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).