REGULATORY TOXICOLOGY AND PHARMACOLOGY **25,** 1–5 (1997)
ARTICLE NO. RT961076

A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data[1]

H.-J. KLIMISCH,[2] M. ANDREAE, AND U. TILLMANN

BASF Aktiengesellschaft, D-67056 Ludwigshafen, Germany

Received November 7, 1996

The evaluation of the quality of data and their use in hazard and risk assessment as a systematic approach is described. Definitions are proposed for reliability, relevance, and adequancy of data. Reliability is differentiated into four categories. Criteria relating to international testing standards for categorizing reliability are developed. A systematic documentation of evaluating reliability especially for use in the IUCLID database is proposed. This approach is intended to harmonize data evaluation processes worldwide. It may help the expert in subsequent assessments and should increase the clarity of evaluation.

INTRODUCTION

Hazard and risk assessment for "existing substances" must be carried out in Europe based on Council Regulation 793/93 (EEC, 1993) and following principles of Commission Regulation 1488/94 (EEC, 1994). All relevant available information/data and corresponding study reports of substances, published in a priority list, must be submitted by the manufacturer/importer using a special software package on disk (IUCLID: International Uniform Chemical Information Database) and as hard copies. During the risk assessment process the assessor must consider whether the supplied data are complete and valid for use in risk assessment. This is particularly important for data on "existing substances" (EINECS, 1981). There may be a number of test results available for each end point but some or all of them may have not been carried out following current standards in toxicology and ecotoxicology.

Before a hazard identification may be performed, the supplied data must be evaluated considering their quality and adequacy for a risk assessment. Some general guidelines on data evaluation were published by the European Commission in a "Technical Guidance Document" (EU, 1994, 1995), based on general principles for data evaluation of the International Coordination of Criteria Document Production (IPCS, 1993). Considerations of the assessment on the quality of data have been described also by OECD (1994). Some experience in the evaluation of data was developed in Germany by the "GDCH-Advisory Committee on Existing Chemicals of Environmental Relevance (BUA)." During the past decade approximately 200 BUA reports were published by this committee composed of representatives of academia, the chemical industry, and government. During this work a systematic approach on data validation was found useful. It gives definitions, discusses a score system with different reliability categories according to validity, defines criteria, and generates a system for standardized documentation of validity evaluation to be used also in data sheets (IUCLID). It appeared to be useful to describe this approach on behalf of BUA in order to initiate harmonization of similar processes in data evaluation worldwide and to facilitate the exchange of experience toward improvement of such approaches. A characterization of the validity of experimental data should also help the expert to assess the effect of end points consistently and thus to increase clarity in hazard or risk assessment processes.

DEFINITIONS

Different terms are being used synonymously to characterize the quality of the data of toxicological and ecotoxicological studies: validation/validity, reliability, adequacy. These terms describe not only procedures to define the quality of test results (data), but also test methods are validated to prove their relevance and reproducibility. Validity statements are also given in assessment processes especially if an expert must decide which data of, for instance, conflicting study results are representative/relevant to describe an effect correctly.

[1] This paper is published on behalf of BUA (GDCH Advisory Committee on Existing Chemicals of Environmental Relevance, Federal Republic of Germany).

[2] To whom reprint requests should be addressed at Department of Toxicology.

0273-2300/97 $25.00

The following definitions are proposed here to be used in hazard and risk assessment processes:

Reliability—Evaluating the inherent quality of a test report or publication relating to preferably standardized methodology and the way that the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings.

Relevance—Covering the extent to which data and/or tests are appropriate for a particular hazard identification or risk characterization.

Adequacy—Defining the usefulness of data for risk assessment purposes. When there is more than one set of data for each effect, the greatest weight is attached to the most reliable and relevant.

