Science of the Total Environment 482 (2014) 358–365

Contents lists available at ScienceDirect

Science of the Total Environment

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

Methods for assigning confidence to toxicity data with multiple values — Identifying experimental outliers

Fabian P. Steinmetz [a], Steven J. Enoch [a], Judith C. Madden [a], Mark D. Nelms [a], Neus Rodriguez-Sanchez [a], Phil H. Rowe [a], Yang Wen [a,b], Mark T.D. Cronin [a,*]

[a] School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, England, United Kingdom
[b] School of Environmental Sciences, Northeast Normal University, Changchun, China

HIGHLIGHTS

- A statistical approach for the quality of data with multiple values was developed.
- Confidence scores were assigned to toxicity data for individual compounds.
- Confidence scores were related to the number of entries and their variability.
- Using higher quality data allowed for the development of more robust QSARs.

ARTICLE INFO

Article history:
Received 16 October 2013
Received in revised form 14 February 2014
Accepted 25 February 2014
Available online xxxx

Keywords:
Data quality
Microtox
Confidence
QSAR
Predictive toxicology
Conflicting data

ABSTRACT

The assessment of data quality is a crucial element in many disciplines such as predictive toxicology and risk assessment. Currently, the reliability of toxicity data is assessed on the basis of testing information alone (adherence to Good Laboratory Practice (GLP), detailed testing protocols, etc.). Common practice is to take one toxicity data point per compound — usually the one with the apparently highest reliability. All other toxicity data points (for the same experiment and compound) from other sources are neglected. To show the benefits of incorporating the "less reliable" data, a simple, independent, statistical approach to assess data quality and reliability on a mathematical basis was developed. A large data set of toxicity values to *Aliivibrio fischeri* was assessed. The data set contained 1813 data points for 1227 different compounds, including 203 identified as non-polar narcotic. Log K_{OW} values were calculated and non-polar narcosis quantitative structure–activity relationship (QSAR) models were built. A statistical approach to data quality assessment, which is based on data outlier omission and confidence scoring, improved the linear QSARs. The results indicate that a beneficial method for using large data sets containing multiple data values per compound and highly variable study data has been developed. Furthermore this statistical approach can help to develop novel QSARs and support risk assessment by obtaining more reliable values for biological endpoints.

1. Introduction

High quality data are preferred for model development in predictive toxicology. They are also required as a benchmark in the assessment of alternative assays and to enable analysis of toxicological pathways. Recently, further toxicity data have become available through the development of the OECD QSAR Toolbox, release of information from dossiers submitted to the European Chemicals Agency (ECHA), the eChemPortal and other sources (Fourches et al., 2010; Przybylak et al., 2012; Péry et al., 2013; Cronin and Schultz, 2003).

When using these expanding resources of toxicity data for risk assessment purposes and modelling, the quality and reliability of the data must be assessed. A given data set could be too "poor" in terms of quality for quantitative structure–activity relationship (QSAR) modelling but still satisfactory for the prioritisation of chemicals for testing or regulatory classification and labelling. Whilst QSAR modelling is dependent on a sensitive statistical analysis, e.g. multivariate regression, to define reasonable descriptors, regulatory use of toxicity data may only need a rough estimation of hazard as a worst case assumption, with extrapolation factors being applied (Nendza et al., 2010).

Reliability is the measure of the extent of repeatability and reproducibility of a toxicity test for a particular chemical (OECD, 2003). As repeatability and reproducibility are not known for most data, a variety of approaches to assign reliability and confidence are used. Assessing

* Corresponding author. Tel.: +44 151 231 2402.
E-mail address: M.T.Cronin@ljmu.ac.uk (M.T.D. Cronin).

http://dx.doi.org/10.1016/j.scitotenv.2014.02.115
0048-9697/

data quality in predictive and computational toxicology is, however, a difficult task (Yang et al., 2013; Przybylak et al., 2012; Klimisch et al., 1997).

There are a number of established criteria to ascertain the reliability of toxicity data. The most commonly applied are those proposed by Klimisch et al. (1997). These authors discussed data attributes such as reliability, relevance and adequacy and provided a scoring system to categorise data into reliability classes:

1. reliable without restriction
2. reliable with restrictions
3. not reliable
4. not assignable.

Recently Przybylak et al. (2012) applied the Klimisch scheme and an updated scoring approach, based on ECHA guidance on information requirements and chemical safety assessment, to "real life" problems of toxicity data harvesting. In the updated work, the focus was on availability of information, consistency of study design, adherence to Good Laboratory Practice (GLP), test chemical identity and toxicological data.

Whilst reliability (the backbone of an experiment and the resulting toxicity data), and relevance (the usefulness of the resulting data for the desired purpose such as risk assessment) in principle require interpretation by experts, the determination of reliability of data can be as well supported by methods of "weighting" the data (Yang et al., 2013; Przybylak et al., 2012; Klimisch et al., 1997). When dealing with large sets of toxicity data, from multiple sources, there is often more than a single data entry for each compound. In this investigation these data entries are referred to as "conflicting data". Even for a well-defined assay such as the acute fish toxicity test, considerable variability in potency is seen within the results for the same compound (Hrovat et al., 2009). If toxicity data are to be extracted for modelling from the increasing number of databases then criteria to identify reliable values are required. In particular, it would be helpful to be able to score data for reliability. In this way, it may be possible to combine what may be considered to be low quality data to obtain a more reliable score.

