoint was relevant for effects on population level, *i.e.* for this criterion at least two participants judged it to be "fulfilled" while at least two others judged it as "not fulfilled" (criterion 5, Fig. 3). This is not surprising; studies investigating histopathological alterations are not used on a regular basis when setting EQS-values since there are diverging opinions whether these type of alterations also could have an effect on populations. Reliability aspects where participants disagreed related to test substance and solvent concentrations, biomass loading, and whether sufficient data were available to allow for additional calculations (criterion 13, 16 and 20, Fig. 4). For the first two criteria, the disagreement was due to different understandings of the study, and for the last criterion due to different views on what is needed to re-calculate data. The conclusions from the NORAP workshop participants regarding the reliability and relevance of the ecotoxicity study are in line with the conclusion from the EQS dossier from the Sub-group on Review of the Priority Substances,[19] as well as the European Commission's SCHER Committee.[20]

Contribution to structured and transparent evaluation of studies

The participants at the workshop generally agreed that the output from the SciRAP evaluation provides a useful basis for conclusions regarding the reliability and relevance of toxicity and ecotoxicity studies. For example, to categorise studies into different categories for reliability and relevance, or for ranking studies according to their "relative" reliability and/or relevance if a larger body of evidence is being evaluated. Importantly, the SciRAP approach was considered to facilitate discussions between evaluators to explain and possibly resolve differences in their evaluations. As such, these methods are considered to contribute to increasing transparency in the study evaluation process. Guidance for categorising studies is currently available for ecotoxicity and nanoecotoxicity studies but not for *in vivo* toxicity studies and was identified as an important factor to increase transparency and consistency in evaluations further.

Another perceived benefit of the SciRAP approach was as a means of illustrating any consistent weaknesses in design, conduct and/or reporting of toxicity and ecotoxicity studies. These insights could be further used to develop guidance for researchers, especially for how to report studies in a way that meets the reliability and reproducibility requirements set in regulatory assessments. Reporting details of study conduct and results is critical in the regulatory setting, since insufficient reporting hampers study evaluation and, consequently, its use in hazard or risk assessment of chemicals.

Some participants considered the two-step process of first weighing and then judging each criterion cumbersome. However, it was considered important to be able to adjust the weight of individual criteria, as they may be more or less critical to the evaluation of studies in different cases. Further, guidance for how to interpret and use the results of the SciRAP

Evaluation criteria	ID 1	ID 2	ID 3	ID 4	ID 5	ID 6	ID 7
1. Is the species tested relevant for the compartment under evaluation?	2	2	2	3	2	2	3
2. Are the organisms tested relevant for the tested substance?	2	2	2	3	2	2	3
3. Are the reported endpoints appropriate for the regulatory purpose?	2	2	2	3	2	1	3
4. Are the reported endpoints appropriate for the investigated effects or the mode of action of the test substance?	2	2	2	2	2	2	3
5. Is the effect relevant on a population level?	2	2	2	3	2	1	2
6. Are appropriate life stages studied?	2	2	2	3	2	1	3
7. Is the magnitude of effect statistically significant and biologically relevant for the regulatory purpose (e.g., EC10, EC50)?	2	2	2	3	2	2	3
8. Are the experimental conditions relevant for the tested species?	2	2	2	3	2	2	3
9. Is the exposure duration relevant and appropriate for the studied endpoints and species?	2	2	2	3	2	2	2
10. If recovery is studied, is this relevant for the framework for which the study is evaluated?	2	N/A	2	2	N/A	2	N/A
11. In case of a formulation, other mixture, salts, or transformation products, is the substance tested representative and relevant for the substance being assessed?	2	2	2	3	N/A	2	N/A
12. Is the tested exposure scenario relevant for the substance?	2	2	2	2	2	3	2
13. Is the tested exposure scenario relevant for the species?	2	2	2	2	2	3	2

Fig. 3 Relevance evaluation of the provided ecotoxicity study, as reported by 7 different evaluators (ID 1–7). Green, yellow, red and grey indicate that the criterion was fulfilled, partially fulfilled, not fulfilled or not determined, respectively. The numbers indicate the weight attributed to each criterion where 2 is default, and 1 or 3 means the criterion is considered less important or specifically important in this case, respectively. N/A = the criterion is considered not applicable in this case and has been removed from the evaluation.

