Regulatory Toxicology and Pharmacology 88 (2017) 227–237

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Relevance and reliability of experimental data in human health risk assessment of pesticides

Johanna Kaltenhäuser*, Carsten Kneuer, Philip Marx-Stoelting, Lars Niemann, Jens Schubert, Bernd Stein, Roland Solecki

German Federal Institute for Risk Assessment (BfR), Max-Dohrn-Str. 8-10, 10589 Berlin, Germany

ARTICLE INFO

Article history:
Received 30 December 2016
Received in revised form
19 June 2017
Accepted 22 June 2017
Available online 24 June 2017

Keywords:
Relevance
Reliability
Uncertainty
Test guidelines
Publication
Risk assessment
Pesticides
Plant protection products
Biocides
Toxicology

ABSTRACT

Evaluation of data relevance, reliability and contribution to uncertainty is crucial in regulatory health risk assessment if robust conclusions are to be drawn. Whether a specific study is used as key study, as additional information or not accepted depends in part on the criteria according to which its relevance and reliability are judged. In addition to GLP-compliant regulatory studies following OECD Test Guidelines, data from peer-reviewed scientific literature have to be evaluated in regulatory risk assessment of pesticide active substances. Publications should be taken into account if they are of acceptable relevance and reliability. Their contribution to the overall weight of evidence is influenced by factors including test organism, study design and statistical methods, as well as test item identification, documentation and reporting of results. Various reports make recommendations for improving the quality of risk assessments and different criteria catalogues have been published to support evaluation of data relevance and reliability. Their intention was to guide transparent decision making on the integration of the respective information into the regulatory process. This article describes an approach to assess the relevance and reliability of experimental data from guideline-compliant studies as well as from non-guideline studies published in the scientific literature in the specific context of uncertainty and risk assessment of pesticides.

1. Introduction

The human health risk assessment of pesticides is an essential part of the approval of active substances (AS) or the authorisation of plant protection products (PPP) and biocidal products (BP) in Europe according to the European legislation (Regulations (EC) No 1107/2009 and (EU) No. 528/2012). Detailed listings of all data requirements are part of this legislation (e.g. Regulations (EU) No. 283/2013, (EU) No. 284/2013 (EU, 2013a; EU, 2013b)). In a complete dossier all data requirements have to be addressed by the applicant. This can be achieved either by using studies performed according to test guidelines and under GLP, which are often property of the applicant and remain unpublished, or based on research studies published in the scientific literature. In any case, data used for regulatory decisions have to be appropriate for the respective purpose (relevant) and trustworthy because of their quality (reliable).

Very often 200 studies or more are submitted for the assessment of human health, including toxicology, residues, application safety and classification & labelling. This does not take into account the assessment of the AS for efficacy and environmental effects, which can easily double this amount.

The evaluation of data reliability itself is a key point which can influence data selection, and thereby also the credibility and usefulness of a regulatory assessment. Therefore, a transparent evaluation tool for determining the relevance and reliability of study results is necessary.

Mandatory studies according to data requirements have to be performed according to harmonised OECD test guidelines (TG) or EU test methods. Furthermore, these studies have to be conducted according to Good Laboratory Practice (GLP) principles. Such studies are described in the following as “guideline-compliant studies”. In addition, current EU legislation mandates regulatory agencies to take published data (e.g. peer-reviewed scientific publications) into consideration for human health risk assessment of pesticides (EC, 2009; EU, 2012). A literature search and review of the available publications has therefore become a mandatory part

* Corresponding author.
E-mail address: johanna.kaltenhaeuser@bfr.bund.de (J. Kaltenhäuser).

http://dx.doi.org/10.1016/j.yrtph.2017.06.010

of the regulatory process. In contrast to the prescribed endpoints for obligatory experimental studies, many scientific publications do neither adhere to harmonized TGs nor to GLP principles. Such studies are described in the following as “non-guideline studies”.

These non-guideline studies can constitute an important part of the database used for risk assessment, especially for previously approved substances with a long history of use. In contrast, for newly developed synthetic molecules or micro-organisms, such published data are often very limited. As a consequence, the databases for regulatory risk assessment consist of a mix of guideline-compliant studies as well as non-guideline studies to varying degrees.

The available scientific information is then subject to evaluation by member state and EU authorities (EFSA, ECHA) and provides the basis for the resulting risk assessment report. In this report, the data are presented and evaluated, and the conclusions drawn by the respective authority for the proposed use are stated.

Key characteristics of a high-quality risk assessment including transparency, reproducibility and usefulness were recently summarized and integrated into the “*Guide for Judging the Quality of an Assessment*” (Fenner-Crisp and Dellarco, 2016).

The evaluation of the quality of data on which regulatory decisions are based is a crucial point. The present paper aims to propose and discuss criteria for relevance and reliability of toxicological data that should be considered when information is used for regulatory purposes. These criteria were compared with those that are included in chosen existing tools for study evaluation as well as with principles laid out in OECD TGs. Focus was put on experimental toxicological studies with pesticides, especially non-guideline studies, leaving aside epidemiological, residue or environmental studies, although it is expected that the same basic principles and analogous criteria could be applied to these studies also.

2. Systems for evaluation of data quality

The criteria used for the assessment of relevance and reliability are a central issue in the process of systematic literature reviews (see Fig. 1).

Different systems have been developed and applied for the evaluation of data quality in the past. For the assessment of chemicals, a now widely known system was developed by Klimisch et al. (1997). In this approach the most important parameter for unrestricted reliability was seen in the adherence to harmonised TGs and GLP principles. Studies are assigned to four categories: 1 - Reliable without restriction; 2 - Reliable with restriction; 3 - Not reliable; 4 - Not assignable. Today, a modification of these principles is recommended by ECHA for the assessment of biocide AS, as well as for chemicals under REACH (ECHA, 2011; ECHA, 2015). One important criticism of the criteria in Klimisch et al. (1997) is that they introduce a bias in favour of the use of GLP- and TG-studies (Buonsante et al., 2014; Myers et al., 2009). Critics also claim that when these criteria are applied without adjustment to non-guideline studies, results may often be categorized as “reliable with restriction” or “not reliable”, despite being of high scientific value.

In the EU, a wide consensus was reached among member states that categories leading to decisions on reliability have to be filled with more specific, transparent and appropriate criteria (EC, 2015). Thus, further development and harmonisation of criteria is urgently needed.

Several tools have been developed to assess the reliability of studies, including ToxRTool (Schneider et al., 2009) and SciRAP (Molander et al., 2014). Both consist of a series of specific questions concerning key points of the described experiments, which have to be answered by scoring. These systems allow a more transparent documentation of the study evaluation by the assessor and do not emphasize the use of harmonised TGs.

ToxRTool (Toxicological data Reliability assessment Tool) is an MS Excel based tool with comprehensive systems for scoring of *in vitro* as well as *in vivo* studies. It makes clear reference to the four categories used by Klimisch, but contains a more specifically phrased questionnaire (Schneider et al., 2009).

