EUFRAM

Concerted action to develop a European Framework for probabilistic risk assessment of the environmental impacts of pesticides[1]

Work Package 11

PRELIMINARY PAPER ON POOLING DATA FOR PROBABILISTIC APPROACHES[2]

June 2003

Gerhard Görlitz (BCS-GmbH) [3], Andy Hart (CSL), Colin Brown, (CU), Mick Hamer (Syngenta)

1 CONTENTS

1 Contents 1
2 Introduction 2
3 General approach 2
4 Entities 2
4.1 Data required for probabilistic approaches 2
4.2 Categories how pooled data can be used 3
4.3 Data owners 3
4.4 Issues with data accessibility 4
4.5 Pooling options 5
4.6 Quality assurance 5
5 Workplan 6
6 Conclusion 7

[1] EUFRAM is coordinated by CSL (www.csl.gov.uk) and supported by European Commission's 5th Framework Programme (www.cordis.lu), contract number QLK5 - CT 2002 01346. Further information and news about EUFRAM is provided at www.eufram.com.

[2] EUFRAM partners welcome public comment and discussion on this document. Comments may be contributed via the EUFRAM website at www.eufram.com. This document is the responsibility of its publishers and in no way represents the views of the European Commission or its services.

[3] Work Package leader.

2 INTRODUCTION

One of the statements at the 2001 EUPRA workshop was: "probabilistic methods are data hungry" and as possible solutions were proposed:

a) the sharing of generic data

b) "Promote the use of unpublished proprietary data in the development of generic databases"

The remit for work package 11 can be considered as the result of these ideas:

This paper will identify the main type of existing data that might be useful for developing probabilistic methods and identify those types for which data access is difficult, e.g. data generated by industry for registration purposes, which could without prejudice to current pesticide authorisations be pooled for meta-analyses that are needed to support the implementation of probabilistic approaches. The paper will also propose a plan of action to investigate ways of pooling these data, including calendar for consultation with data owners and investigating the roles that might be played by organisations such as OECD, the EU Environment Agency and industry associations.

3 GENERAL APPROACH

The proposed general approach is first to collect the information on the data where accessibility is an issue, to identify the data owners and their interest and to explore which options for pooling exist.

This inventory will need to be discussed with the other workgroups (especially WP4, WP5, WP10) to check for completeness and technical suitability.

In a second step the different data owners and organisations shall be consulted and a framework developed, on which concrete proposals for individual data types / data bases can be developed.

4 ENTITIES

4.1 Data required for probabilistic approaches

Principally the data required for probabilistic risk assessment can – for the purpose of data access - be categorized as follows:

Generic data

geographical

biological/ecological

meteorological

agronomical

[more?]

Substance specific data (incl. metabolites)

Chemical

e-fate

ecotox & biological

Product specific data

Market data

Use data

4.2 Categories how pooled data can be used

1. Input data for characterising generic assessment scenarios

   (e.g. the use of soil, climate data etc to define probabilistic versions of the FOCUS scenarios). These data are required if we are to develop robust probabilistic methods for representing spatial and temporal variability.

2. Data to develop generic or surrogate estimates for substance or product specific data

   a. Estimation of pooled variances for SSDs (Luttik and Aldenberg 1996)

   b. QSARs to replace toxicity data

   c. acute/chronic ratios

   d. uncertainty or extrapolation factors for parent/metabolite toxicity

   These data may reduce the number of new studies required (e.g. by enabling estimation of an SSD with very few studies, or avoiding the need for testing metabolite toxicity),

3. Data for validation or testing of probabilistic assessments

   a. Measured concentrations to compare with PEC distributions,

   b. Data on impacts in mesocosms or field studies to compare with predicted impacts.

   This can help by building confidence in probabilistic methods (if predictions are confirmed), and enabling us to detect and correct model components that lead to poor predictions.

The nature of the benefits differs between these types of use. This will on one side influence the willingness of data owners to share, on the other side this will have a profound impact on the techniques that can be used for data pooling.

4.3 Data owners

In a similar way the owners of the data can be grouped by their interest into:

Research Institutions

a) publicly funded

b) privately funded

c) mixed funding (might today be the rule for public research institutions)

Public service institutions

(e.g. meteorological offices)

Other commercial enterprises

Agrochemical Industry

Government and supranational agencies (including regulatory bodies)

In this it is of special interest, to clarify precisely what the limitations are on the use that EU and memer state regulatory authorities can make of data submitted by companies – does this provide any scope for use in research to improve probabilistic approaches? Is this the same in all countries or does it depend on national law? In the USA, does the Freedom of Information Act provide scope for accessing regulatory data for research purposes?

4.4 Issues with data accessibility

Data accessibility can be restricted by several and interlinked factors: These can be categorized:

Legal

property rights

data protection

data confidentiality

licensing and copyright law

competition law

Commercial

cost of data generation

profit from data use

damage from data misuse

Technical

data formats (e.g. non digitized "paper" data)

Immaterial value (e.g. value of a database for a research institute as basis for future research projects)

In all data accessibility discussions it is very important , to consider that the value of the data to the data holder may far transcend the costs for their generation. This holds true for all commercial enterprises, for which the data either support their present business or are a source of future business. But this is also true for a research institute which intends to use a database as a basis for future research.

