Toxicology Letters 189 (2009) 138–144

Contents lists available at ScienceDirect

Toxicology Letters

journal homepage: www.elsevier.com/locate/toxlet

"ToxRTool", a new tool to assess the reliability of toxicological data

Klaus Schneider[a,*], Markus Schwarz[a], Iris Burkholder[b], Annette Kopp-Schneider[b], Lutz Edler[b], Agnieszka Kinsner-Ovaskainen[c], Thomas Hartung[d], Sebastian Hoffmann[e]

[a] Forschungs- und Beratungsinstitut Gefahrstoffe GmbH (FoBiG), Werthmannstraße 16, 79098 Freiburg, Germany
[b] Department of Biostatistics, C060, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[c] In Vitro Methods, ECVAM, Institute for Health and Consumer Protection, E.C. Joint Research Centre, Ispra, Italy
[d] Doerenkamp-Zbinden, Chair for Evidence-Based Toxicology, Center for Alternatives to Animal Testing, Johns Hopkins University, Bloomberg School of Public Health, Baltimore, United States
[e] TÜV Rheinland BioTech GmbH, Nattermannallee 1, 50829 Köln, Germany

ARTICLE INFO

Article history:
Received 25 March 2009
Received in revised form 14 May 2009
Accepted 19 May 2009
Available online 27 May 2009

Keywords:
Data quality
Reliability
Klimisch categories
Relevance
In vitro data
In vivo data

ABSTRACT

Evaluation of the reliability of toxicological data is of key importance for regulatory decision-making. In particular, the new EU Regulations concerning the registration, evaluation, authorisation and restriction of chemicals (REACH) and classification, labelling and packaging (CLP) according to the new globally harmonised system (GHS) rely on the integration of all available toxicological information. The so-called Klimisch categories, although well established and widely used, lack detailed criteria for assigning data quality to categories. A software-based tool (ToxRTool) was developed within the context of a project funded by the European Commission to provide comprehensive criteria and guidance for reliability evaluations of toxicological data. It is applicable to various types of experimental data, endpoints and studies (study reports, peer-reviewed publications) and leads to the assignment to Klimisch categories 1, 2 or 3. The tool aims to increase transparency and to harmonise approaches of reliability assessment. The tool consists of two parts, one to evaluate *in vivo* and one to evaluate *in vitro* data. The prototypes of the tool were tested in two independent inter-rater experiments. This approach allowed the analysis of the performance of the tool in practice and the identification and minimisation of sources of heterogeneity in evaluation results. The final version, ToxRTool, is publicly available for testing and use.

1. Introduction

The use of existing data on toxicological properties of chemicals is necessary in various regulatory contexts on grounds of animal welfare considerations (avoidance of unnecessary animal experiments) as well as for economic reasons. The new European chemicals policy, Regulation EC 1907/2006 (REACH), places strong emphasis on the use of existing data. According to the relevant REACH guidance documents (ECHA, 2008), the registration process should always start with a thorough evaluation of all available data with regard to whether the information is reliable and sufficient to fulfil the information requirements. Also, under the new globally harmonised system for classification and labelling of chemicals (GHS) existing data will be the most important source for classifying substances with respect to their hazardous properties.

However, existing data vary largely in quality and consequently the objective evaluation of data quality gains importance. The Occupational and Public Health Specialty Section of the US Society of Toxicology recently expressed the concern that differences in the evaluation of data quality may counteract the efforts to harmonise classification of substances under GHS (SOT, 2007).

In 1997, Klimisch et al. published a categorisation system with the aim of assigning toxicological data to one of four reliability categories (see Table 1). The authors introduced the following definitions (Klimisch et al., 1997):

- *Reliability*: evaluating the inherent quality of a test report or publication relating to preferably standardised methodology and the way the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings;
- *Relevance*: covering the extent to which data and/or tests are appropriate for a particular hazard identification or risk characterisation; and
- *Adequacy*: defining the usefulness of data for hazard/risk assessment purposes. When there is more than one set of data for each effect, the greatest weight is attached to the most reliable and relevant.

These terms and the reliability categories are widely applied in various regulatory programmes such as the OECD High Production

* Corresponding author. Tel.: +49 761 38608 12; fax: +49 761 38608 20.
E-mail address: klaus.schneider@fobig.de (K. Schneider).

0378-4274/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.toxlet.2009.05.013

Table 1
Reliability categories according to Klimisch et al. (1997).

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

Volume Programme and REACH and are explicitly referred to in the REACH guidance documents (ECHA, 2008).

Whereas some aspects of data quality were mentioned by Klimisch et al. (1997), explicit criteria allowing conclusions to be reached on the reliability categories are lacking. Studies performed according to recent guidelines and under conditions of good laboratory practice (GLP) are generally considered to be of high reliability (category 1). However, existing data may originate from times when such quality-assured generation of data was not yet common or were published in peer-reviewed journals that differ with regard to the level of documentation required. In particular, guidance is lacking on how to distinguish reliability categories 2 and 3 (for definition see Table 1), which in the regulatory context is the most crucial differentiation. Data of reliability category 1 or 2 can be used as stand-alone information to cover a specific endpoint, whereas category 3 information at best may serve as additional information in weight-of-evidence approaches.

