Regulatory Toxicology and Pharmacology 82 (2016) 173–185

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Commentary

A quantitative weight of evidence methodology for the assessment of reproductive and developmental toxicity and its application for classification and labeling of chemicals

Wolfgang Dekant [a, *], James Bridges [b]

[a] Department of Toxicology, University of Würzburg, 97078 Würzburg, Germany
[b] University of Surrey, Guildford, UK

ARTICLE INFO

Article history:
Received 20 May 2016
Received in revised form
11 August 2016
Accepted 7 September 2016
Available online 10 September 2016

Keywords:
Weight of evidence
Classification and labeling
Reproductive and developmental toxicity
Systematic review

ABSTRACT

Hazard assessment of chemicals usually applies narrative assessments with a number of weaknesses. Therefore, application of weight of evidence (WoE) approaches are often mandated but guidance to perform a WoE assessment is lacking. This manuscript describes a quantitative WoE (QWoE) assessment for reproductive toxicity data and its application for classification and labeling (C&L). Because C&L criteria are based on animal studies, the scope is restricted to animal toxicity data. The QWoE methodology utilizes numerical scoring sheets to assess reliability of a publication and the toxicological relevance of reported effects. Scores are given for fourteen quality aspects, best practice receives the highest score. The relevance/effects scores (0 to four) are adjusted to the key elements of the toxic response for the endpoint and include weighting factors for effects on different levels of the biological organization. The relevance/effects scores are then assessed against the criteria dose-response, magnitude and persistence of effects, consistency of observations with the hypothesis, and relation of effects to human disease. The quality/reliability scores and the relevance/effect scores are then multiplied to give a numerical strength of evidence for adverse effects. This total score is then used to assign the chemical to the different classes employed in classification.

1. Introduction

The assessment of the hazardous properties and the potential human health risks from exposure to chemicals often requires the evaluation of a large number of studies providing information of widely differing nature and for hazard assessment and risk characterization. Lines of evidence may range from observational studies, such as human epidemiology or toxicity testing in animals, to mechanistically oriented studies or non-experimental approaches such as read across. The challenge for assessors is to utilize these different lines of evidence in a systematic, transparent and consistent way to arrive at an integrated and scientifically valid conclusion.

Due to these issues, many hazard assessments and risk characterizations of chemicals have either focused on worst case findings or have followed a standardized regimen using the results of the mandatory toxicity testing as central pillar for conclusions. Such studies follow specific testing protocols and are commonly considered as the key studies in hazard assessment due to quality control and detailed reporting (ECHA, 2015; Klimisch et al., 1997). Concerns have been voiced regarding the reliance on guideline toxicity studies since a significant part of the available information on potential hazards may not be considered (Beronius et al., 2014; James et al., 2015; Myers et al., 2009). In response, some regulatory authorities have proposed or stated that they have adopted a "Weight of Evidence' (WoE) to their data assessment (ECHA, 2015). The USA EPA uses the term 'systematic review' instead of WoE (US-EPA, 2005).

The WoE concept originates from the legal field (Weed, 2005). It is applied there to describe a situation where there are large variations in the nature and quality of evidence for a particular claim. WoE defines the process to be used to ascertain whether the information supporting one side of an argument is greater than that supporting the other side. Surprisingly, the legal field provides only very limited advice on the detailed application of WoE.

* Corresponding author. Department of Toxicology, University of Würzburg, Versbacherstrasse 9, 97078, Würzburg, Germany.
E-mail address: dekant@toxi.uni-wuerzburg.de (W. Dekant).

http://dx.doi.org/10.1016/j.yrtph.2016.09.009
0273-2300/

Glossary of terms

Weight of evidence method (WoE)	The identification and objective analysis (using predefined, scientifically justified criteria) of all potentially relevant studies, for their quality and relevance in testing a hypothesis (problem formulation)
Quantitative weight of evidence (QWoE)	The identification, objective analysis and numerical scoring (using predefined scientifically justified criteria) of all potentially relevant studies, for both their quality and relevance in testing a hypothesis (problem formulation)
The hypothesis	Generally, takes the form of "does chemical of interest X cause adverse effects Y under conditions Z". Conditions may include exposure levels and duration, species of interest, adverse effects are defined as by WHO/IPCS
Endpoints	The measured and modelled findings used to identify and characterize adverse effects Y
Quality	The reliance that can be placed on the findings of each study for the purpose of critically testing the hypothesis
Relevance	The utility of the findings of each study on adverse endpoints for the purpose of critically testing the hypothesis
Lines of evidence	The different types of investigation used to critically test the hypothesis (e.g. observations in man, targeted toxicity testing in animals, in vitro experiments determining molecular endpoints, and in silico predictions of toxicity based on read-across or quantitative structure activity relationships)
Weighting of endpoints	A multiplier that is applied to the relevance/effect scores to reflect the relative importance of different types of endpoint and/or different lines of evidence in support of the hypothesis
Strength of evidence	This score is derived by multiplying the final relevance/effects score by the quality/reliability score for a particular study
Overall weight of evidence	This is a summation of the findings from all suitable studies. It may be presented graphically as a plot of relevance/effects against quality scores or as an average numerical value with ranges