The evaluation needs expert judgment and should be clear, so that the use made of a particular data set is clearly justified and understood by others. Agreement on standardized criteria for characterizing and differentiating the quality of data (their reliability, relevance, and adequacy) may be useful for a broader understanding and acceptance worldwide. Such evaluation of the quality of individual studies/data is a step in compiling data in the form of a "data sheets" for a substance (IUCLID, etc.) for hazard or risk assessment purposes. Such data sheets have the intention of making available all toxicological and ecotoxicological data about a substance and keeping them updated to the actual state of knowledge. Furthermore, if information about the quality of the individual test/data is given in such a data sheet, this would help to identify more easily those data preferably used for risk assessment.

CATEGORIES OF RELIABILITY

Test data of toxicological and ecotoxicological laboratory studies may be available as described in

—individual test reports
—publications (literature)
—review articles
—abstracts of presentations
—any other short information (safety data sheets, handbooks, etc.).

The more that details on methodology, test procedures, and analytics are documented, the easier an evaluation of their reliability should be in general. The amount of information presented will thus provide the basis for deciding on the reliability of data reported. Tests conducted and reported according to internationally accepted test guidelines (EU, EPA, FDA, OECD) and in compliance with the principles of Good Laboratory Practice (GLP) should have the highest grade of reliability and should be used as reference standards when evaluating the reliability of tests generated prior to the requirements of GLP and the international standardization of testing methods.

Our approach proposes to indicate a measure of the study/data reliability. Therefore, the quality of laboratory studies and of data from the literature may be differentiated and thus classified according to four categories of reliability.

The following categories/codes of reliability seem to be adequate:

Code	Category
1	Reliable without restriction
2	Reliable with restrictions
3	Not reliable
4	Not assignable

An additional Code 5 may be added to identify information/data which were *not evaluated* according to their reliability (special studies on, for instance, pharmacologic or mechanistic effects) without particular relevance for hazard/risk assessment.

The following definitions of these categories were found practicable to differentiate reliability (Codes 1–4):

1. Reliable without Restriction

This includes studies or data from the literature or reports which were carried out or generated according to generally valid and/or internationally accepted testing guidelines (preferably performed according to GLP) or in which the test parameters documented are based on a specific (national) testing guideline (preferably performed according to GLP) or in which all parameters described are closely related/comparable to a guideline method.

2. Reliable with Restrictions

This includes studies or data from the literature, reports (mostly not performed according to GLP), in which the test parameters documented do not totally comply with the specific testing guideline, but are sufficient to accept the data or in which investigations are described which cannot be subsumed under a testing guideline, but which are nevertheless well documented and scientifically acceptable.

3. Not Reliable

This includes studies or data from the literature/reports in which there are interferences between the measuring system and the test substance or in which organisms/test systems were used which are not relevant in relation to the exposure (e.g., unphysiologic pathways of application) or which were carried out or generated according to a method which is not acceptable, the documentation of which is not sufficient for an

assessment and which is not convincing for an expert judgment.

4. Not Assignable

This includes studies or data from the literature, which do not give sufficient experimental details and which are only listed in short abstracts or secondary literature (books, reviews, etc.).

CRITERIA FOR RELIABILITY CATEGORIES

In order to help in assigning a study to a category/code of reliability, some criteria should be considered more specifically, according to which the quality of the study in relation to standard methods and the scope of the documentation are assessed. Depending on the type of study, a differentiated evaluation by the expert is required: In the case of acute studies, the requirements may generally be interpreted more flexible and broadly than, for example, in the case of carcinogenicity studies. The following general criteria should be considered.

The standard methods recommended, e.g., by OECD or EU, are used as a reference. GLP principles should preferably be considered so that the reproducibility and acceptance according to the state-of-the-art of the results are guaranteed as far as possible. If a complete report is available or if the test, although not performed according to national/international standard methods, is described sufficiently and carried out according to a scientifically acceptable standard, the studies may be assessed as "reliable" as well. This also applies to literature publications. The basic data (test organisms, data on the method, and on the scope of the investigations) should be available and documented in the data set of the substance especially if a standard method was not used. Data on the purity of a substance are necessary particularly if impurities may have a substantial influence on the toxicity. This can be assessed only on a case-to-case basis. Information on dose/concentration is essential. Even if some criteria of an international standard are not met, the expert may decide that the study is "reliable with restrictions" and may be used for a risk assessment.

