Regulatory Toxicology and Pharmacology 72 (2015) 94–101

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Evaluation of the ToxRTool's ability to rate the reliability of toxicological data for human health hazard assessments ☆

D. Segal [a,*], S.L. Makris [a], A.D. Kraft [a], A.S. Bale [a], J. Fox [a], M. Gilbert [c], D.R. Bergfelt [d], K.C. Raffaele [e], R.B. Blain [f], K.M. Fedak [f], M.K. Selgrade [f], K.M. Crofton [b]

[a] EPA Office of Research and Development (ORD), National Center for Environmental Assessment, United States
[b] EPA ORD National Center for Computational Toxicology, United States
[c] EPA ORD National Health and Environmental Effects Research Laboratory, United States
[d] Dept. of Biomedical Sciences, Ross University School of Veterinary Medicine, St. Kitts, St. Kitts and Nevis
[e] EPA Office of Solid Waste and Emergency Response, United States
[f] ICF International Environment and Social Sustainability Division, United States

ARTICLE INFO

Article history:
Received 30 December 2014
Available online 14 March 2015

Keywords:
ToxRTool
Systematic review
Human health risk assessment
Risk of bias
Data quality

ABSTRACT

Regulatory agencies often utilize results from peer reviewed publications for hazard assessments. A problem in doing so is the lack of well-accepted tools to objectively, efficiently and systematically assess the quality of published toxicological studies. Herein, we evaluated the publicly available software-based ToxRTool (Toxicological data Reliability assessment Tool) for use in human health hazard assessments. The ToxRTool was developed by the European Commission's Joint Research Center in 2009. It builds on Klimisch categories, a rating system established in 1997, by providing additional criteria and guidance for assessing the reliability of toxicological studies. It also transparently documents the study-selection process. Eight scientists used the ToxRTool to rate the same 20 journal articles on thyroid toxicants. Results were then compared using the Finn coefficient and "AC1" to determine inter-rater consistency. Ratings were most consistent for high-quality journal articles, but less consistent as study quality decreased. Primary reasons for inconsistencies were that some criteria were subjective and some were not clearly described. It was concluded, however, that the ToxRTool has potential and, with refinement, could provide a more objective approach for screening published toxicology studies for use in health risk evaluations, although the ToxRTool ratings are primarily based on study reporting quality.

Published by Elsevier Inc.

1. Introduction

Leading scientific organizations have expressed a need for standardizing and documenting criteria used to evaluate the quality of toxicological studies considered for use in human health hazard evaluation and risk assessments (Rooney et al., 2014). Many governmental and regulatory organizations, both within the United States (e.g., U.S. Environmental Protection Agency, Agency for Toxic Substances and Disease Registry, California Environmental Protection Agency, National Institute of Environmental Health Science (NIEHS) National Toxicology Program (NTP), and Food and Drug Administration) and internationally (e.g., World Health Organization International Programme on Chemical Safety, European Chemicals Agency) rely on published peer-reviewed literature to develop health assessments of chemicals. Yet, at times, the quality of published data has been unknown or questionable (Schneider et al., 2009). Therefore, the National Academy of Sciences (NRC, 2011) has indicated that applying standard study quality criteria would improve the transparency and consistency of risk assessments. In addition, the European Union's REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) guidance documents (ECHA, 2008) call for the thorough evaluation of all data, at the start of the chemical registration process, to determine if they are reliable and sufficient for use in regulatory decisions. Rating of data quality and reliability were also identified as major challenges by a Society of Toxicology (SOT) expert group, especially when underlying data are not available, as is typically the case in published papers (SOT, 2007). Thus, having a reliable and objective tool to evaluate study quality and document the process would improve risk assessment methods. Currently, several efforts, including those of the European

☆ The views expressed in this paper are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

* Corresponding author. Fax: +1 703 347 8692.
E-mail address: segal.deborah@epa.gov (D. Segal).

http://dx.doi.org/10.1016/j.yrtph.2015.03.005
0273-2300/Published by Elsevier Inc.

Commission (EC), NIEHS NTP (Rooney et al., 2014) and University of California, San Francisco (Woodruff and Sutton, 2011), are underway for developing such a tool. The EC-based efforts are the most advanced and have resulted in the publicly available software-based ToxRTool (Schneider et al., 2009), which is the focus of this evaluation.

The ToxRTool builds on the Klimisch categories that code toxicological studies based on reliability criteria. Klimisch et al. (1997) defined *reliability* as "the inherent quality of a test report or publication relating to preferably standardized methodology and the way the experimental procedures and results are described to give evidence of the clarity and plausibility of the findings." In the Klimisch system, a category code of "1" indicates that a study is "reliable without restriction" and adhered to valid and/or internationally accepted testing guidelines (preferably performed according to good laboratory practices [GLP]). A "2" indicates that a study is "reliable with restriction" and usually was not performed according to GLP, but nonetheless is well documented and scientifically acceptable. A code of "3" means that a study is "not reliable," due to major scientific flaws or lack of documentation, and a "4" indicates that a paper is not assignable because it does not contain primary data (Klimisch et al., 1997).