SciRAP (Science in Risk Assessment and Policy), which focuses on *in vivo* studies, proposes a more integrated approach allowing assessment of both relevance and reliability. It uses scoring for the questions, which are separated into reporting quality and methodological quality, but does not lead to a final score for the whole study. According to the authors, one of the reasons is to avoid dismissal of studies as a result of too strict criteria (Beronius et al., 2014; Molander et al., 2014).

A very broad and comprehensive overview on frameworks used for evaluating relevance and reliability has recently been published by Roth and Ciffroy (2016). Ågerstrand and Beronius (2016) reviewed the regulatory basis for the implementation of systematic review approaches in many regulatory fields.

3. Relevance

Relevance evaluation determines whether a study or publication should be included or excluded for a specific regulatory purpose or whether a weight of evidence approach should be used when addressing a precisely formulated question. In systematic review approaches, an initial relevance check is carried out based on titles and abstracts of retrieved literature. *Per se*, all data that contain information on the substance or product under assessment and that concern the problem under assessment are relevant. However, the actual use for regulatory purposes depends also on reliability of the data.

According to EFSA, studies relevant for regulatory purposes are those that address the data requirement(s) set out in the respective regulations on hazard identification, hazard characterisation or exposure assessment (EFSA, 2011). ECHA defines relevance as “the extent to which data and tests are appropriate for a particular hazard identification or risk characterization” (ECHA, 2011). It is important to understand that the relevance of a study depends mainly on the scientific or regulatory question under assessment and the suitability of the study to address this question. Studies meeting regulatory data requirements will be most likely considered relevant but relevance is not confined to those. In contrast, studies which exceed data requirements or address additional issues may be also of scientific and regulatory importance.

Important criteria for assessing the relevance of information for toxicological risk assessment have been proposed in three guidance documents for chemicals, PPP and BP (ECHA, 2011; ECHA, 2015; EFSA, 2011). Based on these approaches, a set of questions addressing relevance was compiled (Table 1), which has to be addressed prior to reliability within an iterative process. If the study is considered not relevant, it will not be necessary to assess its reliability.

4. Reliability

Reliability evaluation influences the weight that is attributed to the presented results. Consequences of reliability scoring depend upon the whole data package and have to be decided case-by-case for each dossier. Even when no studies or publications of unrestricted reliability are available, a weight of evidence evaluation can still allow one to draw sound and robust conclusions from available and congruent data with restricted reliability (ECHA, 2010).

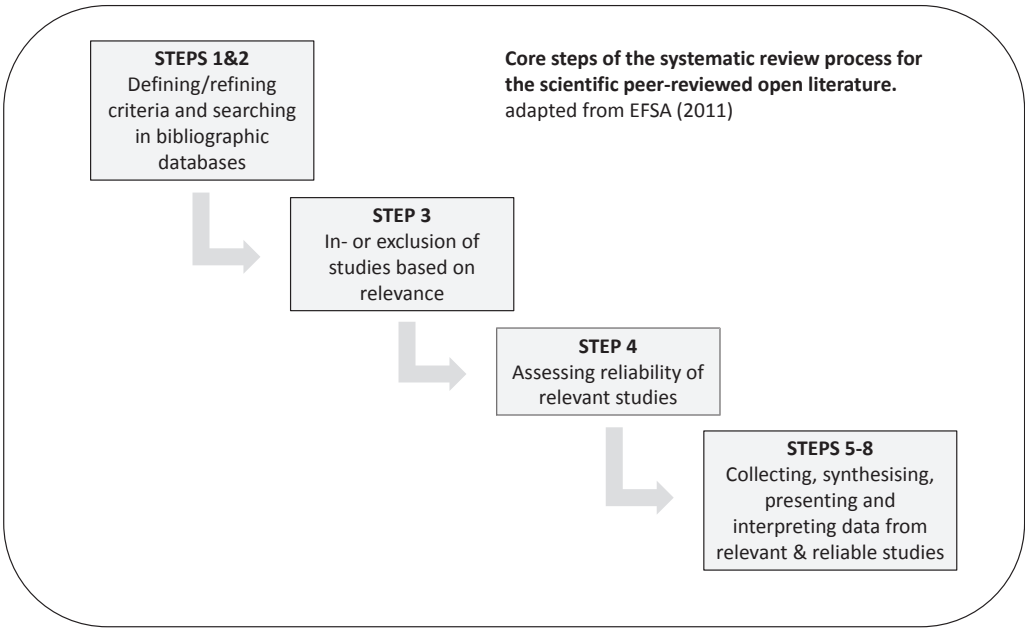

Fig. 1. Core steps of the systematic literature review process for scientific peer-reviewed open literature in pesticides assessment as proposed by EFSA. Adapted from EFSA (2011). The first three steps result in the identification of publications with relevance to the assessed question by suitable search strategies. Subsequently, a reliability evaluation is performed (step 4). Steps five to eight concern data collection and synthesis, presentation and interpretation of results as well as final conclusions to be drawn. All studies assessed as relevant and reliable have to be included into the risk assessment. As a standard procedure, such a systematic literature review is performed for the respective period of interest usually 10 years before application.

Table 1
Compilation of relevance criteria for toxicological studies in regulatory assessments.

Questions helping to evaluate the relevance for regulatory purposes
1 Which regulatory question is addressed? Is the study adequate for addressing the formulated regulatory questions (e.g. hazard identification, hazard characterisation, derivation of reference values, mechanistic aspects, interspecies differences, etc.)?
2 Is the test item representative for the substance/mixture under evaluation? It should appear plausible that the results obtained with the test item can reflect the properties of the substance or product under assessment (for transferability criteria see e.g. EC, 2012; EU, 2012).
3 Is the test system suitable for addressing the formulated regulatory questions? Characteristics of the study design, including test organism/*in vitro* model, have to be taken into account with respect to the regulatory question addressed.
4 Are the methods adequate for the investigation of the endpoint(s)? The methods used may be more or less relevant for addressing the regulatory question. For example, effects like activation of a signalling cascade measured by mRNA expression of marker genes may point to proliferative changes, but are considered less relevant than the proof on cellular level.
5 Is the route of exposure suitable to sufficiently characterise a potential effect? For a thorough toxicological evaluation, data on all relevant exposure routes are needed. For pesticides especially, oral route for dietary exposure and dermal/inhalation for application safety are important.
6 Are test concentrations relevant in the context of the addressed problem? The applied doses should be appropriate to evaluate the risk with regard to the expected exposure. Although TGs may require the highest dose to induce adverse effects independent of expected exposure, for the prediction of risks at the level of expected human exposure such excessive doses may be irrelevant.
7 Is the duration of exposure appropriate for the endpoint(s) being investigated? The duration of exposure should be appropriate for addressing the particular endpoint (e.g. acute response vs. chronic response, developmental susceptibility windows, etc.).

According to EFSA, study reliability "concerns methodological quality and refers to the extent to which a study is free from bias and its findings reflect true facts" (EFSA, 2011). ECHA defines reliability as "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. Reliability of data is closely linked to the reliability of the test method used to generate the data" (ECHA, 2011; Klimisch et al., 1997).