To provide such guidance, a research project was initiated by ECVAM, the European Centre for the Validation of Alternative Methods, of the European Commission's Joint Research Centre in Ispra, Italy. The project was aimed at developing criteria and a transparent methodology for evaluating the quality of existing experimental toxicological data. While a methodology has been proposed for the development of data quality criteria for ecotoxicological data (Hobbs et al., 2005), similar initiatives have not been taken for toxicological data.

As both relevance and adequacy of data depend on the specific regulatory context in which it is to be used, the project focused on the inherent quality, i.e. the reliability of toxicological data, which is independent of the use of these data for regulatory purposes. The objective was to establish a set of assessment criteria, with which to evaluate toxicological data, and implement these criteria in a readily available and user-friendly tool facilitating the documentation of the assessment. A reliability assessment tool called ToxRTool (**Tox**icological data **R**eliability Assessment **Tool**) was developed in this project intended to be used by scientists routinely dealing with reliability assessment of toxicological data. Ultimately the tool will increase transparency and provide guidance for more harmonised approaches to data quality evaluations. This effort is in agreement with recent activities focussing on the use of evidence-based approaches in toxicological practice ("evidence-based toxicology", EBT, www.ebtox.org) (Hoffmann and Hartung, 2006).

2. Methods

2.1. Development of the tool

The primary objective of the tool is to provide a transparent methodology for assigning data from a toxicological study to Klimisch categories 1, 2 or 3 by assessment against specific, weighted criteria (as detailed below). The tool has been designed to be applicable to original data only, and assignment to Klimisch category 4, which refers to secondary sources and handbooks, has not been implemented in the tool.

The criteria were initially developed by compilation of a list of parameters with potential impact on the data quality of a study. Parameters were retrieved from evaluating guideline requirements as well as publications and reports related to this subject. From this initial list, it became obvious that the tool had to be organised in two parts, i.e. one for *in vivo* data and one for *in vitro* data. Nevertheless, both parts of the tool were developed in parallel following the same structure. Parameters were reformulated into questions (criteria) and divided into five groups according to whether they related to: (1) test substance identification, (2) test organism (system) characterisation, (3) study design description, (4) study result documentation, and (5) plausibility of study design and results.

Following approaches described by the US Environmental Protection Agency (US EPA, 1999) and under REACH (ECHA, 2008) a minimum set of information in each group which is considered to be indispensable for reliable data, was defined. The criteria addressing such information are marked in red (in the following referred to as 'red criteria'). Non-compliance with at least one red criterion leads to Klimisch category 3, irrespective of the total score (i.e. sum of scores for all criteria) achieved. If all red criteria are met, individual scores are summed up and lead to the assignment of data to Klimisch categories 1 (at least ca. 80% of maximum score) and 2 (at least ca. 60% of maximum score). Lower total scores result in category 3. The cut-off values for categories 1 and 2 were arbitrarily set to provide adequate ranges for all categories, reflecting current use of the Klimisch categories.

Prototype 1 of the tool consisted of 25 criteria each for the *in vivo* and *in vitro* part, to be answered with "yes" (score 1) or "no" (score 0). Criteria were supplemented with explanations to support unambiguous and common understanding of their meaning.

At the initial stages of the project, an expert group, convened by ECVAM and consisting of eight toxicologists from national authorities, industry and academia was established. This group accompanied and supported the tool development and assessment over the entire project period (Table 2). Its main tasks were to provide toxicological guidance and feedback in order to safeguard the scientific soundness, appropriateness and applicability of the tool.

The tool-guided assignment to Klimisch categories 1, 2 or 3 under consideration of the red criteria is the primary result of the tool. Once a Klimisch category is assigned on basis of the criteria, the tool offers the option to assign a Klimisch category according to personal judgement. If this personal judgement differs from the tool outcome, the evaluator is requested to document the reasons.

In addition to criteria for assessing data reliability, documentation of observations with importance to relevance is included as an option at the bottom of each worksheet. This documentation does not have an impact on the reliability assessment. It allows the informal recording of observations, made during the evaluation, which may be of importance for future regulatory or other purposes.

2.2. Design of rater experiments

Once prototype 1 was established after consultation with the expert group, the tool was embedded in a Microsoft Office Excel® 2003 file to be evaluated and developed further by means of two inter-rater experiments. In these experiments, the tool was applied to a set of selected case studies by volunteer toxicologists (raters) not involved in the project. Therefore, the electronic version consisted of several worksheets. In the introductory worksheet raters were asked to give some personal details such as name and contact information as well as their overall opinion on the tool after having completed the rater experiment (Fig. 1). The second worksheet

Table 2
Members of the expert group.

Name	Affiliation
Neil Carmichael	ECETOC AISBL, Brussels, Belgium
Karl-Heinz Cohr	DHI Water, Environment, Health, Hørsholm, Denmark, on behalf of EUROTOX, Risk Assessment Speciality Section
Cees de Heer	RIVM Centre for Substances and Integrated Risk Assessment Bilthoven, The Netherlands
Sebastian Hoffmann (coordinator, chair)	(then) ECVAM, Institute for Health and Consumer Protection, EC Joint Research Centre, Ispra, Italy
Dinant Kroese	TNO Quality of Life, Zeist, The Netherlands
Franz Oesch	Institute of Toxicology, University of Mainz, Mainz, Germany, on behalf of EUROTOX, Risk Assessment Speciality Section
Iona Pratt	Food Safety Authority of Ireland, Dublin, Ireland
Hans-Bernhard Richter-Reichhelm	Bundesinstitut für Risikobewertung (BfR), Berlin, Germany

was purely informative (no input required), giving some explanations about the tool. Finally, one worksheet per study had to be filled in.

