Study Rating – is Reliability enough? An appeal for revisiting Relevance and Adequacy, particularly in environmental risk assessment (ERA)

R. Arno Wess
IES Ltd, Benkenstrasse 260, 4108 Witterswil, Switzerland

EUROTOX 2018,
Brussels, Belgium, September 4th

Introduction

The challenge of decision making between several and unequivocal experimental data occurred in regulatory with the systematic hazard and risk evaluation of existing chemicals, namely the OECD High Production Volume program and eventually the EU REACh legislation. The working experience resulted in the recommendation of a systematic approach (Klimisch et al. 1997), considering not only Reliability but also Relevance and Adequacy of data. The regulatory implementation in the REACh legislation and the IUCLID database tool for data submission were anyhow restricted to the assignment of a Reliability score, thus rating seemingly only the intrinsic quality of the available study information. Nonetheless the integrated IUCLID implementation and also rating hints given in the Klimisch et al. (1997) publication do not blind out adequacy and relevance, but these aspects appear within the Reliability criteria (e.g. “Unsuitable test system” as a reason for unreliability). Due to endpoint specific refinement of the criteria for Reliability evaluation, also the publication of Küster et al. (2009) became a standard in regulatory, specifically for the environmental risk assessment (ERA) of pharmaceuticals. Finally the Reliability evaluation methodology has been further developed to gain increased reproducibility of the ratings (Kase et al. 2016).
Although criteria to be fulfilled for assignment to a certain category are stated in the publications, the final evaluation is done by expert judgment, leaving considerable space for interpretation. Thus, evaluation needs to be clearly justified and must be understandable to others, especially when data is assigned to a category of low quality. In result, a state-of-the-art conclusion on the Reliability of a study is a work and documentation intensive task, but as the Reliability of an irrelevant and/or inadequate study is completely meaningless, these points may be worth to be checked and documented before undertaking investigations on the Reliability.
Studies correctly ratable as “reliable without restriction”, can still be unsuitable for assessment. This poster revisits thus particularly the topics Relevance and Adequacy. It is based on the experiences made and the studies checked for regulatory use during many years in consultancy. The proposed definitions represent a synthesis of the above cited publications and own experiences. Based on this a more structured rating procedure is encouraged.

Reliability, which refers to and scores the intrinsic value

A key asset of the data quality evaluation is the **inherent quality** of the accessible documentation, e.g. a test report or publication, relating to preferably standardized methodology, the formal validity of the test method, applied quality assurance (Good Laboratory Practice) and completeness of the information. The evaluation of the study data is conducted by expert judgment, and the quality is described by assigning a study to one of four codes of reliability: 1 (reliable without restrictions); 2 (reliable with restrictions), 3 (not reliable) and 4 (not assignable). While 1 to 3 describe the scientific inherent quality or value of the study data, 4 (not assignable) covers studies whose quality and reliability cannot be entirely evaluated due to insufficient documentation.
Klimisch et al (1997) suggest an additional code 5 for identifying not evaluated studies, which may be the case because they are obviously without relevance for hazard/risk assessment but may e.g. pop up in literature searches (e.g. because of a coding error or ambiguity). This rating should be explained in the Relevance statement, because it does not score the intrinsic quality.
In case additional information becomes available (e.g. laboratory raw data), the reliability score may be corrected and the rating statement has to refer also to the source of this data.

Relevance for hazard identification and risk characterization

It covers the extent to which the information gained through a study is appropriate for a particular **hazard identification** or risk characterization. Such hazard or risk aspects depend on the type of evaluation and are e.g. represented by a regulatory demand such as “Toxicity to birds”. It may turn out that a study result is not reliable but contains relevant information, e.g. if the exposure is not clear but strong effects in comparison with the positive control are reported. Relevance is thus (regulatory-) context-dependent and is valid only for a specific type of evaluation (e.g. REACH registration in the <1000 t/a band), which must be clearly stated. In an ERA for human pharmaceuticals, such study would be irrelevant because the endpoint is not to be assessed (however the catastrophic effects of diclofenac to the vulture population in India has promoted the implementation of the ERA requirement for human medicinal products in the EU). Another example is the sensitivity assessment, where experimental data are overruled by the simple Nickel content based legal criterion, which makes thus any study abundant.