Thus, in contrast to relevance, reliability is an inherent property of a study, which includes the use of well-founded scientific approaches, the avoidance of bias within the study design and faithful study conduct and documentation.

4.1. Good Laboratory Practice (GLP) and Good Scientific Practice (GSP)

The implementation of "Good Laboratory Practice" (GLP) principles was originally motivated by fraudulent practices leading to falsified study results being fed into regulatory processes (Budiansky, 1983). GLP aimed to preclude such practices and improve the quality of regulatory decisions. Adherence to GLP principles is also an essential basis for the "mutual acceptance of data", intended, among other purposes, to avoid duplication of animal testing (OECD, 1997). Despite these clear advantages, putting a high weight on adherence to GLP principles has been heavily criticised as this would lead to a lower acceptance of non-GLP

studies for regulatory purposes, thereby excluding or disregarding many non-guideline studies from research institutes as these might not hold GLP certificates (Buonsante et al., 2014; Myers et al., 2009). Recommendations for "Good Scientific Practice" (GSP), which can to some degree be considered an equivalent to GLP recommendations, have been developed for the academic environment, e.g. by the German research council (*Deutsche Forschungsgemeinschaft*, DFG) or the Second World Conference on Research Integrity in 2010 (DFG, 2013; ESF, 2011). These recommendations give guidance on the documentation and storage of primary and secondary data, on publication of results and require that institutions implement rules and structures assuring compliance to given standards. Similar GSP rules are increasingly implemented around the globe (BBSRC, 2013; MRC, 2012; NHMRC, 2007; NSF, 2009; PRCR, 2011). In addition, criteria for good reporting of animal studies can be found in the ARRIVE guideline or the gold standard publication checklist (GSPC) (Hooijmans et al., 2010; Kilkenny et al., 2010). In contrast to GLP, however, there are typically no audits to monitor adherence to GSP rules. Still, many scientific journals today require a declaration by the authors that these practices were respected. Such a declaration may be considered an appropriate equivalent to a GLP compliance statement and should suffice to assure an appropriate level of raw data documentation in the laboratory even though raw data is not necessarily part of the publication.

Nevertheless, neither adherence to GLP nor to GSP guarantees methodological quality or error-free experimentation and data analysis. These aspects require further attention in the reliability assessment.

4.2. Evaluation of guideline-compliant studies

For guideline-compliant studies, the respective TGs and guidance documents provide harmonised "check lists" to assess the reliability of the study report. Deviations may reduce study reliability as described in Fig. 2. The reliability of a study may be enhanced if the extent and sources of uncertainty are clearly addressed. Current guideline-compliant studies need to fulfil the following points that support the reliability of their data:

i) The experiments are carried out under GLP conditions.
ii) The methods used have been validated and their comparability, reproducibility, specificity and sensitivity has been confirmed.
iii) All data are submitted in a non-aggregated form, allowing one to check the results of single measurements (e.g. blood parameters or body weight).
iv) Information considered necessary for judging reliability of the presented data is requested in chapter "Data and Reporting" of the OECD TGs.

If no deviations from the TG are evident, reliability of a guideline-compliant study may be assumed. Deviations from TGs will not always have an impact on the assessment of an endpoint and studies carried out according to outdated TGs may deviate from current ones, but often still contain all necessary information for acceptance. If deviations are identified, they have to be reported, explained and transparently justified. If a study is flawed by severe deficiencies, it may become unacceptable for regulatory purposes (see Fig. 2). This appraisal is subject to expert judgement.

4.3. Evaluation of non-guideline studies

Evaluation of reliability and inclusion of non-guideline studies in a regulatory context needs to be performed and documented in a transparent and reproducible manner that is generally comparable to existing approaches for evaluation of guideline-compliant studies.

Most published toxicological studies have not been designed specifically to screen for hazard properties of chemical substances or to satisfy regulatory data requirements. Thus, the methodology used in the most published toxicological studies varies considerably and cannot easily be compared with the recommendations of a (OECD) TG.

Only in certain areas may general quality criteria, similar to those for guideline-compliant studies, be applicable to published studies (for example when standardised study designs are widely used as per *in vitro* gene mutation assays).

Instead, for the majority of studies, more generic criteria are needed. Such criteria are proposed in the following section. A process scheme is presented in Fig. 3, analogous to the one for guideline-compliant studies.

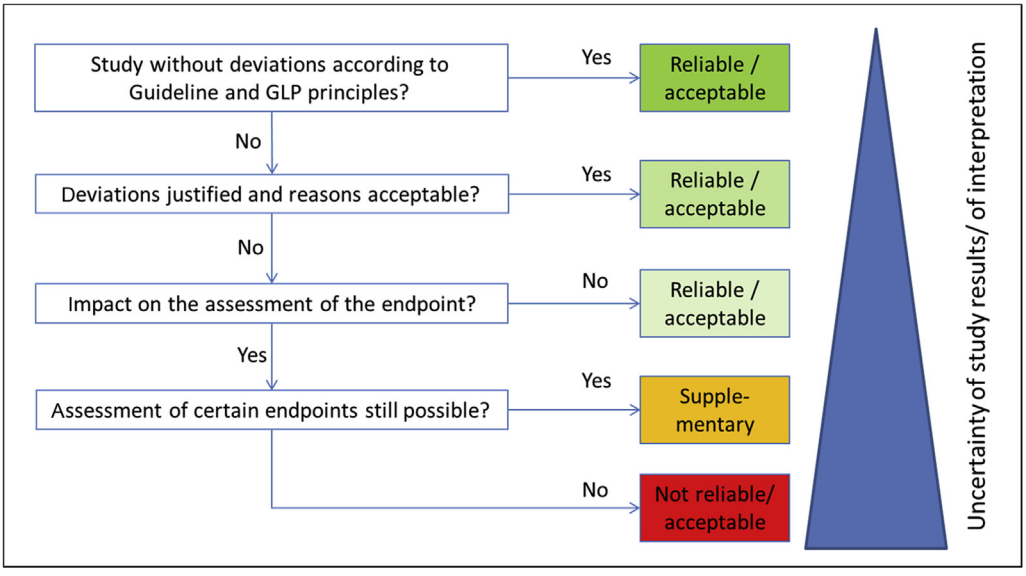

Fig. 2. Decision tree for the acceptability of guideline-compliant studies (expert judgement).

4.4. Criteria for guideline-compliant and non-guideline studies

Criteria for the evaluation of reliability of guideline-compliant or non-guideline studies are grouped into five topics and discussed in the following sections.

4.4.1. Test item identification

Basic information on the identity of the test item such as the substance or product name is a core criterion in terms of relevance for deciding on inclusion/exclusion of the particular study. However, more detailed information on the test item is required to evaluate whether the results of an assay can reliably be attributed to the actual substance or mixture under evaluation (see Table 2).