Evaluation criteria	ID 1	ID 2	ID 3	ID 4	ID 5	ID 6	ID 7
1. Is a guideline method (e.g., OECD/ISO) or modified guideline used?	2	2	2	2	2	2	1
2. Is the test performed under GLP conditions?	2	2	2	1	2	2	1
3. If applicable, are validity criteria fulfilled (e.g. control survival, growth)?	2	2	2	3	3	2	2
4. Are appropriate controls performed (e.g. solvent control, negative and positive control)?	2	2	2	3	2	2	3
5. Is the test substance identified clearly with name or CAS-number? Are test results reported for the appropriate compound?	2	2	2	3	2	2	2
6. Is the purity of the test substance reported? Or, is the source of the test substance trustworthy?	2	2	2	3	2	2	2
7. If a formulation is used or if impurities are present: Do other ingredients in the formulation exert an effect? Is the amount of test substance in the formulation known?	2	2	2	2	2	2	N/A
8. Are the organisms well described (e.g. scientific name, weight, length, growth, age/life stage, strain/clone, gender if appropriate)?	2	2	2	2	2	2	2
9. Are the test organisms from a trustworthy source and acclimatized to test conditions? Have the organisms not been pre-exposed to test compound or other unintended stressors?	2	2	2	2	2	2	2
10. Is the experimental system appropriate for the test substance, taking into account its physico-chemical characteristics?	2	2	2	3	2	2	3
11. Is the experimental system appropriate for the test organism (e.g., choice of medium or test water, feeding, water characteristics, temperature, light/dark conditions, pH, oxygen content)? Have conditions been stable during the test?	2	2	2	2	2	2	3
12. Were exposure concentrations below the limit of water solubility (taking the use of a solvent into account)? If a solvent is used, is the solvent within the appropriate range and is a solvent control included?	2	2	2	3	2	2	2
13. Is a correct spacing between exposure concentrations applied?	2	2	2	2	2	2	1
14. Is the exposure duration defined?	2	2	2	3	2	2	3
15. Are chemical analyses adequate to verify concentrations of the test substance over the duration of the study?	2	2	2	2	2	2	3
16. Is the biomass loading of the organisms in the test system within the appropriate range (e.g. < 1 g/L)?	2	2	2	2	2	2	3
17. Is a sufficient number of replicates used? Is a sufficient number of organisms per replicate used for all controls and test concentrations?	2	2	2	2	2	2	2
18. Are appropriate statistical methods used?	2	2	2	3	2	2	2
19. Is a concentration-response curve observed? Is the response statistically significant?	2	2	2	2	2	2	3
20. Are sufficient data available to check the calculation of endpoints and (if applicable) validity criteria (e.g., control data, concentration-response curves)?	2	2	2	3	2	2	3

Fig. 4 Reliability evaluation of the provided ecotoxicity study, as reported by 7 different evaluators (ID 1–7). Green, yellow, red and grey indicate that the criterion was fulfilled, partially fulfilled, not fulfilled or not determined, respectively. The numbers indicate the weight attributed to each criterion where 2 is default, and 1 or 3 means the criterion is considered less important or specifically important in this case, respectively. N/A = the criterion is considered not applicable in this case and has been removed from the evaluation.

approach in the hazard and risk assessment process, and how to express uncertainty in the evaluation, was considered needed. Current work is ongoing to update and enhance the SciRAP online platform based on feedback from workshop participants.

Conclusions

The purpose of the SciRAP approach for evaluating study reliability and relevance is not to eliminate expert judgment from hazard and risk assessment of chemicals but to provide tools that promote structured and transparent application of evaluators' expertise in this process. The experience with this approach so far indicates that it contributes towards that purpose, and has specifically brought to light the usefulness of such tools in facilitating discussions between evaluators/risk assessors to explain and possibly resolve differences in evaluations. Training with the tools and the availability of guidance for evaluation, as well as categorisation into reliability/relevance categories, are important factors that are likely to contribute to consistency between evaluations. By using structured and transparent approaches like SciRAP it is possible to make the most of expert judgment, a crucial and inherent part of evaluating evidence for hazard and risk assessment of chemicals. Ongoing and future work to further develop the SciRAP platform include developing and testing a method for evaluating *in vitro* studies, providing further guidance for evaluators, especially in relation to the interpretation of the evaluation results, as well as improving the functionality of the platform by including a possibility to save evaluations online.

Conflict of interest

The authors are the developers of the SciRAP tool. SciRAP is available online free of charge and does not generate any revenue. The authors do not have any other conflicts of interest.

Acknowledgements

The SciRAP workshop was made possible by grants from the Nordic Chemical Group (NKG) of the Nordic Council of

Ministers and organised with the support of the Nordic Risk Assessment Project (NORAP). The authors would specifically like to thank everyone who participated in the workshop for contributing with their time and expertise, as well as Annika Hanberg and Christina Rudén for contributing to the organisation of the workshop and the workshop report.