Following analysis of the findings and conclusions of the first inter-rater experiment, in which prototype 1 was used, an improved prototype 2 was developed. Subjecting prototype 2 to a second inter-rater experiment (with raters different from the first experiment) allowed the effectiveness of the improvements to be assessed.

In consultation with the expert group case studies (11 *in vivo* studies and 11 *in vitro* studies) were selected to represent a wide range of study types and endpoints (Table 3). Furthermore, the intent was that they should represent a broad spectrum of data quality (to allow for statistical evaluation of inter-rater agreement), including studies performed according to guidelines, detailed peer-reviewed publications, older publications from the 1950s and 1960s, short reports as well as studies summarising results for large numbers of substances. Data quality of the *in vivo*/*in vitro* studies, at this stage judged subjectively by one or more scientists from the expert group and/or the project group, was assumed to cover the three Klimisch categories. Raters were asked to evaluate the selected case studies, taking into consideration relevant cross-references cited in the methodological sections of the publications. To allow for comparison, the same set of case studies was used in both inter-rater experiments.

Raters of the first experiment were recruited by approaching scientific groups, industrial associations and the organisations of the expert group's members. For the second inter-rater experiment, ECVAM invited scientists from mailing lists compiled in the context of the Joint Research Centre's initiative towards evidence-based toxicology. In both experiments, different professional sectors (industry, regulatory authorities, and academia) were well represented, and raters were from several, mainly European countries. In the first inter-rater experiment, nine raters evaluated the *in vivo* studies and eleven raters the *in vitro* studies. In the second inter-rater experiment, twelve evaluations of *in vivo* studies and 17 evaluations of *in vitro* studies were performed. As instructed, participating raters evaluated all case studies for a specific subset (*in vitro* or *in vivo*).

Results of the rater experiments were analysed statistically. The kappa coefficient was used as quantitative statistical measure for the assessment of agreement for categorical outcomes. As agreement was evaluated for more than two raters, Fleiss' kappa (Fleiss et al., 2003) was used. The kappa coefficient is a more robust measure than simple percent agreement calculation since it takes into account the agreement occurring by chance. It calculates the degree of agreement in classification over that which would be expected by chance. For both rater experiments, agreement between the raters' tool outcome was assessed, i.e. inter-rater agreement, for the 11 studies (*in vitro* and *in vivo*, respectively) and was denoted as overall kappa. Strength of agreement was judged by means of the interpretation table published by Landis and Koch (1977). They proposed the following as standard for strength of agreement for the kappa coefficient: −1 to 0.0 poor, 0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial and 0.81–1.0 almost perfect agreement.

In order to describe all data obtained from the multi-item questionnaire, bar plots were produced displaying the number of ratings in the three Klimisch categories for each case study. The different reliability categories were depicted by different patterns, and the numbers of ratings given by each rater on the studies were displayed.

To identify criteria with the potential for improvement in the subsequent version of the tool, individual criteria were categorized into three classes: satisfactory, borderline and unsatisfactory, based on the number of indecisive ratings amongst the raters. The number of indecisive ratings for a criterion was determined according to the following algorithm: For every criterion, and for every study, the number of concordant and discordant ratings was determined. If the number of discordant ratings for one study exceeded a pre-specified limit, the rating of the study was labelled as indecisive.

The in-depth analysis of results for all criteria allowed the identification of reasons for differences between rater results.

3. Results

3.1. Inter-rater experiment 1

Substantial heterogeneity was observed in the evaluation results for both the *in vitro* (Fig. 2A) and *in vivo* case studies (Fig. 2B). For most case studies, the categories assigned by the tool as result of individual raters assessments, ranged over all possible reliability categories (1–3). Only three *in vitro* studies and two *in vivo* studies achieved results ranging over two categories and only one *in vivo* study (study no. 6) was consistently considered to be not reliable (category 3) by all nine raters of the *in vivo* part. Overall, heterogeneity was somewhat less for *in vivo* compared to *in vitro* studies. The personal judgement of evaluators on the reliability of case studies showed a similar heterogeneity (Table 4). Personal judgement, which has been asked for after having applied the tool to the case

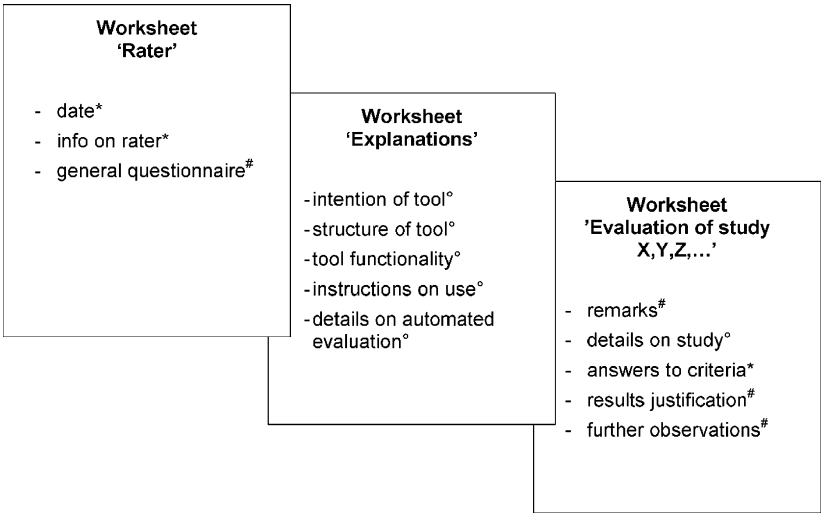

Fig. 1. Schematic overview of ToxRTool prototypes as used in inter-rater experiments (* mandatory data input; # optional data input; ° information (no input)/automated results).