Regarding use of WoE in toxicology, marked differences also exist in both its purpose and in the way it is conducted (Linkov et al., 2009; Lutter et al., 2015; NRC, 2011; NRC, 2014; Rhomberg et al., 2013; Rooney et al., 2014; SCENIHR, 2012). For example, ECHA (2015) defines WoE as "*the process of considering the strengths and weaknesses of various pieces of information in reaching and supporting a conclusion concerning a property of the substance*". While ECHA considers "reliability, adequacy, relevance and quality" as essential components of a WoE approach (ECHA, 2010), little guidance is given other than to apply Klimisch et al. (1997) criteria for quality. ECHA provides no guidance as to how to integrate complex and sometimes contradictory research findings into the overall evaluation process. In addition, the term 'Weight of Evidence' is often used indiscriminately by regulatory agencies to support conclusions on hazard and risk without indications of how it was applied. Consequently, hazard assessments and risk characterizations may yield widely differing outcomes (Agerstrand et al., 2014; Beronius et al., 2010; Golden et al., 2003; Ruden, 2001a, b).

Since the purpose of applying WoE is to make the scientific judgments clearer, more consistent and less susceptible to bias, it is important to have a transparent and well-developed WoE framework (Lutter et al., 2015). WoE has been defined as *the testing of a hypothesis (problem formulation) using predefined scientifically justified criteria for quality, relevance, and impact of observations* (Lutter et al., 2015). In addition, WoE assessments should be based on the best available science, be objective and consider all available data sources in a transparent and reproducible manner. It is therefore inappropriate to give priority to particular data without a detailed justification or to adopt a conservative/precautionary approach without providing the rationale of the nature and dimension of the uncertainties that have led to its adoption (Guzelian et al., 2005).

While general WoE-approaches may be applied to a variety of issues in toxicology, the WoE presented here is quantitative (thus termed quantitative WoE, QWoE) and is developed specifically for application to hazard classification mandated by EC regulation 1272/2008 (EC-Regulation, 2008). As outlined before, a WoE-analysis requires a well-framed question or hypothesis, which, in this case is "Does the chemical have a capacity to induce adverse effects on reproductive endpoints" due to the focus on classification and labeling (C&L) and addresses hazard rather than risk because C&L is hazard-based. The field of reproductive toxicity is particularly appropriate to a QWoE evaluation since chemicals suspected of inducing reproductive effects often have complex and controversial databases (EFSA, 2015). Since clear indications for effects of a specific chemical on reproductive endpoints in humans are usually not available, animal toxicity studies are the major supportive data for C&L. Therefore, the QWoE developed concentrates on an assessment of the animal toxicity studies for purposes of C&L. As the CLP legislation follows the international UN Globally Harmonized System (GHS) applied by the EU, Japan, Korea, and a number of other countries for classification, this QWoE approach is widely relevant.

The QWoE relies on a numerical scoring system that consists of two key components, a score for quality/reliability and a score for (toxicological) relevance and effects (detailed in Fig. 1). Study quality scores encompass appropriateness of the scientific methodology, statistical approach, and reporting based on best practice (Table 1). Relevance/effects scores capture to what extent the observations support the hypothesis, i.e. a chemical induces adverse effects to reproduction and provide a quantitative measure of effects induced by the chemical of interest. Development of relevance/effects scores includes weighting observations based on their toxicological significance since the scoring of complex biological response patterns needs to consider the relative weight of an observation based on its adversity. For C&L, adversity of reported effects (WHO/IPCS, 2004) is the key determinant for classification. The relevance/effects score therefore integrates both the nature of the observation and the extent of support for the conclusion that the observation is toxicologically and biologically meaningful and adverse. The final strength of evidence score for a study is then obtained by totaling the strength scores for all observations and integrated into different classes of evidence to support an adverse effect that may serve as the basis for classification. It is important that scoring sheets for all studies both regarding quality and relevance/effects are openly available and include the rationale for assigning specific scores.