Toxicity Studies

The following information/data should generally be available and reported for *animal studies* which were not carried out according to an international/national standard method:

—Data/information on the test animals (species, strain, sex, age);
—Purity/composition/origin of the test substance;
—Number of animals evaluated;
—Scope of the investigations per animal (for instance, clinical chemistry, hematology, organ weights, pathology or histopathology) and description of the methods;
—Description of the changes/lesions observed;
—Control group or historical control data of the laboratory;
—Description of the test conditions;
—Description of the route and doses of administration (preferably including analytical verification);
—Dose/concentration relationship if possible.

The following data/information should be available for *in vitro studies* which were not carried out according to an international/national standard method:

—Description of the test system and test method in details;
—Purity/composition/origin of the test substance;
—Data on the dose/concentration differentiated according to the toxicity of the test substance on the test system; information on volatility;
—Data on secondary effects which may influence a result (solubility, impurities, pH shifts, influence on the osmolarity, etc.);
—Appropriate negative/positive controls as integral parts of the test;
—References on adequacy of the method should be given or generally known.

The usefulness will be particularly influenced by the adequacy of the method.

Ecotoxicity Studies

For assessing the *reliability* of ecotoxicological studies, which are not carried out according to national/international test guidelines, the following items should be screened (expert judgment):

Acute studies.

- Clear description of the test procedure (complete documentation)
- Specification of the test substance (purity, by-products)
- Data on the test species and the number of individuals tested
- Data on the measured parameters (including definitions)
- Data on exposure period
- Use of emulgators/solubilizers[3]
- Data on concentration control analysis[3]
- Data on neutralization of samples[4]
- Data on physical and chemical test conditions (pH value, conductivity, light intensity, temperature, hardness of water)

[3] Especially in case of poorly soluble and unstable substances.
[4] In case of basic and acid substances.

• Determined effect concentrations (EC/LC/NOEC/LOEC)
• Data on the statistical evaluations (including method)
• Data on dosing the test substance (static, semistatic, flow through system).

Additional items in case of chronic studies.

• Information about the investigated period of the life cycle of the test animals
• Data on feeding of test animals.

DOCUMENTATION OF RELIABILITY CATEGORIES IN DATA SHEETS (IUCLID)

A short justification should be given in writing for assigning data of a study to a code/category of reliability. This should help in making such an expert decision transparent and understandable. For codes/categories 1 and 2 only short phrases may be necessary to justify such an assignment: for instance, "OECD Guideline study: GLP," etc. A more detailed justification should be given particularly for studies which are assigned to Code 3 (unreliable). The justification must be documented. If the data are compiled in a data sheet (IUCLID), this may preferably be reported in such a computer-based system in an additional field "reliability" under each individual test. The responsible "European Chemicals Bureau" has generated a software package for the latest IUCLID version 2.12 (ECB, 1996) according to the following patterns:

Example: Reliability, Code number (wording) justification statement.

Code/Category of Reliability

Reliability 1. (Reliable without restriction) short free text phrases, for instance:

—Guideline study (OECD, etc.)
—Comparable to guideline study
—Test procedure according to national standards (DIN, etc.).

Reliability 2. (Reliable with restrictions) short free text, for instance:

—Acceptable, well-documented publication/study report which meets basic scientific principles
—Basic data given: comparable to guidelines/standards
—Comparable to guideline study with acceptable restrictions.

Reliability 3. (Not reliable) more detailed free text.

—Method not validated
—Documentation insufficient for assessment
—Does not meet important criteria of today standard methods
—Relevant methodological deficiencies
—Unsuitable test system.

Reliability 4. (Not assignable) short free text

—Only short abstract available
—Only secondary literature (review, tables, books, etc.).

Relevance/Adequacy

As described the evaluation of reliability is performed considering certain formal criteria using international standards as references. It should clearly be stated that it is not the intention of this procedure to automatically exclude all unreliable data from further consideration by experts in risk assessment. The classification into different reliability categories should help the assessor especially in cases when conflicting results regarding one end point are reported. In such cases results of studies with a higher reliability should have greater weight for being used in risk assessment.