Table 3
Case studies selected for inter-rater experiments.

Study type	No.	Reference	Study type
In vivo studies	1	Ranaldi et al. (2001)	Inhalation genotoxicity study
	2	Belpoggi et al. (1997)	Oral carcinogenicity study
	3	Farr et al. (2001)	Teratogenicity study (short report)
	4	Kuschner et al. (1997)	Human experimental acute inhalation study
	5	Warren (1958)	Acute toxicity data (old study)
	6	Weil et al. (1963)	Skin irritation data (old study)
	7	Harris et al. (1992)	Reproductive toxicity study
	8	Carmichael et al. (2000)	Knock-out model carcinogenicity study
	9	Grassian et al. (2007)	Short-term inhalation study (nanoparticles)
	10	Kitamura et al. (1999)	Toxicokinetic study
	11	Study number ##; Local lymph node assay (LLNA) in mice with ##, 2007, unpublished	Skin sensitisation study (guideline study)
In vitro studies	1	Ishidate et al. (1984)	Genotoxicity (gen mutation) study (concise report on many substances)
	2	Morris and Heflich (1984)	Study on induction of sister chromatide exchanges
	3	Kodama et al. (1980)	Genotoxicity (chromosome aberrations) study
	4	Riddell et al. (1986)	Cytotoxicity study in mammalian cells
	5	Perkins et al. (1996)	Skin corrosion study
	6	Price et al. (1996)	Hepatotoxicity study
	7	Florin et al. (1980)	Genotoxicity (gen mutation) study (concise report on many substances)
	8	Tirelli et al. (2007)	Membrane barrier penetration study
	9	Barcelos et al. (2007)	Comet assay
	10	Kejlová et al. (2007)	Phototoxicity study
	11	Kandárová et al. (2006)	Skin corrosion study

studies and therefore cannot be considered independent of the tool outcome, differed from the tool outcome only in a few cases.

In-depth analysis of the evaluations revealed that heterogeneity between raters was caused by failure of identifying available information in the study report, by including aspects of relevance in the reliability evaluation, by different judgements on study quality and by ambiguous phrasing of criteria, which were then understood differently by raters.

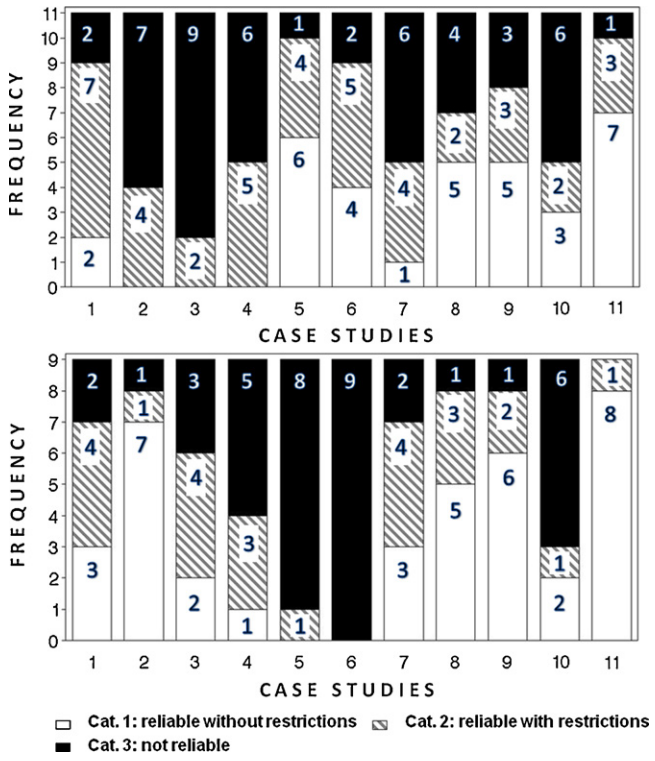

Fig. 2. Results of inter-rater experiment 1: reliability categories obtained by raters when using the tool for A: *in vitro* and B: *in vivo* studies (figures in bars for each case study are numbers of raters with the respective outcome, study numbers are according to Table 3).

Based on the results of the first inter-rater experiment, the assessment tool was modified, linguistically changed and a second version of the tool consisting of 18 (*in vitro*), respectively, 21 (*in vivo*) criteria was developed. In the second testing phase, a new set of raters applied this second version of the tool to the same set of case studies as chosen in the first testing phase.

3.2. Inter-rater experiment 2

The heterogeneity in rater results was lower as compared to the first inter-rater experiment, but was still substantial (see Table 4). As in the first inter-rater experiment, variability was higher for the *in vitro* part than for the *in vivo* part.

Differences between raters were more pronounced for reliability categorisation (Fig. 3A and B for *in vitro* and *in vivo* studies, respectively) than for total scores (Fig. 4A and B for *in vitro* and *in vivo* studies, respectively). This was caused by a high number (compared to inter-rater experiment 1) of red criteria answered with "no", thus leading to reliability category 3 despite a high total score. For example, the *in vivo* case study by Ranaldi et al. (2001) obtained total scores of at least 60% (i.e. assignment to category 1 or 2) from all raters, but was assigned to category 3 by five raters due to failing one or more red criteria. Reasons were: one out of 12 raters failed to identify the test substance, two raters missed in the publication information on test concentrations, and two raters considered that positive controls are necessary for this type of study. Two further raters found the study design inappropriate to achieve the study objectives. As two of these raters each identified two red criteria as not being met, this led to reliability category 3 in five evaluations in total.