Another interesting aspect of the issue of data quality control and assurance was investigated by Ruusmann and Maran (2013) who undertook an extended data harvest for the *Tetrahymena pyriformis* inhibition of growth assay (Tetratox assay). They analysed the "timelines" associated with the reporting of chemical structures and experimental data and so examined when and how certain data were reported in the scientific literature over time. These authors came to the conclusion that mathematical manipulation (rounding, building averages, etc.) and, of course, human error has led to differences in the data reported. For some compounds, there are many toxicity data from the same test, there is, however, no unified strategy to select which of the data to use, or how to use them. Often these toxicity data for the same compound have a normal distribution that makes it relatively easy to define a representative value via the median or arithmetic mean. Data which fall outside the normal distribution may be termed "data outliers", i.e. they may be subject to considerable experimental error. Fig. 1 illustrates the issues of the presence of a data outlier in reducing certainty in the calculation of the mean or median.

In principle, the arithmetic mean is a good way to consolidate associated data points to a single value. Here, every data point is taken into consideration, in equal parts, to build a new value — the arithmetic mean. In contrast the median is the middle value of a distribution. When dealing with high individual spreads, the median is the more stable approach (Rowe, 2007).

Confidence scoring is based on the number and variability of conflicting data. In this context, the relative standard deviation (RSD; sometimes referred as coefficient of variation), as a quotient of standard deviation and arithmetic mean, expresses the variability of a data set of toxicity values for one compound (Rowe, 2007). Thus a high number of entries per compound and a low RSD lead to high confidence and vice versa.

In order to investigate the role of variability in toxicity databases and explore the possibility of applying statistical approaches to identify reliable toxicity data, historical toxicity data, measured in the Microtox assay (and its precursors), were considered. Such data have been published since the early 1980s (e.g. Dutka and Kwan, 1981; Chang et al., 1981; Bulich et al., 1981; King and Painter, 1981; Curtis et al., 1982; Yates and Porter, 1982; DeZwart and Slooff, 1983; Ribo and Kaiser, 1984) and by the company Beckman Instruments, Inc. (now Beckman Coulter, Inc.). The *Aliivibrio fischeri* toxicity assay (Microtox) is a standardised aquatic toxicity test based on the marine bacterium *A. fischeri* (also known as *Photobacterium fischeri* and *Vibrio fischeri*). The photo-luminescent bacteria are exposed to a chemical at different concentrations with the reduction of light emitted being regarded as the effect. The results from the Microtox assay include the concentration of a compound where light intensity is reduced by 50% (EC_{50}). The pT value is the negative logarithm of the EC_{50}, for the purposes of this paper the units are in mmol L^{-1}, and the measurement has historically been taken at different exposure times (5, 15 and 30 min) (Kaiser and Palabrica, 1991). As the *A. fischeri* toxicity assay is a well standardised study, little experimental variability is assumed. However, there are some data which can be regarded as low quality, which may be attributed to interlaboratory variation and experimental error. Cronin and Schultz (1997) furthermore suggested that there is no significant influence of exposure times (5, 15 and 30 min) on the toxicity of compounds which act by non-polar narcosis. In this study non-polar narcosis is taken to be a non-specific mechanism of acute toxicity brought about by membrane perturbation (van Wezel and Opperhuizen, 1995; Ellison et al., 2008). As such, in aquatic toxicology, it is well established that the octanol/water partition coefficient (K_{OW}) is strongly related to

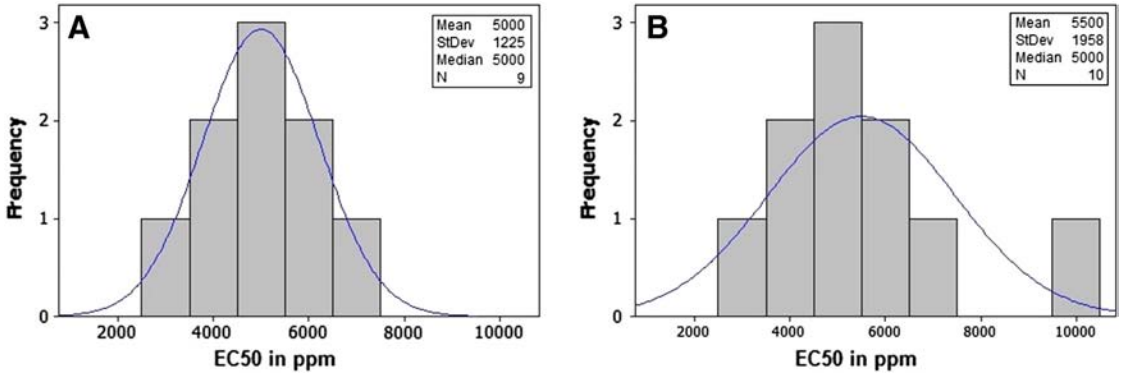

Fig. 1. Normal distribution bell for a sample data set (representative EC_{50} values from different sources for one compound) with an "optimal" normal distribution (A) and with a data set containing an outlier in the upper range (B) demonstrating the skew it may bring to the distribution.