If for example results of *in vitro* tests are available (positive and negative Ames test), the test with the higher reliability may be more relevant. Therefore the assessment of relevance is very important and only the expert can decide which test describes the effect "correctly." Ames tests carried out according to International Testing Guidelines and GLP but using different purities of the substance may lead to positive and negative results depending on the reactivity and quantity of impurities. Both tests may have a high reliability but only the test without the reactive impurity may be relevant if this is the chemical to be used. Only this test should be considered as adequate for risk assessment.

The *relevance* of an ecotoxicological study should be elucidated in the light of the following questions:

• Is the testing strategy (organism, exposure scenario) aligned with the occurrence and the persistance of the test substance in the environment (target compartment)?
• Is it possible to derive useful ecotoxicological information from data obtained from experiments with nonstandard organisms (specialist, spread)?
• Are physical/chemical properties of the test substance (stability against hydrolytic and photolytic attacks, volatility, solubility) sufficiently considered before planning the test design?

Data with lower reliability may also be used as *supporting* information especially if the results are comparable or in the similar range; even in a case where only data with limited reliability are available, they may be used for definitive assessments of risk if the assessor considers these data as relevant (plausible) for risk assessment. For instance, LD_{50} values from studies with

rats, rabbits, and dogs, each with limited information on methodology, were considered as of limited reliability or even unreliable. But despite these reliability limitations, the assessor may use such data for risk assessment if the LD_{50} are within an acceptable range, evaluated in combination such that they are relevant (plausible), and show an only low interspecies variability. The same applies to an carcinogenicity study with too small a number of mice but showing a carcinogenic effect similar to that obtained in a reliable guideline study on rats.

It is not the aim of this paper to define criteria when studies with a restricted quality may nevertheless be used for hazard or risk assessment. This can only be decided by expert judgment on a case-by-case basis. Also data on structurally related compounds (SAR) should be used to define the relevance and adequacy of test results. All available experimental data as compiled in a data sheet (IUCLID) should be considered in risk assessment because only the totality of data will increase clarity of the conclusions. Thus limitations of publishing only a "definitive data set" appear to limit clarity and worldwide understanding. The relevance and adequacy of all the data used in a risk assessment process should be defined by expert judgment in a comprehensive report. Thus conclusions on relevance to humans of effects observed in studies in animals must be explained to make this interpretation clear and to gain a better understanding of the mechanism of action of a substance.

The proposed systematic approach to define and differentiate reliability of data should help experts worldwide to decide about relevance of the data for humans and their adequacy in risk assessment processes.

REFERENCES

EEC (1994). Commission Regulation No. 1488/94 of 28 June 1994 laying down the principles for the assessment of risks to man and the environment of existing substances. *Off. J. Eur. Communities* **37**(L 161), 3–11.

EEC (1993). Council Regulation No. 793/93 of 22 March 1993 on the evaluation and control of the risks of existing substances. *Off. J. Eur. Communities* **36**(L 84), 1–7.

EINECS (1981). European Inventory of Existing Commercial Chemical Substances (commercially available in the EEC between 01.01.71 until 18.09.81).

EU (1994). European Commission, Directorate—General Environment, Nuclear Safety and Civil Protection: Risk Assessment of Existing Substances, Technical Guidance Document (XI, 919/94-EN).

EU (1995). European Commission: Risk Assessment of New and Existing Substances; Technical Guidance Document Draft October.

ECB (1996). Address of Dr. W. Karcher, CEC Joint Research Centre, European Chemicals Bureau, T.P. 280, I-21020 ISPRA (Varese) Italy; IUCLID: Additional Features of Version 2.12 (25.01.96).

IPCS (1993). Meeting report on "International Co-ordination of criteria Document Production," Annex 5.

OECD (1994). Revised Draft SIDS Manual (OECD Secretariat) EXCH, Manual 9405 DOC July.