The composition of commercially available products is often modified over time as a result of adaptations to technical progress or market needs. Also, products of the same name sold in different countries may have different compositions. Without sufficient information, the toxicological properties found in a respective study might be wrongly attributed to the AS in such cases. Sometimes, information in the title or abstract of publications gives the impression that the pure AS was investigated where in fact the test item was a complex commercially available PPP or BP (e.g. Daruich et al., 2001; George et al., 2010; Karabay and Oguz, 2005). These are no problems affecting the reliability of a study, but they certainly affect relevance for a certain regulatory question. Thus, the assessment of relevance may occasionally need to be revised during evaluation of reliability.

4.4.2. Test species and in vitro models

Transparent and comprehensive description of the test organism of an *in vivo* study or the model used in an *in vitro* study is in general essential information, not only for any subsequent regulatory use. Ågerstrand et al. (2011) examined nine studies on aquatic toxicity with regard to the description of the test organism and concluded that none of the publications covered all aspects under consideration. A common drawback is that certain information might not have been considered important by the author at the time of publication, but would be needed later by a regulatory authority to make the best possible use of the data, particularly when there is conflicting information from other sources.

For *in vivo* studies, the identification of the species, strain and sex of animals is required for rating systems as well as in the OECD TGs for the different study types. According to OECD TGs, also age, weight at start of the test as well as housing conditions and usually also source of the animals have to be covered. The questionnaire of ToxRTool includes similar criteria. SciRAP criteria in addition include very detailed questions on housing and feeding conditions that can provide background information on exposure to possible endocrine active substances.

It is well known that the genetic background can influence test results. Beside defined genetic modifications, it has also been shown in both mice and rats that intra-strain differences exist for inbred and outbred strains (e.g. Bryant et al., 2008; Langer et al., 2011).

For *in vitro* studies, the situation is even more complex, as the multitude of models shows: for example, whole organs or organ slices, tissue samples or explants, cell cultures, bacteria or cell free assays. The criteria of ToxRTool include questions on the description of the test system, on the source/origin, on test system properties and on conditions of cultivation and maintenance. (Schneider et al., 2009). Also in OECD TG for *in vitro* tests and other *in vitro* guidelines, many criteria have been listed that can be used as guidance. Table 3 summarises information required on test species and *in vitro* models.

4.4.3. Study design

Sufficient information on study design is requested by all existing rating systems and TGs. Important features for description of *in vivo* study design are given in Table 4.

In a comparison carried out by Segal et al. (2015) using ToxRTool as rating system, the criterion “Are sufficient details of the administration scheme given to judge the study” was one of those with the most inconsistent responses among different evaluators. This underlines how important (and difficult) it is to describe the exposure conditions sufficiently. Information on study design is essential for the assessor to judge the appropriateness and limitations of the specific test.

For *in vitro* systems it is not feasible to present a comprehensive list of criteria meeting all demands and only few evaluation systems address *in vitro* studies. In principle, the basic considerations for *in vivo* studies are applicable also to *in vitro* studies, some specific items to be considered as applicable are included in Table 5.

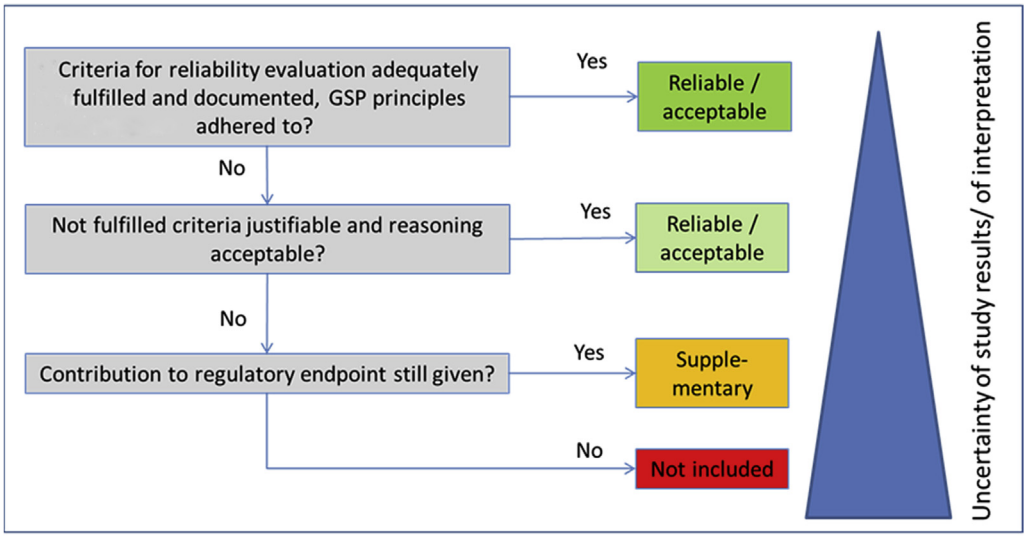

Fig. 3. Decision tree for the acceptability of non-guideline studies (expert judgement).

Table 2
Test item identification.

No.	Criterion	Description and remarks
1	Is the test item unequivocally identified?	- CAS No., IUPAC name in addition to synonyms/trivial names - Isomers/racemic mixtures, if applicable - For formulations: Content of relevant ingredients, formulation type - Source of the product: manufacturer, lot/batch number.
2	Is the purity of the active substance or its content in the product stated?	- Minimum content (purity) of the substance - Numerical descriptors preferred (% w/w, g/kg). - If applicable: isomeric ratios, presence of enantiomers, diastereomers
3	Is information on impurities of the test substance given?	- Information on potentially toxicologically relevant impurities (e.g. from the manufacturer via lot/batch number) that may influence the test result - Impurities present in the test substance/formulation should cover the impurity profile of the substance under assessment
4	Are relevant degradation products specified?	- Substances susceptible to deterioration may produce degradation products which affect the toxicological profile.

Table 3
Test species and *in vitro* models.

Test organism		
No.	Criterion	Description and Remarks
1	Are the main features of the animal model described?	- Species, strain, sex - Supplier, breeding (inbred, outbred) - Age and body weight at the beginning of the experiments
2	Is additional information about the strain given, which might be necessary?	- Genetic background; genetic modifications, if present; - Proof of absence of pathogens, e.g. specific pathogen free status etc. - Sub-strain, if applicable - Historical control data on background incidences
3	Are housing conditions described in detail?	- Information on feed, cages, bedding material, temperature, humidity, etc.
In vitro models		
No.	Criterion	Description and remarks
1	Is the *in vitro* model described sufficiently?	- Basic information for cell cultures: origin, provider, conditions under which the cells were kept and replicated. - For assays using explants, organ slices or dermatomed skin, detailed information on the type, source and on the animal or human donors.
2	Are specific features described?	- If genetic variations or modifications exist, these should be stated.
3	Is necessary information on maintenance and culture of the test system given?	- Information on culture conditions and material; temperature, CO_2, humidity, media, supplements, etc. including information on batches or lots since these may vary and affect results.

4.4.4. Documentation of test results

To be considered acceptable for regulatory purposes, test results need to be presented in a sufficiently accurate, detailed and comprehensible manner, irrespective of the study type.