In-depth analysis showed that approximately 67% (12 of 18 criteria) and 81% (17 of 21 criteria) for the *in vitro* and the *in vivo* part, respectively, were answered "satisfactorily" in prototype 2, compared to 56% (14 of 25 criteria) and 60% (15 of 25 criteria) for the *in vitro* and the *in vivo* part of prototype 1. Unsatisfactorily answered criteria were often those, which were relevant for some study types, but not or only partly applicable to other study types. For example, the question on documentation of animal housing and feeding conditions of animals led to widely different responses from raters in the case of acute toxicity studies. Generally, this information is considered less relevant for acute studies than for long-term studies.

Table 4
Results of the statistical evaluation of inter-rater experiment 1 and 2: overall kappa for tool-guided Klimisch categorisation and for rater's personal judgement (*N*: number of raters).

	In vitro				*In vivo*			
	N	Kappa	95% confidence interval	Interpretation	*N*	Kappa	95% confidence interval	Interpretation
1. Inter-rater experiment								
Tool outcome	11	0.1033	[0.047, 0.160]	Slight agreement	9	0.2473	[0.177, 0.318]	Fair agreement
Personal judgement	11	0.1354	[0.079, 0.192]	Slight agreement	9	0.1846	[0.114, 0.255]	Slight agreement
2. Inter-rater experiment								
Tool outcome	17	0.1506	[0.107, 0.194]	Slight agreement	12	0.2953	[0.242, 0.348]	Fair agreement
Personal judgement	17	0.1868	[0.149, 0.224]	Slight agreement	12	0.1710	[0.119, 0.223]	Slight agreement

However, for such a study an evaluator strictly applying the criterion would probably have considered it as not fulfilled, while an evaluator applying this criterion in a more flexible way would have considered it as met. In the *in vitro* part of the tool, the (red) criterion addressing positive controls was a major source of heterogeneity. Positive controls were not included in several of the case studies, but in a number of these cases (e.g. in the presence of other substances tested in parallel with positive results) raters considered the absence of positive controls not as indispensable. In addition, questions focusing on study results and plausibility of study design were answered more heterogeneously in the case of *in vitro* studies compared to *in vivo* studies.

3.3. Consequences for the final tool version: ToxRTool

Observations from the inter-rater experiment 2 were used to adjust and improve the tool. Only a few criteria were found to contribute to the heterogeneity of rater results by causing misunderstandings or ambiguous interpretation. For those criteria either the explanations were modified or ambiguous parts were re-phrased. For example, it was observed that the term "study objectives" was interpreted in several ways by raters and was mixed up with overall objectives for using the data ("data relevance"). Therefore, criterion 20 of the *in vitro* part (and also the respective criterion 17 of the *in vivo* part) "*Is the study design chosen appropriate for the study objective*?" was changed to "*Is the study design chosen appropriate for obtaining the substance-specific data aimed at*?". Despite these linguistic improvements the final tool, called ToxRTool, is very close to prototype 2, whose performance was tested in the second inter-rater experiment.

The criteria (and accompanying explanations) of ToxRTool are available online as supplementary material to this publication.

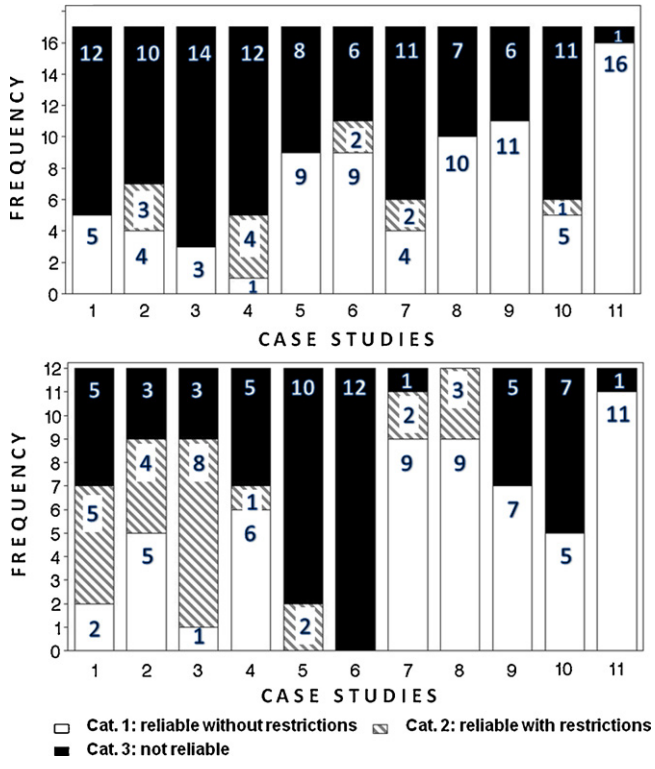

Fig. 3. Results of inter-rater experiment 2: reliability categories obtained by raters when using the tool for A: *in vitro* and B: *in vivo* studies (figures in bars for each case study are numbers of raters with the respective outcome, study numbers according to Table 3).