4.5 Pooling options

Depending on the type of data, the interest of the data owner and the purpose for which the data are to be used, different option might exist to resolve accessibility issue:

Cost sharing (e.g. acquisition or generation of data by an industry association for use by a group of companies)

Sharing of interest

> This solution might be favoured, when the commercial value of a data package is difficult to assess (e.g. 2 agrochemical companies exchanging regional soil, wheather and topographical or ecological data on a quid pro quo basis).

Technical solutions

> anonymization of data
>
> generation of meta data

For each of these options the applicability to the different data types and data owners needs to be analysed, determining the technical, practical and legal restrictions (e.g. competition law) that exist for them.

Some of the issues which arise from data sharing arrangements are given below:

- What would happen in a data-sharing arrangement if only some companies participated? Would other companies not be allowed to use the pooled data? Would they be charged a fee?
- How will a data-sharing arrangement cope with some partners having more data to share than others? Could it be assumed that those with more to share probably have more actives and therefore get more benefit? Or would one need some accounting of contributions and benefits?
- Would owners be more likely to contribute to a project with a defined plan for analysing the data and generating specific outputs? This could then be viewed like a joint research project than a data pooling exercise. For example, one could envisage a Focus-style project to develop probabilistic exposure scenarios, including data pooling but perhaps also data collection (e.g. satellite data). Another example might be a joint project to estimate extrapolation factors for metabolite toxicity, with companies contributing existing tox data for parents and their metabolites. There are funding programs that could contribute to the costs of such projects (e.g. LINK in the UK, and EU schemes).

4.6 Quality assurance

A very general issue is the quality assurance of pooled data. While it will in all cases be necessary, to establish confidence, that the pooling procedure did not introduce artefacts which have a significant impact on the use of the data and that the pooled data properly reflect the original data set and are not biased either due to a selection of the data used in the pooling procedure or due to the pooling procedure itself. This will be especially difficult in those cases, where pooling was choosen to maintain the confidentiality of the original data, since in these cases no retrospective check will be possible, and the suitability of the pooled data will have to be established a priori.

A few examples may illustrate the different kinds data may be biased due to a pooling procedure:

- Frequently monitoring studies are targeted towards problem areas and therefore give a biased representation of the full data.
- Some data holders may have an interest to include only data “favourable” to a certain position.
- The pooling procedure itself may introduce statistical artefacts, for example weather generators, which are calibrated with a limited set of real wheather data and then used to generate large volumes of simulated weather data for use in exposure models. While these algorithms give generally good estimates of the normal range of weather conditions, they are far less satisfying where the frequency and amplitude of extreme events is concerned. Such programs are therefore not recommended, where the extremes of the distribution curve are of prime interest (i.e. run-off events).

In all cases, where access to the original data is restricted, special care will have to be taken, that the procedures are properly documented and define the use areas for the data thus generated.

5 WORKPLAN

In order to resolve the above mentioned issues, the following stepwise systematical approach is proposed:

1) Inventory of required data (data types) and identification of accessibility issues

The inventory should comprise information on:

- Data required
- Purpose (see 3 main types of purpose described above)
- Volume and quality of the data required (In general, the more data the better. Quality is a more difficult question, and may be determined by the perceptions of regulators who may expect new regulatory tools to be based only on data of regulatory quality. This is a question investigated in DEMETRA (www.demetra-tox.net), an EU project on QSARs led by Emilio Benfenati of Institute Mario Negri with both industry and regulators
- Data owner(s)
- Critical issue with accessibility

There are very many potentially useful types of data and we don’t necessarily know yet exactly what is required, what quality etc., so it could be rather difficult to make a comprehensive inventory. Anyway, if we can identify some solutions for data access, hopefully they will be fairly generic and could apply to all types of data? If so, would it be enough to identify a few contrasting types of data that are a high priority, and use these as examples to explore the access issues?

2) Assessment of pooling techniques and their applicability to those data where accessibility is an issue. Might we benefit from consulting some company people with close involvement in legal and commercial issues already at this stage?

3) Inventory of the relevant data owners, grouped by the owners interest

4) Consultation of the data owners and of organisations (OECD, environment agencies, industry associations)

5) Work out of a proposal on data access and data pooling, identification of unresolved issues.

6 CONCLUSION

Probabilistic methods will most certainly increase the incentives for data pooling. Currently, assessments are deterministic and there is only limited incentive for reducing uncertainty. Probabilistic assessments up to now have mostly concentrated on quantifying variability (especially in exposure) as a means of relaxing worst-case assumptions – this favours sharing of generic data but perhaps not specific data. In future, probabilistic assessments may do more to quantify uncertainty (in both exposure and effects). In theory, this would provide a clear incentive for sharing all types of data, as reductions in uncertainty would be fully recognised. In praxis however, large obstacles are still to overcome, since on one side data owners have legitimate concerns regarding the proper sharing of costs and benefits, the protection of intellectual property and commercial interests on the other side the users of pooled data must be assured that these data are “fit for purpose” and not biased due to the underlying selection and pooling procedures.