Adequacy to conclude an assessment (of intrinsic toxicity or environmental fate)

This aspect is about the usefulness of data for a risk **assessment level or purpose**. The information given may be reliable and relevant but not sufficient for assessment, e.g. may primary biodegradation be reliably evidenced in a good documented report and this fact is relevant for the identification of persistence related risks but not sufficient for assessment as ultimate degradation or the identity of the transformation products are required at the assessment level. In a Weight of Evidence approach, several individually inadequate studies may -taken together- reach adequacy.
Adequacy is result-dependent, e.g. strong effects in a preliminary developmental toxicity screen (OPPTS 870.3500) may suffice for GHS/CLP classification whereas absence of effects would make the study inadequate.

Table 1. Combinations of rating results for Relevance, Adequacy and Reliability and conclusions for use in assessment and regulation

Reliability	Relevance	Adequacy	Example(s)	Possible Improvement
1 or 2	Yes	Yes	GLP study, suitable for the fully characterized test item, according to a study protocol as recommended by regulatory guidance	Not required
	Yes	No	- Assessment of ready biodegradability (but not for the overall environmental fate); Use of the wrong test item, e.g. a dissociating pharmaceutical formulated as an organic salt with a readily biodegradable and toxicologically almost irrelevant organic cation or anion, used in the biodegradation screening-test; Deficit: It will be present as two substances, one of them feeding the test microorganisms. - Persistency assessment; A radiolabel in a carboxylic acid group, being subject to readily decarboxylation	- New study with the test salt with an inorganic counterion - New study with ring labelling
	No	Yes	Not possible, irrelevant studies can never be adequate.	
	No	No	Test artifacts, i.e. if the experimental protocol is unsuitable for a substance, deliver typical cases: - Aquatic toxicity of test items releasing metal cations in case the test medium contains chelating agents, which lack in the environment and influence the bioavailability. - The publication of Hill (1965), if adapted for (eco)-toxicity as e.g. available in a first layout from Wess (2016), may provide some help for the identification of hints for test artifacts.	None
3	All combinations		GLP study, suitable for the incompletely characterized test item, according to study protocol as recommended by regulatory guidance, Deficit: Stereochemical composition and origin unknown	- New study with fully characterized test item
4	All combinations		GLP study, suitable for the ambiguously characterized test item, according to study protocol as recommended by regulatory guidance, Deficit: Only trivial name, which is in use for two different chemicals, stated as test substance, but supplier & Lot number known	Clarification by statement from the supplier

Table 2. Improvement options and Dependencies of the evaluation topics

Topic	Fixable?	Context/Regulatory dependant?	Result dependant?	Assessment level or purpose dependant?	Dependent from data accessibility?
Reliability	Potentially	No	No	No	Yes
Relevance	No	Yes	No	No	No
Adequacy	No	Yes	Potentially	Yes	No

References:

- Hill AB (1965). The environment and disease: Association or causation? Proc Royal Soc Medicine 58:295–300.
- Kase R, Korkaric M, Werner I, Ågerstrand M (2016). Criteria for Reporting and Evaluating Ecotoxicity Data (CRED): comparison and perception of the Klimisch and CRED methods for evaluating reliability and relevance of ecotoxicity studies. PMID 27752442 PMCID PMC5044958, DOI 10.1186/s12302-016-0073-x Environ Sci Eur 28(1):7.
- Klimisch HJ, Andreae M, Tillmann U (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. PMID 9056496 DOI 10.1006/rtph.1996.1076 Regul Toxicol Pharmacol 25:1–5.
- Küster A, Bachmann J, Brandt U, Ebert I, Hickmann S, Klein-Goedicke J, Maack G, Schmitz S, Thumm E, Rechenberg B (2009). Regulatory demands on data quality for the environmental risk assessment of pharmaceuticals. PMID 9607869 DOI 10.1016/j.yrtph.2009.07.005 Regul Toxicol Pharmacol 55(3):276-80.
- Wess RA (2016). The Question of Causation and Adequacy — Iron as an Example of Intrinsic Toxicity and other Effects. PMID 26632140 DOI 10.1002/ieam.1722 Integr Environ Assess Manag 12(1):202-4.