References

1 M. Agerstrand, M. Breitholtz and C. Rudén, 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.*, 2011, **23**, 17.

2 Scientific Committee on Emerging and Newly Identified Health Risks. Memorandum on the use of the scientific literature for human health risk assessment purposes – weighing of evidence and expression of uncertainty. 2012. Available on-line: https://ec.europa.eu/health/scientific_committees/emerging/docs/scenihr_s_001.pdf.

3 B. Wandall, Values in science and risk assessment, *Toxicol. Lett.*, 2004, **152**, 265–272.

4 D. L. Weed, Weight of evidence: a review of concept and methods, *Risk Anal.*, 2005, **25**, 1545–1557.

5 A. Beronius, C. Ruden, H. Hakansson and A. Hanberg, Risk to all or none? A comparative analysis of controversies in the health risk assessment of Bisphenol A, *Reprod. Toxicol.*, 2010, **29**, 132–146.

6 J. V. Tarazona, D. Court-Marques, M. Tiramani, H. Reich, R. Pfeil, F. Istace and F. Crivellente, Glyphosate toxicity and carcinogenicity: a review of the scientific basis of the European Union assessment and its differences with IARC, *Arch. Toxicol.*, 2017, DOI: 10.1007/s00204-017-1962-5.

7 European Chemicals Agency, Guidance on information requirements and chemical safety assessment. Chapter R.4: Evaluation of available information. 2011. Available on-line: https://echa.europa.eu/documents/10162/13643/information_requirements_r4_en.pdf.

8 L. Molander, M. Agerstrand, A. Beronius, A. Hanberg and C. Rudén, Science in Risk Assessment and Policy (SciRAP): An Online Resource for Evaluating and Reporting In Vivo (Eco) Toxicity Studies, *Hum. Ecol. Risk Assess.*, 2015, **21**, 753–762.

9 C. Moermond, R. Kase, M. Korkaric and M. Agerstrand, CRED - Criteria for Reporting and evaluating ecotoxicity Data, *Environ. Toxicol. Chem.*, 2015, **35**, 1297–1309.

10 N. B. Hartmann, M. Agerstrand, H.-C. H. Lûtzhoft and A. Baun, NanoCRED: A transparent framework to access the regulatory adequacy of ecotoxicity data for nanomaterial - Relevance and reliability revisited, *NanoImpact*, 2017, **6**, 81–89.

11 A. Beronius, L. Molander, C. Ruden and A. Hanberg, 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.*, 2014, **34**, 607–617.

12 European Chemicals Agency, Annex XV report – Identification of Dicyclohexyl phthalate (DCHP) as SVHC, 2016, Available online: https://www.echa.europa.eu/documents/10162/0f8a6fd1-835f-4686-8cca-bcbbdc137b73.

13 R. Kase, M. Korkaric, I. Werner and M. Agerstrand, 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.*, 2016, **28**, 7.

14 L. Molander, A. Hanberg, C. Ruden, M. Agerstrand and A. Beronius, Combining web-based tools for transparent evaluation of data for risk assessment: developmental effects of bisphenol A on the mammary gland as a case study, *J. Appl. Toxicol.*, 2017, **37**, 319–330.

15 L. Ivry Del Moral, L. Le Corre, H. Poirier, I. Niot, T. Truntzer, J. F. Merlin, P. Rouimi, P. Besnard, R. Rahmani and M. C. Chagnon, Obesogen effects after perinatal exposure of 4,4′-sulfonyldiphenol (Bisphenol S) in C57BL/6 mice, *Toxicology*, 2016, **357–358**, 11–20.

16 J. Schwaiger, H. Ferling, U. Mallow, H. Wintermayr and R. D. Negele, Toxic effects of the non-steroidal anti-inflammatory drug diclofenac. Part I: histopathological alterations and bioaccumulation in rainbow trout, *Aquat. Toxicol.*, 2004, **68**, 141–150.

17 A. Beronius, M. Agerstrand, C. Rudén and A. Hanberg, SciRAP workshop report: Bridging the gap between academic research and chemicals regulation – the SciRAP tool for evaluating toxicity and ecotoxicity data for risk assessment of chemicals, Nordic Working Papers, Nordic Council of Ministers, Copenhagen, 2017, 33 pp.

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

19 European Commission Sub-Group on Review of the Priority Substances List. Diclofenac EQS dossier. 2011. Available online: https://circabc.europa.eu/sd/a/d88900c0-68ef-4d34-8bb1-baa9af220afd/Diclofenac%20EQS%20dossier%202011.pdf.

20 European Commission Scientific Committee on Health and Environmental Risks, Opinion on Chemicals and the Water Framework Directive: Draft Environmental Quality Standards, Diclofenac. 2011, available online: https://ec.europa.eu/health/scientific_committees/environmental_risks/scher_09-13/opinions_en.