In principle such an assessment is performed prior to publication by the respective journal for published scientific literature. However, there are currently no common standards that would guarantee the same level of assurance across journals and publications. For studies on endpoints covered by OECD (or other) TGs, detailed requirements for reporting of results can be found in the respective guideline documents. These recommendations as well as requirements common to the various TGs may be used to guide the reliability evaluation of studies for which such TGs do not exist. A compilation of important questions is given in Table 6.

4.4.5. Evaluation including statistical methods

The statistical processing of data allows formalising the results and weighing their relevance in consideration of the study design. However, inappropriate statistical analysis of research data has been heavily criticised in the past. This was mainly attributed to formal application of statistical tests and generic significance levels (Ioannidis, 2005). An analysis of 30 recent publications from toxicological journals showed that few reports provided justifications for the statistical method selected and many failed to describe tests for normality and equal distribution when applying ANOVA or Student's T-test (Na et al., 2014). It was noted that the positive predictive value, i.e. the probability that a positive finding is a true positive, also depends on the pre-study odds (i.e. the quality of the hypothesis tested) and the power of a study (Ioannidis, 2005). Similarly, repeated and/or parallel multiple testing of a large range of endpoints increases the likelihood of false positives. Publication bias, i.e. that test results showing an effect are more likely to be reported, may then lead to an underestimation of the impact of repeated and parallel testing. Conversely, positive relationships may be missed as the result of high variability and/or unfavourable signal-to-noise ratios or application of inappropriate statistical methods (Ioannidis, 2005) and low group sizes in *in vivo* testing can be one reason for low statistical power and a lack of reproducibility. Accordingly, TGs and accompanying guidance frequently advise one to ensure that i) the statistical test chosen is appropriate for the data to be evaluated and ii) interpretation of the statistics takes into account the degree of certainty of the test result and the biological context (or at least relevance) of the finding. Extensive guidance has been developed on the topic and can be found for example in the OECD Guidance document no. 116 on the conduct and design of chronic toxicity and carcinogenicity studies (OECD, 2012). Although this document specifically relates to the chronic and carcinogenicity tests in animals, key principles apply to the evaluation of study results in general, including those published in the open literature.

In descriptive statistics, calculation of means remains the most popular approach in toxicology to describe a tendency, accompanied by standard deviation or standard error of mean (Na et al., 2014). However, depending on the distribution of the data, the

Table 4
Study design: ***In vivo* studies**.

Study setup		
1	Is information on the dose groups given?	- Identity of dose groups and controls - Number of animals per dose group (affects statistical power).
2	Are frequency and duration of exposure adequate?	- Information on frequency of applications and total duration of exposure - Adequacy of frequency and duration for the aim of the study
3	Is information on time-points of observations sufficient?	- Adequacy of time points for the individual toxicological endpoint under investigation - Documentation of timing of observations (if necessary, relevant persistence of effects)
4	Were appropriate controls included?	- Depending on the endpoints, negative or positive controls or both. - Number of control animals, important for statistical power - Comparable housing and handling conditions to treated groups.
5	Was the aim of the study stated and were selected parameters suitable?	- Full description of parameters evaluated - Adequacy of selected parameters to address the endpoints
Administration		
6	Is the administration route stated and appropriate?	- Information on route and details on administration. - Representativeness and appropriateness (e.g. in light of phys.-chem. properties) to reach target organs in sufficient quantities for a given route.
7	Are the doses administered or concentrations in application media given?	- Clear indication, whether given test concentrations refer to the whole formulation or to the active compound. - Information on the test item concentrations, if possible analytically verified
8	Is the vehicle described and adequate?	- Description of vehicle used to prepare dosing solution to conclude on potential vehicle effects; e.g. on absorption rate.
9	Were potential confounding factors addressed and/or obviated?	- Analysis of potential confounding factors, which may influence the outcome of a toxicity study (e.g. phytoestrogen content of the diet; solvents that induce liver toxicity, feed contaminants, etc.).
Methods		
10	Are the methods for the determination of all parameters clearly described and adequate?	- Conclusive and detailed description of methods used in the research laboratories (validated or confirmed in inter- or intra-lab- comparisons, e.g. ring trials, etc.). - Information whether the method is widely accepted or whether modifications have been carried out in the reporting lab.
11	Are the analytical methods clearly described and adequate?	- Statement of specificity and sensitivity of the analytical methods.

Table 5
Study design: ***In vitro* studies**.

Study setup		
1	Is information on the replicates given?	- Number of replicates affects statistical power.
2	Is the number of treated cells provided?	- May affect the sensitivity of a test; in particular those aimed at detecting rare events, e.g. in genotoxicity testing
3	Are culture conditions adequate and described sufficiently	- For example temperature, CO_2, humidity - May influence the study result, in particular variability/sensitivity.
4	Were adequate controls included?	- Depending on the endpoints, negative or positive controls or both. - Number of controls has to be sufficient for statistical power. - Culture and handling to mimic that of treatment groups.
5	Are parameters measured adequate for investigation of the endpoints?	- Predictivity of changes in measured parameter(s) for the toxicological endpoint. - Note: *In vitro* methods may address different parameters than *in vivo* studies.
Administration		
6	Is the method of application/details on the experimental set-up given?	- Comprehensible and clear description of the method of application and set-up. - Sufficient information about frequency of applications and total duration of exposure - Appropriateness of type, frequency and duration of exposure for the aim of the study
7	Are the applied doses or concentrations in test medium given?	- Clear indication, whether given test concentrations refer to the whole formulation or to the active compound. - Information on the test concentrations, if possible analytically verified.
8	Is the vehicle described and adequate?	- Description of the vehicle used to prepare dosing solution to conclude on potential vehicle effects on the test system itself or on the potency of the test item.
9	Were potential confounding factors addressed and/or obviated?	- Potential confounding factors, which may influence the outcome (e.g. compounds in cell culture material; FCS batches; solvents that induce liver toxicity like ethanol, etc.).
Methods		
10	Are the methods for the determination of all parameters clearly described and adequate?	- Sufficient detail on the methods required to assess reliability of the study results. - Validation of methods, external/internal quality controls. - Comparability to reference methods. - Impact of method modifications.
11	Are the analytical methods described and adequate?	- Specificity and sensitivity of any analytical methods, e.g. for determination of test item concentrations, if applicable.

median and quantiles may be more robust. If data are not normally distributed, are skewed or contain outliers, mean and standard deviation will not properly represent the sample and transformation of the data may be appropriate. Assay data is routinely evaluated by null hypothesis testing. Any such test is based on assumptions, e.g. on the independence of samples and random sampling, or the modality and symmetry of distribution of the population data. If these assumptions are not met, the test may produce unreliable results, which should be considered when judging the statistical test strategy. Recommendations for appropriate statistical tests are provided in Appendix V of the OECD Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies (OECD, 2002) for example and in flowcharts in chapter 4 and Appendix 1 of OECD Guidance document no. 116 (OECD, 2012).

Table 6
Documentation of test results.