4. Discussion

In the present study, a tool aimed to increase the transparency and to harmonise approaches in reliability assessment of *in vivo* and *in vitro* toxicological data has been developed in a step-wise approach and tested in two independent inter-rater experiments. The study allowed also to get an insight in the possible sources of variability in the quality assessment of toxicological data that, to our knowledge, has not been carried out before. Hobbs et al. (2005) published results on similar, small-scale inter-rater experiments applied to data quality criteria for ecotoxicological data from two publications. A list of 20 criteria was established, with criteria weighted differently (3, 4, 5 or 10 points) and applied to the two case studies by 23 raters. Boundaries between "high" and "acceptable quality" and between "acceptable" and "unacceptable quality" were set at 80% and 50% of total score, respectively. Based on the observations from the first trial the criteria were slightly modified and re-applied by seven of the 23 raters to the same case studies, resulting in 16% reduction in variation. This reduction is not discussed in detail (e.g. in terms of statistical significance), but it has to be assumed that by re-choosing rater in the second trial, a dependency is introduced, which does not allow to unambiguously relate this reduction to the modifications of the criteria. In addition, both case studies were judged by Hobbs et al. and by most of the raters as being of acceptable quality. Therefore, performance of the criteria when applied to studies belonging to other categories or being borderline could not be assessed.

In the present study, the thorough testing of our tool in two independent inter-rater experiments allowed for a detailed analysis of various aspects of reliability evaluations. Different reasons can be discerned for the substantial heterogeneity observed in the evaluation results: first, instructions were not followed or errors were made when answering criteria; second, raters did not always succeed in addressing reliability only, but considered also aspects of relevance; third, raters had differences in opinion about data quality (with respect to weighting of pieces of information or to plausibility of study design and results), and finally, some criteria were ambiguously phrased causing different interpretations and/or misinterpretation.

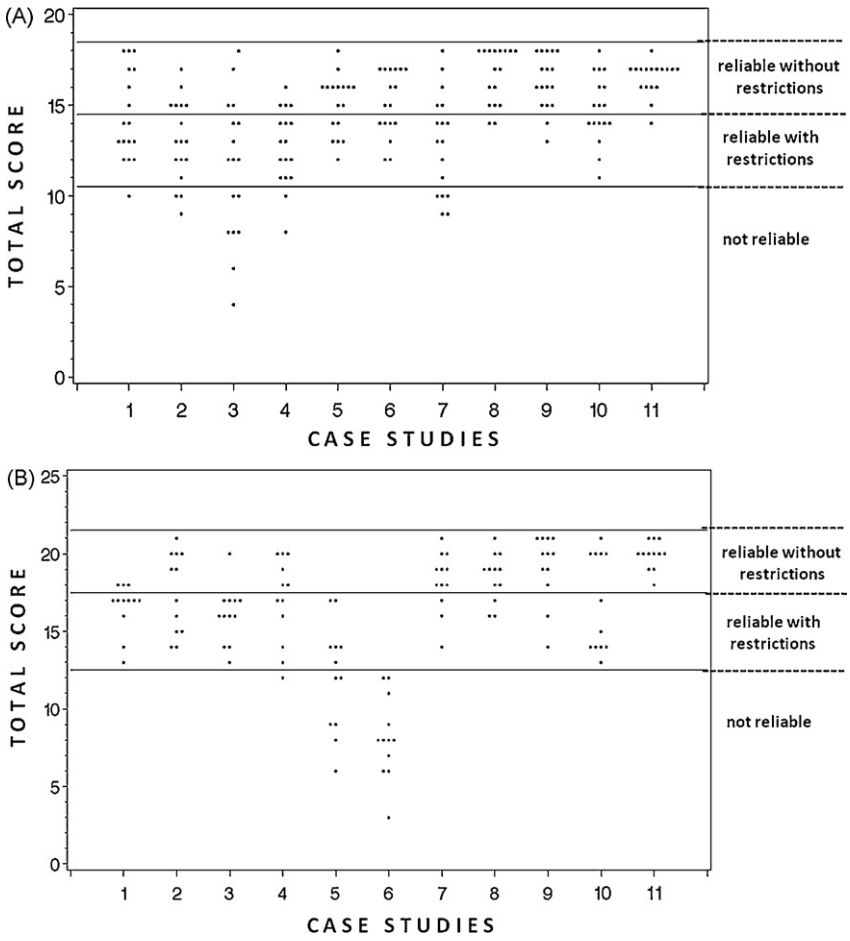

Fig. 4. Results of inter-rater experiment 2: total scores (TS) for A: *in vitro* and B: *in vivo* studies (each point represents a result obtained by a rater for a specific case study, study numbers according to Table 3).

The guidance given on how to use the tool, the choice of wording, grammar, the scope and logical complexity of questions, and in particular the weighting of the questions were identified as critical issues that influenced the heterogeneity. Training might improve the handling of the tool and decrease the frequency of individual errors. Another source of heterogeneity is the degree to which raters included elements of relevance and adequacy into their rating and how they weighted those against reliability. The concepts of reliability and relevance and their discrimination need to be disseminated and discussed more thoroughly.

With respect to the third source of heterogeneity, the tool can help to identify differences in opinions between evaluators and can facilitate discussions about their reasons. This may be especially helpful in substance-information-exchange-fora (SIEFs), where exchange of data is negotiated between potential registrants of substances under REACH. The tool can be used to transparently discuss and document the value of a study for a registration dossier.

Heterogeneity arising from criteria ambiguously phrased was step-by-step minimised taking into account the observations made by the raters during the experiments. Especially, criteria applicable to only some but not all study types were identified as sources of variability and were carefully considered during the stepwise improvement of the tool. Rephrasing of some criteria led to more homogenous results for the total scores of *in vivo* studies. Results were less homogenous with respect to reliability categories. This difference can be partly explained by the way the "red criteria" were used: several raters did not consider these pivotal criteria with the detailed scrutiny needed. Individual mistakes (e.g. failure to find the respective information in the study report, albeit available) occurred with similar frequency for red and normal criteria. Whether this situation prevails when ToxRTool is used routinely for regulatory purposes should be the subject of further investigations.

Two observations made during the second inter-rater experiment may lead to further modifications of the tool in the future, if supported by more experience. First, in the *in vitro* part, the (red) criterion requiring the use of positive controls was often the reason for assigning data to reliability category 3. Several raters expressed doubts that the lack of positive controls should result in this outcome. If the tool would follow their argumentation, this criterion should be changed from a red criterion to a "normal" one. Another potential change concerned the last group of criteria ("plausibility of study design and results") in both parts of the tool. Several raters in their comments pointed out that criteria checking completeness of documentation have too much weight compared to criteria addressing plausibility of the study. To correct this imbalance when further developing the tool the score for the two criteria in this group addressing plausibility could be increased.

In both inter-rater experiments heterogeneity of results was higher for *in vitro* studies than for *in vivo* studies. Based on the observation that heterogeneity was especially high for criteria asking for plausibility of study design it seems that the lack of generally accepted guidelines for some types of *in vitro* studies lead to uncertainty about suitable study designs. Nevertheless, it is apparent that in some of the *in vitro* case studies the basic information was insufficiently documented. This leads us to the conclusion that the requirements for publication of *in vitro* studies should be improved or made more stringent. By considering the criteria as minimum

documentation requirements, a further benefit from ToxRTool may be to provide guidance for publication in scientific journals.

In addition to the evaluation exercise, raters were asked to indicate the time needed to apply the tool. Time for application of the tool (not considering the time to read and understand the case study report) was on average about 20 min. This is considered acceptable in terms of practicability for routine use.

Asked for their general opinion on the tool, a great majority of raters in both experiments (and for both parts of the tool) found it useful (90%), user-friendly (82%) and transparent (78%). Comments made by raters clearly show that the guidance provided (on the type of issues to be considered and the systematic way of covering them) was appreciated. Some raters asked for a more differentiated tool, e.g. development of tools for specific study types (e.g. reproductive toxicity studies) or for specific regulatory purposes (e.g. classification and labelling). Some raters suggested that the tool should specifically consider whether data were obtained in compliance with recent guidelines and under GLP conditions. This proposal was not taken up as ToxRTool follows the approach to treat all kind of data equally and to base the decision on the reliability of data using only information provided in the study report. Based on the experiences made with the tool, studies carried out according to recent guidelines under GLP conditions will generally be assigned to reliability category 1. Hence, no extra consideration for these studies, e.g. by an additional criterion, seems necessary and justifiable.

The process of evaluation and improvement of ToxRTool in a stepwise approach was both feasible and successful. The development of ToxRTool provided valuable insight into the critical components of the evaluation of reliability of toxicological data and revealed several sources of heterogeneity that might influence this evaluation. The background of the evaluator determined, e.g. by country of origin, education, professional experience as well as a differential understanding of the questions in a non-native language may be an important source of heterogeneity.

Moreover, during the assessment of the tool it became evident that a better guidance tailored for toxicological studies is needed to facilitate a clear distinction between reliability and relevance in toxicological studies. ToxRTool was developed to improve transparency in the assessment of reliability of toxicological data and can be seen as a prototype for further development of rating instruments with the potential to be integrated into the formal consensus processes of evidence-based toxicology (EBT).

The ToxRTool has been made publicly available from the ECVAM website (http://ecvam.jrc.it, section "Publications") to encourage testing of its practical applicability.

Acknowledgements

The project was funded by ECVAM (contract no. CCR_IHCP_C433199_XO). The valuable comments to the manuscript by Claudius Griesinger (ECVAM) are appreciated. We are most grateful to the members of the expert group and all participants of the inter-rater experiments for their contributions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.toxlet.2009.05.013.

References

Barcelos, G.R., Shimabukuro, F., Maciel, M.A., Cólus, I.M., 2007. Genotoxicity and antigenotoxicity of cashew (*Anacardium occidentale* L.) in V79 cells. Toxicol. In Vitro 21, 1468–1475.

Belpoggi, F., Soffritti, M., Filippini, F., Maltoni, C., 1997. Results of long-term experimental studies on the carcinogenicity of methyl tert-butyl ether. Ann. N. Y. Acad. Sci. 837, 77–95.

Carmichael, N.G., Debruyne, E.L.M., Bigot-Lasserre, D., 2000. The p53 heterozygous knockout mouse as a model for chemical carcinogenesis in vascular tissue. Environ. Health Perspect. 108, 61–65.

ECHA, European Chemicals Agency, 2008. REACH "Guidance on Information requirements and Chemical Safety Assessment", Chapter R.4 "Evaluation of available information" http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_en.htm.