To aid in the placement of studies into reliability categories, Klimisch et al. (1997) outlined nine criteria for in vivo studies and six for in vitro studies. Although these criteria are provided, explicit guidance was not included. Additionally, only those studies following recent guidelines and GLPs are generally assigned to the "reliable without restriction" category. Therefore, high-quality studies that were conducted prior to the publication of GLP standards might be unfairly downgraded to a lower Klimisch score. In addition, many academic studies are non-GLP, yet still meet high scientific standards. To address these ambiguities, the European Centre for the Validation of Alternative Methods (ECVAM), of the European Commission Joint Research Centre, initiated a project to provide better guidance in applying the Klimisch categories by developing more explicit criteria that did not require a study to follow GLP to be rated as high quality. ECVAM also sought to create a more elaborate documentation process to make the selection of toxicological studies more transparent. As a result ECVAM developed ToxRTool (Toxicological Data Reliability Assessment Tool), a Microsoft® Office Excel spreadsheet-based tool publicly available on the internet (Schneider et al., 2009). Because the ToxRTool is intended for the evaluation of literature studies with primary data, category "4" from the Klimisch categories was excluded.

An initial draft version of ToxRTool was developed and then refined following an evaluation in which the resulting categorizations by different scientists for the same studies were found to be excessively heterogeneous (Schneider et al., 2009). This was followed by a second evaluation that resulted in a few minor linguistic adjustments to some of the criteria. The final version of ToxRTool contains two templates in the form of Excel© spreadsheets with a series of questions on a variety of components of the study report. One template was provided for in vitro studies and includes 18 criteria. A separate template for in vivo studies includes 21 criteria. The ToxRTool does not provide a template for epidemiologic studies and, as such, is not intended for their evaluation.

The purpose of this study was to evaluate the strengths and weaknesses of ToxRTool and to determine if it provides the consistency needed for review of studies under consideration for human health hazard assessments. The present analysis was conducted to determine if eight scientists evaluating the same set of studies with the ToxRTool would arrive at similar reliability scores. These studies represented a small set of papers specific to thyroid disrupting chemicals typical of those encountered for use in human health assessments.

2. Methods

2.1. Study selection

An independent expert selected 20 (11 in vivo and 9 in vitro) peer-reviewed studies on thyroid disrupting chemicals from the open literature (Table 1). These studies were chosen from a large eLibrary of over 1000 primary literature reports of thyroid disrupting chemicals (Baker et al., 2013). From this eLibrary, a small set of papers were selected based on the following subjective criteria: (1) quality of the reporting for the chemicals and methods used, (2) numbers of animals for in vivo papers, (3) number of replicates for in vitro papers, and (4) reporting quality for the results, including the amount of data presented (group means with variance for all treatments vs. summary calls on results). Publications were selected to range from excellent to low quality with respect to these criteria, as subjectively assessed by the expert. In order not to add any *a priori* bias, the independent expert did not participate in the use of the ToxRTool to rate the reliability of the selected journal articles.

2.2. Use of ToxRTool to evaluate studies

Eight scientists with various levels of experience in the areas of thyroid toxicology and risk assessment evaluated the 20 selected publications with the ToxRTool. One of the raters was an expert on thyroid toxicology, two had moderate levels of expertise on thyroid toxicology and extensive risk assessment experience, and five were toxicologists in other areas, but also had extensive risk assessment experience. Six of the raters were employed by the USEPA and two were employed by ICF International Environment and Social Sustainability Division. All of the raters were unfamiliar with the ToxRTool prior to participation in this project. Thirteen studies were evaluated by eight raters, and seven were evaluated by seven raters. Each rater evaluated the articles using the ToxRTool downloaded from the European Commission's Joint Research Center website: <http://ihcp.jrc.ec.europa.eu/our_labs/eurl-ecvam/archive-publications/toxrtool>.

The raters used the in vitro ToxRTool spreadsheet to indicate whether 18 criteria were met in the following 5 areas: (1) test substance identification, (2) test system characterization, (3) study design description, (4) study results documentation, and (5) plausibility of study design and data. For in vivo studies, the raters used a separate spreadsheet that listed 21 criteria in the same 5

Table 1
Published papers selected for Evaluation with the ToxRTool.

Study (in alphabetical order by study type)	Study type
Christenson et al. (1995)	In vivo
Dalton et al. (2003)	In vivo
Darnerud and Thuvander (1998)	In vivo
Florsheim and Velcoff (1962)	In vivo
Fowles et al. (1994)	In vivo
Fregly et al. (1968)	In vivo
Gray and Kavlock (1983)	In vivo
Sciarrillo et al.(2008)	In vivo
Siglin et al. (2000)	In vivo
Smith et al. (1986)	In vivo
Villeneuve et al. (1979)	In vivo
Cheek et al. (1999)	In vitro
Doerge et al. (1998)	In vitro
Freyberger and Ahr (2006)	In vitro
Gaitan et al. (1983)	In vitro
Hohenwarter et al. (1996)	In vitro
Manzon and Youson (2002)	In vitro
Santini et al. (2003)	In vitro
Sun et al. (2008)	In vitro
Tonacchera et al. (2004)	In vitro