No.	Criterion	Description and Remarks
1	Are the results provided transparently?	- Legends for tables and figures that allow for unambiguous identification of data including allocation of measured values to test item concentration and identity (substance or formulation). - Criteria for any scoring system (e.g. for clinical observations). - Information on potential compromising effects on the test item (e.g. precipitation, degradation or phase separation in the assay matrix) on the test organism/*in vitro* model (e.g. information on cytotoxicity that may compromise detection of genotoxicity).
2	Are results for concurrent controls reported at the same level of detail as for treatment group?	- Full reporting of the results of any negative and positive control or performance standard/reference compound. - Information on controls in sufficient detail for interpretation of results, as well as for statistical analysis, i.e. they should include not only ranges but also mean values, information on distribution of values, standard deviation and the size (n) of the control sample.
3	Are relevant historical controls or reference values/ranges available or referenced?	- Historical controls data obtained under comparable conditions to the study under evaluation (OECD, 2012). - to decide whether concurrent controls are representative (reliable) and deviations in treatment groups from controls are biologically significant - For certain tests, generally accepted decision criteria – usually based on results from validation studies – may replace use of historical control values.
4	Are results provided for all measured endpoints?	- Measured results for all endpoints measured, even if there was no statistically significant effect compared to controls.
5a	Sufficient level of detail– Are numerical values provided?	- Appropriation of all results numerically either as dichotomous, continuous or categorical values, histopathological endpoints and clinical observations as scores. - Graphical representation of data only, i.e. without numerical values, can compromise interpretation and thus reliability.
5b	Sufficient level of detail– Is aggregation of individual data appropriate?	- Aggregation of numerical values for replicates into means and standard deviation (or confidence intervals, ranges, etc.) for normally distributed data if statistical analysis meets all current standards. - Aggregation of data over various endpoints or fractions reduces the level of detail and should be avoided.
6	High level of detail– Are individual data of replicates provided?	- Individual values for biological replicates, in particular for key endpoints supporting the study conclusion. - Individual values assigned to the respective replicate/animal number. - Additional (statistical) analysis if needed. - Potential correlations between various endpoints at the level of the individual. - Provision of individual values.
7	Are raw data included or accessible, e.g. as supplementary material or through a database?	- Access to primary data (i.e. original output files/tables providing OD readings rather than enzyme activity in IU/mL) in order to increase the transparency of the presented information for regulatory risk assessment. - Allows one to perform recalculation of results if needed, including standard curves.

Examination of dose-response relationships may improve sensitivity and confidence in findings. For carcinogenicity studies, trend testing (e.g. using the Cochran-Armitage linear trend test) has become a routine approach (Gart et al., 1986; OECD, 2012). However, it should be noted that the "shape" of the dose-response relationship can vary considerably and prior knowledge about the appropriate regression model is thus required. Inappropriate assumptions may result in a failure to identify a treatment related effect. Guidance on the selection of appropriate models has been developed in the context of benchmark dose modelling and may be applied more generally (U.S.EPA, 2012).

Data evaluation should also take into account known confounders. While reduced survival in cancer bioassays may reduce the number of tumour bearing animals, cytotoxicity may either mask or induce other effects such as mutagenicity. To some extent, the confounding effect may be eliminated by choosing additional statistical tests (e.g., Peto test) or including alternative endpoints in the evaluation, such as time-to-tumour. Normalisation of data can be a viable option, for example, when there are effects on body weight.

Finally, it should be noted that the methodology of data analysis in toxicology is also subject to some, albeit slow modifications over time, reflecting scientific progress in the field (Kobayashi, 2001). Accordingly, results should be reported in sufficient detail to allow for re-analysis if indicated (see also section 4.4.4). Also, the fact that a lack of statistical significance does not prove an absence of an effect or activity is often neglected. To clarify the level of confidence of negative test outcomes, an assessment of the statistical power of an assay is necessary. However, the statistical power is rarely calculated or cited in the study report, even if available from method validation.

Criteria provided in Table 7 can assist when assessing the appropriateness of the statistical method(s) and documentation.

5. Uncertainty

A consideration of the sources of uncertainty that clearly and transparently states what sources of uncertainty were identified and how these affect the final conclusions should be included in an assessment. According to Fenner-Crisp and Dellarco (2016), uncertainty is "a scientific reality that cannot be totally eliminated, and, therefore, must be acknowledged and explained, together with its impact on the risk conclusions and estimates". The WHO/IPCS more specifically defined uncertainty in relation to hazard characterisation as "lack of knowledge regarding the "true" value of a quantity, lack of knowledge regarding which of several alternative model representations best describes a system of interest, or lack of knowledge regarding which probability distribution function and its specification should represent a quantity of interest" (WHO/IPCS, 2014). As outlined by ECHA in REACH Guidance R.19, uncertainty can also relate to the presence or absence of intrinsic hazard properties of a substance (ECHA, 2012). In agreement with this, EFSA described uncertainty as "a general term referring to all types of limitations in available knowledge that affect the range and probability of possible answers to an assessment question" (EFSA, 2016).

Uncertainty in risk assessments can result from uncertainty in the measurement of (no) effect, uncertainty in the exposure

Table 7
Evaluation including statistical methods.

No.	Criterion	Description and remarks
1	Are the descriptive statistics appropriate?	- Mean and standard deviation or standard error of mean. - Depending on the distribution of the data (skewedness, modality, outliers), other descriptors such as median/quantiles/ranges. - Normalisation (e.g. log transformation) of data.
2	Are the statistical methods stated?	- If lacking, conclusions cannot be substantiated and statistical re-assessment should be performed using individual replicate values and appropriate methods.
3	Are the most sensitive statistical methods considered?	- Parametric tests are mostly considered more sensitive compared to non-parametric, but may not be applicable. - Trend-testing/regression across many dose groups may be more sensitive compared to group-wise comparison.
4	Are the statistical tests chosen appropriate?	- Statistical tests make assumptions on sample and population data. If these are violated (appropriate initial testing e.g. for normality and equal variance in case of ANOVA), the test is not appropriate and another test relying on fewer/other assumptions should be used.
5	Have regression models been applied, can they be considered appropriate?	- When regression methods are applied for parameter calculation, e.g. benchmark doses or flux values, there should be either biological or empirical evidence for the appropriateness of the underlying assumption in the specific case.
6	Have confounding effects been taken into consideration?	- Confounding effects can be taken into consideration in the statistical analysis itself and/or interpretation of the statistical evaluation.

estimate, whether predicted or based on measurements, and uncertainty in the risk estimate e.g. resulting from extrapolation between or within species (ECHA, 2012). These kinds of influences, lack of knowledge or lack of data as well as methodological limitations of the underlying study can provide sources of uncertainty. Uncertainty can also result from the transformation of continuous measurement data into dichotomous "yes/no" outputs, for example, as discussed for the LLNA in skin sensitisation testing (Kolle et al., 2013; Leontaridou et al., 2017 and others) largely ignoring the grey-zone in between. This has been addressed in recent revisions of genotoxicity test guidelines with the introduction of the categories clearly positive, positive, equivocal, etc. (OECD, 2016).