Farr, C.H., Reinisch, K., Holson, J.F., Neubert, D., 2001. Potential teratogenicity of di-n-butyltin dichloride and other dibutyltin compounds. Teratogen. Carcinogen. Mutagen. 21, 405–415.

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical Methods for Rates and Proportions, 3rd ed. John Wiley, New York.

Florin, I., Rutberg, L., Curvall, M., Enzell, C.R., 1980. Screening of tobacco smoke constituents for mutagenicity using the Ames' test. Toxicology 18, 219–232.

Grassian, V.H., O'Shaughnessy, P.T., Adamcakova-Dodd, A., Pettibone, J.M., Thorne, P.S., 2007. Inhalation exposure study of titanium dioxide nanoparticles with a primary particle size of 2 to 5 nm. Environ. Health Perspect. 115, 397–402.

Harris, M.W., Chapin, R.E., Lockhart, A.C., Jokinen, M.P., 1992. Assessment of a short-term reproductive and developmental toxicity screen. Fundam. Appl. Toxicol. 19, 186–196.

Hobbs, D.A., Warne, M.S., Markich, S.J., 2005. Evaluation of criteria used to assess the quality of aquatic toxicity data. Integr. Environ. Assess. Manage. 1, 174–180.

Hoffmann, S., Hartung, T., 2006. Toward an evidence-based toxicology. Hum. Exp. Toxicol. 25, 497–513.

Ishidate, M., Sofuni, T., Yoshikawa, K., Hayashi, M., Nohmi, T., Sawada, M., Matsuoka, A., 1984. Primary mutagenicity screening of food additives currently used in Japan. Food Chem. Toxicol. 22, 623–636.

Kandárová, H., Liebsch, M., Spielmann, H., Genschow, E., Schmidt, E., Traue, D., Guest, R., Whittingham, A., Warren, N., Gamer, A.O., Remmele, M., Kaufmann, T., Wittmer, E., De Wever, B., Rosdy, M., 2006. Assessment of the human epidermis model SkinEthic RHE for *in vitro* skin corrosion testing of chemicals according to new OECD TG 431. Toxicol. In Vitro 20, 547–559.

Kejlová, K., Jírová, D., Bendová, H., Kandárová, H., Weidenhoffer, Z., Kolárová, H., Liebsch, M., 2007. Phototoxicity of bergamot oil assessed by *in vitro* techniques in combination with human patch tests. Toxicol. In Vitro 21, 1298–1303.

Kitamura, S., Okamoto, Y., Takeshita, M., Ohta, S., 1999. Reductive metabolism *in vivo* of trans-4-phenyl-3-buten-2-one in rats and dogs. Drug Metab. Dispos. 27, 767–769.

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

Kodama, F., Fukushima, F., Umeda, M., 1980. Chromosome aberrations induced by clinical medicines. J. Toxicol. Sci. 5, 141–150.

Kuschner, W.G., Wong, H., D'Alessandro, A., Quinlan, P., Blanc, P.D., 1997. Human pulmonary responses to experimental inhalation of high concentration fine and ultrafine magnesium oxide particles. Environ. Health Perspect. 105, 1234–1237.

Landis, J.R., Koch, G.G., 1977. The measurement on observer agreement for categorical data. Biometrics 33, 159–174.

Morris, S.M., Heflich, R.H., 1984. A comparison of the toxic and SCE-inducing effects of inhibitors of ADP-ribosyl transferase in Chinese hamster ovary cells. Mutat. Res. 126, 63–71.

Perkins, M.A., Osborne, R., Johnson, G.R., 1996. Development of an *in vitro* method for skin corrosion testing. Fundam. Appl. Toxicol. 31, 9–18.

Price, R.J., Mistry, H., Wield, P.T., Renwick, A.B., Beamand, J.A., Lake, B.G., 1996. Comparison of the toxicity of allyl alcohol, coumarin and menadione in precision-cut rat, guinea-pig, cynomolgus monkey and human liver slices. Arch. Toxicol. 71, 107–111.

Ranaldi, R., Bassani, B., Pacchierotti, F., 2001. Genotoxic effects of butadiene in mouse lung cells detected by an *ex vivo* micronucleus test. Mutat. Res. 491, 81–85.

Riddell, R.J., Clothier, R.H., Balls, M., 1986. An evaluation of three *in vitro* cytotoxicity assays. Food Chem. Toxicol. 24, 469–471.

SOT, Society of Toxicology, 2007. A new language for toxicologists: globally harmonized system of classification and labelling of chemicals (GHS): what do you know about GHS? Comments prepared on behalf of the Society of Toxicology by members of the Occupational and Public Health Specialty Section, http://www.toxicology.org/AI/NEWS/GHS_Comments.pdf.

Tirelli, V., Catone, T., Turco, L., Di Consiglio, E., Testai, E., De Angelis, I., 2007. Effects of the pesticide clorpyrifos on an *in vitro* model of intestinal barrier. Toxicol. In Vitro 21, 308–313.

US EPA, Environmental Protection Agency, 1999. Determining the adequacy of existing data. High Production Volume (HPV) Challenge, http://www.epa.gov/hpv/pubs/general/datadfin.htm.

Warren, K.S., 1958. The differential toxicity of ammonium salts. J. Clin. Invest. 37, 497–501.

Weil, C.S., Condra, N., Haun, C., Striegel, J.A., 1963. Experimental carcinogenicity and acute toxicity of representative epoxide. Am. Ind. Hyg. Assoc. J. 24, 305–325.