When hazard data such as information from a second species or a valid long-term study is lacking, the increased uncertainty is frequently compensated for by application of additional assessment (or uncertainty) factors (ECHA, 2015). However, uncertainty of a value derived from data is usually not explicitly taken into account when following the NOAEC/LOAEC approach. The utility of the benchmark dose approach, providing a lower confidence limit (BMDL) for a possible toxicological effect has been discussed frequently, but a final and harmonised methodology has not yet been broadly implemented.

It should be considered how the inclusion of studies with limited relevance and/or reliability may affect the uncertainty of the whole assessment. For example, when the tested formulation differs from the product under evaluation, the applicability of study results has to be examined according to the specific criteria given in EU Guidance SANCO/12638/2011 (EC, 2012), thus limiting the introduction of uncertainty resulting from the limited relevance of the data. This shows how a thorough evaluation of relevance and reliability is required to identify and control for sources of uncertainty as much as possible.

6. Discussion and conclusion

We conclude that the contribution of non-guideline studies from peer-reviewed scientific literature to regulatory risk assessments could be substantially increased. Implementation of good reporting criteria and using transparent principles for study conduct; for example GSP, would augment the applicability of scientific publications. Within assessment reports or databases used for regulatory decision making, transparent evaluation of relevance and reliability of all available information is of major importance for reproducibility and public confidence and has to be well documented. Even though several systems for evaluation have been published, the need for harmonised criteria is still urgent.

The comparison of requirements included in OECD TGs and in the two evaluation tools SciRAP (Molander et al., 2014) und ToxRTool (Schneider et al., 2009) shows that for non-guideline studies criteria catalogues are also applicable. These tools need to be employed regularly to become more widely accepted by risk assessors. Another recently published overview on important features of a high-quality assessment is the "*Guide for Judging the Quality of an Assessment*" by Fenner-Crisp and Dellarco (2016) in which the importance of applying predefined criteria and a weight-of-evidence approach to address causal relationships in a systematic manner is advocated.

Original guideline-compliant studies from industry are usually not published or otherwise accessible and thus are not subject to critical review by the scientific community or the public (Lutter et al., 2013). It would be more transparent to disclose the original, typically confidential study reports together with the raw data. After extensive public discussions concerning the assessment of the active substance glyphosate, EFSA recently started to release the raw data used in the assessment, following a *public access to document* request[1]. This will enable the professional public in particular to independently asses the conclusions of the regulatory authorities and the details of the assessment process.

In general, regulatory authorities fulfil their responsibility of transparency by making reports available, for example on their websites. These include compilations of the assessed data. Additionally, international authorities like the European EFSA and ECHA provide an immense amount of data on assessed substances (e.g. registered substances database[2], EU Pesticides Database[3]).

Despite this, regulatory reports are cited only very rarely in published literature. In many research facilities, little awareness seems to exist for the fact that scientific assessments by regulatory authorities are indeed applied science based on reliable data. McDonagh et al. (2013) addressed the importance of knowledge from regulatory documents for systematic reviews.

The criteria presented in this paper give insight into the considerations concerning relevance and reliability of openly published studies from a regulatory perspective and introduce improvements for better use of the available evidence especially from published literature for risk assessment purposes. Scientific research data fulfilling the described criteria can and indeed should

[1] (http://www.efsa.europa.eu/en/press/news/160929a).
[2] http://echa.europa.eu/information-on-chemicals/registered-substances.
[3] http://ec.europa.eu/food/plant/pesticides/eu-pesticides-database.

be used in risk assessments. As such, said data might also have a much greater impact on regulatory decision making. The application of the methods and criteria described here could increase the database used in substance approval and product authorisation. In addition, the presented criteria aim at a better understanding of assessments by interested parties or the public.

Acknowledgements

The authors would like to thank Kristin Herrmann, Thomas Kuhl, Denise Kurth, Glenn Lurman, Britta Michalski, David Schumacher and Birgit Wobst for the detailed discussions and helpful contributions.

Transparency document

Transparency document related to this article can be found online at http://dx.doi.org/10.1016/j.yrtph.2017.06.010.

References

Ågerstrand, M., Beronius, A., 2016. Weight of evidence evaluation and systematic review in EU chemical risk assessment: foundation is laid but guidance is needed. Environ. Int. 92-93, 590–596.

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, 1–15.

BBSRC, 2013. BBSRC Statement on Safeguarding Good Scientific Practice. (UK) as assessed on October 08 2015. http://www.bbsrc.ac.uk/documents/good-scientific-practice-pdf/.

Beronius, A., Molander, L., Ruden, 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.

Bryant, C.D., Zhang, N.N., Sokoloff, G., Fanselow, M.S., Ennes, H.S., Palmer, A.A., McRoberts, J.A., 2008. Behavioral differences among C57BL/6 substrains: implications for transgenic and knockout studies. J. Neurogenet. 22, 315–331.

Budiansky, S., 1983. Data falsification trial: drug testing lab was “shambles”. Nature 302, 738.

Buonsante, V.A., Muilerman, H., Santos, T., Robinson, C., Tweedale, A.C., 2014. Risk assessment's insensitive toxicity testing may cause it to fail. Environ. Res. 135, 139–147.

Daruich, J., Zirulnik, F., Gimenez, M.S., 2001. Effect of the herbicide glyphosate on enzymatic activity in pregnant rats and their fetuses. Environ. Res. 85, 226–231.

DFG (Deutsche Forschungsgemeinschaft), 2013. Proposals for Safeguarding Good Scientific Practice. Recommendations of the Commission on Professional Self Regulation in Science. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, ergänzte Auflage.

EC, 2015. Working Group of the Advisory Group on the Food Chain, Animal and Plant Health; Ad hoc Dialogue event on risk assessment of active substances in Plant Protection Products; 24 April 2015. Ref. Ares(2015)2071689–18/05/2015. http://ec.europa.eu/dgs/health_food-safety/dgs_consultations/docs/dgs-consultations_working-groups_20150424_summary_en.pdf.

ECHA, 2010. Practical Guide 2: How to Report Weight of Evidence.

ECHA, 2011. Guidance on Information Requirements and Chemical Safety Assessment; Chapter R.4: Evaluation of Available Information.

ECHA, 2012. Guidance on Information Requirements and Chemical Safety Assessment; Chapter R.19: Uncertainty Analysis.

ECHA, 2015. Guidance on the BPR: Volume III Human Health, Part B Assessment.

EFSA, 2011. Submission of scientific peer-reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009. EFSA J. 9 (2), 2092, 2011.

EFSA, 2016. Guidance on Uncertainty in EFSA Scientific Assessments - Revised Draft for Internal Testing.

ESF, 2011. The European Code of Conduct for Research Integrity. European Science Foundation/ALLEA ALL European Academies, Strasbourg.

EC, 2009. Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 Concerning the Placing of Plant Protection Products on the Market and Repealing Council Directives 79/117/EEC and 91/414/EEC; OJ l 309, 24.11.2009, pp. 1–50.

EC, 2012. Guidance Document on Significant and Non-Significant Changes of the Chemical Composition of Authorised Plant Protection Products Under Regulation (EC) No 1107/2009 of the EU Parliament and Council on Placing of Plant Protection Products on the Market and Repealing Council Directives 79/117/EEC and 91/414/EEC; SANCO/12638/2011, 20 November 2012 rev. 2.

EU, 2012. Regulation (EU) No 528/2012 of the European Parliament and of the Council of 22 May 2012 Concerning the Making Available on the Market and use of Biocidal Products Text with EEA Relevance; OJ L 167, 27.6.2012, pp. 1–123.

EU, 2013a. Commission Regulation (EU) No 283/2013 of 1 March 2013 Setting Out the Data Requirements for Active Substances, in Accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council Concerning the Placing of Plant Protection Products on the market Text with EEA Relevance; OJ L 93, 3.4.2013, pp. 1–84.

EU, 2013b. Commission Regulation (EU) No 284/2013 of 1 March 2013 Setting Out The Data Requirements for Plant Protection Products, in Accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council Concerning the Placing of Plant Protection Products on the Market Text with EEA Relevance; OJ L 93, 3.4.2013, pp. 85–152.

Fenner-Crisp, P.A., Dellarco, V.L., 2016. Key elements for judging the quality of a risk assessment. Environ. Health Perspect. 124, 1127–1135.

Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E., Wahrendorf, J., 1986. Statistical Methods in Cancer Research. Volume III–The design and analysis of long-term animal experiments. IARC Sci Publ, pp. 1–219.

George, J., Prasad, S., Mahmood, Z., Shukla, Y., 2010. Studies on glyphosate-induced carcinogenicity in mouse skin: a proteomic approach. J. Proteomics 73, 951–964.

Hooijmans, C.R., Leenaars, M., Ritskes-Hoitinga, M., 2010. A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. Altern. Lab. Anim. 38, 167–182.

Ioannidis, J.P., 2005. Why most published research findings are false. PLoS Med. 2, e124.

Karabay, N.U., Oguz, M.G., 2005. Cytogenetic and genotoxic effects of the insecticides, imidacloprid and methamidophos. Genet. Mol. Res. 4, 653–662.

Kilkenny, C., Browne, W.J., Cuthill, I.C., Emerson, M., Altman, D.G., 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. 8, e1000412.

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.

Kobayashi, K., 2001. Methods of statistical analysis of quantitative data obtained by toxicological bioassays using rodents in Japan: historical transition of the decision tree. J. Environ. Biol. 22, 1–9.

Kolle, S.N., Basketter, D.A., Casati, S., Stokes, W.S., Strickland, J., van Ravenzwaay, B., et al., 2013. Performance standards and alternative assays: practical insights from skin sensitization. Regul. Toxicol. Pharmacol. 65, 278–285.

Langer, M., Brandt, C., Loscher, W., 2011. Marked strain and substrain differences in induction of status epilepticus and subsequent development of neurodegeneration, epilepsy, and behavioral alterations in rats. [corrected]. Epilepsy Res. 96, 207–224.

Leontaridou, M., Urbisch, D., Kolle, S.N., Ott, K., Mulliner, D.S., Gabbert, S., Landsiedel, R., February 23 2017. Quantification of the borderline range and implications for evaluating non-animal testing methods' precision. Altex. https://doi.org/10.14573/altex.1606271.

Lutter, R., Barrow, C., Borgert, C.J., Conrad Jr., J.W., Edwards, D., Felsot, A., 2013. Data disclosure for chemical evaluations. Environ. Health Perspect. 121, 145–148.

McDonagh, M.S., Peterson, K., Balshem, H., Helfand, M., 2013. US Food and Drug Administration documents can provide unpublished evidence relevant to systematic reviews. J. Clin. Epidemiol. 66, 1071–1081.

Molander, L., Ågerstrand, M., Beronius, A., Hanberg, A., Rudén, C., 2014. Science in risk assessment and policy (SciRAP): an online resource for evaluating and reporting in vivo (Eco)Toxicity studies. Hum. Ecol. Risk Assess. Int. J. 21, 753–762.

MRC, 2012. MRC Ethics Series. Good research practice: Principles and guidelines. Medical Research Council, UK, pp. 1–27.

Myers, J.P., vom Saal, F.S., Akingbemi, B.T., Arizono, K., Belcher, S., Colborn, T., et al., 2009. Why public health agencies cannot depend on good laboratory practices as a criterion for selecting data: the case of bisphenol A. Environ. Health Perspect. 117, 309–315.

Na, J., Yang, H., Bae, S., Lim, K.M., 2014. Analysis of statistical methods currently used in toxicology journals. Toxicol. Res. 30, 185–192.

NHMRC, 2007. Australian Code for the Responsible Conduct of Research. Jointly issued by National Health and Medical Research Council, the Australian Research Council and Universities Australia downloaded from. https://www.nhmrc.gov.au/guidelines-publications/r39. October 08 2015.

NSF, 2009. Responsible Conduct of Research. National Science Foundation (USA); Federal Register, pp. 42126–42128. http://www.gpo.gov/fdsys/pkg/FR-2009-08-20/html/E9-19930.htm. as assessed on October 08 2015.

OECD, 1997. Decision C (97)186/Final of the OECD Council. Council Decision Amending Annex II to the Council Decision Concerning the Mutual Acceptance of Data in the Assessment of Chemicals [C(81)30(FINAL)].

OECD, 2002. Guidance Notes for Analysis and Evaluation of Repeat-dose Toxicity Studies, ENV/JM/MONO(2000)18.

OECD, 2012. Guidance Document 116 on the Conduct and Design of Chronic Toxicity and Carcinogenicity Studies, Supporting Test Guidelines 451, 452 and 453. Series on Testing and Assessment.

OECD, 2016. Test No. 476. In: Vitro Mammalian Cell Gene Mutation Tests Using the Hprt and Xprt Genes. OECD Publishing.

PRCR, 2011. The TRI-agency Framework: Responsible Conduct of Research, pp. 1–17. Panel on Responsible Conduct of Research (Canada). http://www.rcr.ethics.gc.ca/eng/policy-politique/framework-cadre/.

Roth, N., Ciffroy, P., 2016. A critical review of frameworks used for evaluating reliability and relevance of (eco)toxicity data: perspectives for an integrated eco-

human decision-making framework. Environ. Int. 95, 16–29.
Schneider, K., Schwarz, M., Burkholder, I., Kopp-Schneider, A., Edler, L., Kinsner-Ovaskainen, A., et al., 2009. "ToxRTool", a new tool to assess the reliability of toxicological data. Toxicol. Lett. 189, 138–144.
Segal, D., Makris, S.L., Kraft, A.D., Bale, A.S., Fox, J., Gilbert, M., et al., 2015. Evaluation of the ToxRTool's ability to rate the reliability of toxicological data for human health hazard assessments. Regul. Toxicol. Pharmacol. 72, 94–101.
U.S.EPA, 2012. Benchmark Dose Technical Guidance, EPA/100/R-12/001. Washington DC.
WHO/IPCS, 2014. Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterization.