Scientific Committee on Emerging and Newly Identified Health Risks

SCENIHR

Memorandum on the use of the scientific literature for human health risk assessment purposes – weighing of evidence and expression of uncertainty

The SCENIHR adopted this memorandum at its 17th plenary of 19 March 2012

About the Scientific Committees

Three independent non-food Scientific Committees provide the Commission with the scientific advice it needs when preparing policy and proposals relating to consumer safety, public health and the environment. The Committees also draw the Commission's attention to the new or emerging problems which may pose an actual or potential threat.

They are: the Scientific Committee on Consumer Safety (SCCS), the Scientific Committee on Health and Environmental Risks (SCHER) and the Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR) and are made up of external experts.

In addition, the Commission relies upon the work of the European Food Safety Authority (EFSA), the European Medicines Agency (EMA), the European Centre for Disease prevention and Control (ECDC) and the European Chemicals Agency (ECHA).

SCENIHR

This Committee deals with questions related to emerging or newly identified health and environmental risks and on broad, complex or multidisciplinary issues requiring a comprehensive assessment of risks to consumer safety or public health and related issues not covered by other Community risk assessment bodies. Examples of potential areas of activity include potential risks associated with interaction of risk factors, synergic effects, cumulative effects, antimicrobial resistance, new technologies such as nanotechnologies, medical devices including those incorporating substances of animal and/or human origin, tissue engineering, blood products, fertility reduction, cancer of endocrine organs, physical hazards such as noise and electromagnetic fields (from mobile phones, transmitters and electronically controlled home environments), and methodologies for assessing new risks. It may also be invited to address risks related to public health determinants and non-transmissible diseases.

Scientific Committee members

Anssi Auvinen, James Bridges, Kenneth Dawson, Wim De Jong, Philippe Hartemann, Peter Hoet, Thomas Jung, Mats-Olof Mattsson, Hannu Norppa, Jean-Mari.e. Pagès, Ana Proykova, Eduardo Rodríguez-Farré, Klaus Schulze-Osthoff, Joachim Schüz, Mogens Thomsen, Theo Vermeire

Contact:

European Commission
DG Health & Consumers
Directorate D: Health Systems and Products
Unit D3 - Risk Assessment
Office: B232 08/15 B-1049 Brussels

Sanco-Scenihr-Secretariat@ec.europa.eu

ISSN 1831-4783 ISBN 978-92-79-26315-6
doi:10.2772/86755 ND-AS-12-003-EN-N

ACKNOWLEDGMENTS

Members of the working group are acknowledged for their valuable contribution to this opinion. The members of the working group are:

SCENIHR members:

Jim Bridges

Anssi Auvinen

Mats-Olof Mattsson

Hannu Norppa

Joachim Schüz

Theo Vermeire

ABSTRACT

This memorandum is focussed on risk assessment of stressors to which humans may be exposed. The memorandum is intended to complement the draft SCENIHR report on the identification of emerging issues and the work of SCENIHR on the challenges in future risk assessments.

Currently the assessment of data relies on expert judgement and although this approach is well established, how the expert judgement is used, is often not clear to many stakeholders.

The memorandum addresses the following:

- Identification and selection of relevant publications for analysis,
- Weighing the data,
- Expression of uncertainty, and
- Application for risk assessment purposes.

The aim is to use it, wherever appropriate, for the risk assessment activities of the SCENIHR.

A risk assessment requires the evaluation of the evidence across all relevant domains/lines of evidence.

The approach proposed is a staged one involving:

- Individual data sets (e.g. publications),
- Individual lines of evidence,
- Combination of lines of evidence, and
- Characterisation of relevant uncertainties.

Individual papers/data sets that are identified initially but on preliminary examination do not meet the criteria of quality and/or relevance for the purposes of the development of the opinion will appear in the reference list or additional document for the report on which the opinion is based as 'Publications noted but not considered suitable for the purposes of developing the opinion'.

For each line of evidence the additional criteria of utility and comprehensiveness are introduced. This analysis leads to the assignment of individual papers to one of the following categories:

1. 'Publications that are relevant and of sufficient/suitable quality and were important for the development of the opinion'.
2. 'Publications that are relevant and of sufficient/suitable quality but were not judged to be necessary for the development of the opinion'.

Integration of the various lines of evidence (integrative risk assessment) is the final stage. The weighing of the total evidence should be presented for the purposes of clarity and consistency in a standard format. A tabulated form is proposed. Although all lines of evidence are considered for human risk assessment human, animal and mechanistic studies comprise the primary line of evidence along with exposure. The result of the tabulation and its analysis should be expressed for human risk assessment in terms of:

- Strong overall weight of evidence: Coherent evidence from human and one or more other lines of evidence (animal or mechanistic studies) in the absence of conflicting evidence from one of the other lines of evidence (no important data gaps).
- Moderate overall weight of evidence: good evidence from a primary line of evidence but evidence from several other lines is missing (important data gaps).

- Weak overall weight of evidence: weak evidence from the primary lines of evidence.

- Uncertain overall weight of evidence: due to conflicting information from different lines of evidence that cannot be explained in scientific terms.

- Weighing of evidence not possible.

- No suitable evidence available.

Characterisation of the uncertainties in the assessment needs to be specifically expressed in a form that is helpful to the stakeholders. A tiered approach for addressing uncertainties is recommended.

Keywords: Human health risk assessment, weight of evidence, uncertainty, scientific literature

Memorandum to be cited as: SCENIHR (Scientific Committee on Emerging and Newly Identified Health Risks), Memorandum on the use of the scientific literature for human health risk assessment purposes – weighing of evidence and expression of uncertainty, 19 March, 2012

TABLE OF CONTENTS

EXECUTIVE SUMMARY

SCENIHR is committed to ensuring a high quality in all its risk assessments. Such risk assessments inevitably involve a number of scientific judgements. The purpose of this memorandum is to provide transparency to stakeholders of the risk assessment process and to aid consistency between opinions. The focus of this memorandum is on risk assessment of stressors to which humans may be exposed. This memorandum is intended to complement the draft SCENIHR report on the identification of emerging issues and its work on the future challenges in risk assessment.

Currently the assessment of data relies on expert judgement and although this approach is well established, how the data is selected and used, and the consistency of this process between risk assessments, is not clear to many stakeholders.

This memorandum is concerned with the application of the weighing of evidence and the uncertainty involved for risk assessment purposes. It covers:

- the identification and selection of relevant publications for analysis,
- weighing the data,
- the expression of uncertainties, and
- the application for risk assessment purposes.

The aim is to use this memorandum, wherever appropriate, to guide the risk assessment activities of the SCENIHR.

A risk assessment requires the evaluation of the relevant, available evidence and involves a staged process beginning with individual publications and ending with the summation of all domains/lines of evidence.

Individual publications

It is proposed that the acceptability of each publication considered to be relevant should be specifically identified on the following criteria:

Quality

- *Good Scientific quality.* Study is considered to be appropriately designed, conducted and reported, and using valid methodology.
- *Adequate/utilizable scientific quality but with significant limitations*, i.e. Scientifically acceptable but some important deficiencies in the design and/or conduct and /or the reporting of the experimental findings
- *Inadequate scientific quality*, i.e. Serious concerns about the design or conduct of the study
- *Not assignable*, i.e. insufficient detail to make an evaluation.

Relevance

- *Direct relevance,* i.e. addressing the agent (stressor), model and outcome of interest
- *Indirect relevance*, i.e. addressing a related agent (stressor), model or outcome of interest
- *Insufficient relevance,* i.e. not useful for the specific risk assessment being conducted.

Typically, the data search methods used will identify many papers that could be used. A preliminary screening is then needed in order to focus on those relevant for the specific purposes of the development of the opinion. Papers that are identified initially but on preliminary examination do not meet the criteria of quality and/or relevance for the purposes of the development of the opinion should be cited in the reference list or additional document for the report on which the opinion is based as:

"Publications noted but not considered suitable for the purposes of developing the opinion".

Lines of evidence

The next stage is the assessment of the weight of evidence for each line of evidence.

For most risk assessments a number of lines of evidence need to be considered i.e.:

Exposure

- exposure measurements,
- mathematical modelling,
- toxicokinetics.

Hazard assessment

- epidemiology studies,
- human volunteer studies,
- other human data,
- animal studies,
- *in vitro* studies,
- *in silico* studies,
- mathematical modelling,
- mechanistic/mode of action studies.

Studies should be classified into those that:

- indicate the presence of an effect,
- indicate the absence of an effect,
- are consistent with either the presence or absence of an effect.

Two criteria are used:

Consistency is defined as the agreement on the outcome between different studies for each line of evidence. The following categories may be identified:

- High – most studies show findings in the same direction;
- Medium – the majority of studies are a mixture of findings in the same direction and consistent with either outcome;
- Low – little agreement between studies.

Utility is defined as usability for the purposes of developing the risk assessment. A matrix is proposed that integrates consideration of both relevance and quality. The following categories may be identified:

- High overall relevance,
- Moderate overall relevance,
- Low overall relevance.

This analysis leads to the assignment of individual papers to one of the following categories:

- 'Publications that are relevant and of sufficient/suitable quality and were important for the development of the opinion'
- 'Publications that are relevant and of sufficient/suitable quality but were not judged to be necessary for the development of the opinion'.

Integration of all lines of evidence

Integration of the various lines of evidence (integrative risk assessment) is the final stage It involves several steps, the details will vary according to the data available:

- Describe the nature of the data, including endpoints considered.
- Evaluation of exposure. Combining modelling and monitoring data.
- Evaluation of hazard data. Combining in vivo, *in vitro* and *in silico* data also combining animal/human data. This stage also includes dose-response and internal exposure modelling and extrapolation based on the critical study or studies.
- Mode(s) of action. The plausibility of the observed or hypothetical mode(s) of action and its validity for extrapolation purposes particularly between species.
- Quantifying the risks (including the statistical analysis) using the hazard and exposure data. Overall impact on man and on the environment**.**

At each stage, a narrative justification should be provided for the final conclusions. It should highlight possible knowledge gaps and other uncertainties.

Dimensions of risk that may need to be expressed include the severity of the effect/outcome (nature of the adverse effect) and the likelihood of its occurrence. Both of these aspects should be addressed in depth, where possible, during in risk characterisation.

The weighing of the total evidence should be presented, for the purposes of clarity and consistency in a standard format. A tabulated form is proposed. Although all lines of evidence are considered for human risk assessment human, animal and mechanistic studies comprise the primary lines of evidence along with exposure. The result of the tabulation and its analysis should be expressed as follows:

Strong overall weight of evidence:

Coherent evidence from human findings and one or more other lines of evidence (animal or mechanistic studies) in the absence of conflicting evidence from one of the other lines of evidence (no important data gaps).

Moderate overall weight of evidence

Good evidence from a primary line of evidence but evidence from several other lines is missing (important data gaps).

Weak overall weight of evidence*:*

The data itself is poor or insufficient.

Uncertain overall weight of evidence*:* due to conflicting information from different lines of evidence that cannot be explained adequately in scientific terms.

Weighing of evidence not possible

No suitable evidence available.

Expression of uncertainty

In principle, uncertainty analysis can be incorporated during the development of an opinion or added after the opinion is completed. Incorporation during development of a risk assessment should be the approach of choice wherever possible. The strategies for the two approaches are different. In the former a bottom up approach is used whereas in the latter a top down approach may be employed. The expression of uncertainty should not be more sophisticated than the expression of the risks, i.e. if the risk is expressed in a qualitative form so should the uncertainty. Conversely if the risk is expressed in numerical terms so should the uncertainty.

The focus of the expression of uncertainties should be on those aspects which could substantially impact on the risk assessment outcome and have not been allowed for through the use of default (uncertainty) factors**.**

1. PURPOSE OF THIS MEMORANDUM

The main aims of this memorandum are to provide greater transparency in the risk assessments carried out by the Scientific Committee, to be an aid to consistency between opinions and to be helpful to stakeholders. The memorandum draws on the methodology section of previous opinions of the SCENIHR.

It is intended to be used as:

i) guidance for risk assessments carried out by SCENIHR working groups and as the framework for a record of how SCENIHR has conducted individual risk assessments.

ii) a basis for subsequent published guidelines for the risk assessment work of SCENIHR and the other scientific committees.

iii) a contribution of SCENIHR to the developing global dialogue on risk assessment methodology.

It is anticipated that in the light of experience, involving a wide range of stressors and exposure situations the SCENIHR may wish to further develop the procedure set out below.

2. INTRODUCTION

To optimise their value, it is vital that risk assessments are accurate, transparent and readily utilisable. To this end the potential sources of error need to be recognised and addressed (see annex).

The assessment of the risk to human health requires the use of a number of distinct lines of evidence (in vivo, in vitro, in silico, population studies, modelled and measured exposure data etc.). The challenge for risk assessors is to utilise these types of evidence in a systematic and consistent way in order to arrive at an integrated assessment of the risk to human health from a particular agent(s)/stressor(s). To achieve a valid, scientifically defensible, assessment involves expert judgement. In many risk assessments, prepared by the scientific committees, it can be difficult for the stakeholders to follow how individual papers have been selected and weighed, how the findings have been integrated to reach the final conclusions and any uncertainties regarding the conclusions.

There are increasing demands to improve the transparency of risk assessments. This memorandum is concerned with improving the transparency on how papers are selected, evaluated, the conclusions reached and any uncertainties expressed. It draws on a number of previous publications on the subject, for example in regard to epidemiological studies the Bradford Hill criteria for assessing evidence. A similar approach has been used for eco-epidemiology. It has to be recognised that efforts to improve transparency through more detailed documentation will inevitably require additional time, with consequent resource implications.

Various bodies have used their own criteria for weighing of evidence to be applied in specific circumstances, for example IARC, IPPC, ICRP.

In some of its recent opinions SCENIHR has sought to explain the general procedures used for examining different types of evidence and in one opinion (on the risks to health from the use of energy saving light bulbs) it has used a scoring system to evaluate the human studies).

Linkov et al. (2009) have identified seven stages in the integration of data using a weight of evidence approach namely:

i) *Listing evidence* (Presentation of individual lines of evidence without attempting integration)

ii) *Best professional judgement* (qualitative integration of multiple lines of evidence)

iii) *Causal criteria* (a criterion based methodology for determining cause effect relationships)

iv) *Logic*. Standardised evaluation of individual lines of evidence based on qualitative logic methods)

v) *Scoring*. (Quantitative integration of multiple lines of evidence using simple weighting or ranking).

vi) *Indexing*. (integration of multiple lines of evidence into a single measurement based on empirical models)

vii) *Quantification*. (Integrated assessment using formal decision analysis and statistical methods.)

Currently most opinions of the scientific committees, including those of the SCENIHR, are in category i) or ii).

Critical issues in deciding on a future approach for the weighing of data are:

- While a common approach may be very suitable for the evaluation of standard data packages (which for a number of RA's is a legal requirement) can it also be applied to the work of committees, such as the SCENIHR, which deal with a broad diversity of stressors and with varying amount and level of evidence?
- Is it possible to use a standardised methodology and still retain any necessary flexibility?

It is probably not appropriate to jump from a level i) or ii) to a level vii) approach without gaining some experience of at least one intermediary level. The approach used in this opinion is aimed at a level v) -methodology that could evolve into a level vi) or vii) as more experience is gained.

This memorandum provides transparent criteria for:

- The identification of possible sources of relevant publications
- Weighing the data for its quality and relevance
- Introducing progressively, a formal scoring system for lines of evidence
- The integration of assessments of different lines of evidence
- The expression of uncertainty

Five stages may be identified for the weighing of data:

1. Identification of the possible sources of data and data gaps in relation to the aim of the assessment (section 3)
2. Initial screening of these data sources to identify those that are relevant to address the question(s) posed by the Commission Services (section 4)
3. Assessment of individual data sources (section 5)
4. Weighing of the individual lines of evidence (section 6)
5. Weighing of the totality of evidence (section 7)
6. Expression of uncertainty (section 8)

3. IDENTIFICATION OF POSSIBLE SOURCES OF DATA AND DATA GAPS

A critical aspect of the risk assessment process is the framing of the questions asked of the committee. This issue is addressed in a recent opinion (Improvements in Risk assessment 2012) and is not further considered here.

3.1. Selection of data to be analysed

It is very important that the Committee identifies very clearly throughout its work on a particular opinion the data it sought to access, how the data was accessed and any restrictions that were placed on the data access. It should be noted that often the Committee/ working group will only need to conduct a rather restricted literature search in order to answer the specific questions asked of it by the Commission services.

Initial electronic literature searches may be a starting point for data gathering. Where this is the case, the data bases and search engines used include for example PubMed, Toxline, Chemical and Biological Abstracts, and Google Scholar. In each opinion the search engines used should be identified along with the search terms used.

For each line of evidence (see section 6) the data sources will differ. It is very important that (as far as is practicable) all the relevant data is identified which is appropriate to providing the answer to the questions asked by the Commission. Inevitably, this is subject to practical constraints of accessibility, time available to reach the opinion and the language in which a publication appears.

Wherever possible, the Committee relies primarily on original refereed publications. This is obviously not always possible, e.g. in evaluating a product that has yet to be marketed.

According to the issue being addressed, the Committee may utilize one or more of the following:

- Reports and opinions of other scientific bodies. It should be noted that any references that are intended to be cited from these reviews must be cross checked before being included in the reference list;
- Meta-analysis and systematic reviews;
- Reports of various governmental and international bodies, e.g. WHO, FAO, JECFA, IARC, OECD, WMO, NIEHS;
- Reports of stakeholder bodies, e.g. ILSI, ECETOC, WWF;
- Stakeholder submissions such as confidential data provided by companies (see below);
- Non-refereed publications;
- Personal communications. Such communications can only be used if supported by raw data and details of the methodology used;
- Modelled data such as read across and exposure estimations.

In general, secondary sources, i.e. reports of the work of others (this is not intended to include meta-analysis), should only be considered if there is insufficient peer reviewed, published scientific data to provide an opinion.

3.2. Use of confidential data

For the purposes of the work of SCENIHR, confidential data should be used only when:

- The data will be made publically available in the near future;

- There is insufficient peer reviewed published data to answer the questions asked by the Commission services;
- It is a reputable source and the work is conducted to GLP or an appropriate indicator of quality. In order to accept such data the source of the data should be required to identify when the data was generated and to confirm that they know of no other data which would conflict with the information provided. Full details of methodology used need to be provided along with the opportunity to access the raw data;
- It should be made clear in any request for access to confidential reports that the data can only be considered if the provider agrees that the summary of the evaluation of the data by the WG may be incorporated in the text of the report (or as an appendix) and thereby made available to all stakeholders;
- The Commission services retain the confidential files, where these have been used to generate an opinion.

In respect of the latter point the protocol used by the German MAK committee is very appropriate, i.e. unpublished internal company data in the form of complete study reports is used when necessary. These are identified in the reference list at the end of the document. The validity of the information is checked. The unabridged reports are made available to the MAK Commission and are filed in the Commissions Scientific Office. Access to the company's reports is however not made available to third parties. Information required by a third party about the company's reports, cited in the commission documents, is supplied in writing by the chairman of the commission at their discretion. It is understood that a similar process is used by IPCS for its monographs.

4. INITIAL SCREENING OF DATA SOURCES

Prior to the screening of data sources it is important to consider all aspects of the risk(s) under consideration since incomplete identification of the risk(s) and/or risk factors may lead to an inappropriate literature search.

4.1. Issues that need to be considered

Typically, a substantial number of papers will be identified that are of possible interest. An initial screening process is needed to identify those that are suitable for the purposes of answering the questions. posed by the Commission Services

4.2. Criteria

There is no universally, formal and transparent procedure for the acceptability of data for risk assessment purposes. However the acceptability of a paper for the purposes of its use to answer a question(s) from the Commission Services can be based on the criteria proposed by Klimisch et al., (2007) and the OECD Manual for the investigation of HPV chemicals:

- *Relevance/potential importance*. This defines whether a set of data (e.g. a publication) has the potential to contribute to answering the questions asked by the Commission Services.
- *Quality/validity/reliability*. Quality is a general term covering the way the work has been conducted.

This covers:

- The suitability of the experimental design and the application of the methods and models,
- Whether or not the findings were reproducible between experiments (reliability). The assessment of reliability is among others based on the completeness and detail of reporting and referencing, whether peer review took place, and whether the authors worked under GLP/GCP or other audited schemes. In principle the more details of the methodology and results obtained are provided the greater the confidence in the publication's reliability.

Papers that are identified initially but do not meet the criteria of relevance, reliability and validity for the development of the opinion should appear in the reference list or additional document for the report on which the opinion is based as:

"Publications noted but not considered suitable for the purposes of developing the opinion".

5. ASSESSMENT OF INDIVIDUAL DATA SOURCES

A primary purpose of this memorandum is to identify a '*level of acceptability'*, i.e. the contribution of a publication to the knowledge base, for the purpose of developing an opinion. This 'level' is likely to vary according to the nature and extent of the evidence available. It should not necessarily be the case that for the purposes of a particular risk assessment scientifically significant data of questionable/low quality is completely disregarded. Rather the implications of the findings may be considered and addressed in the text while taking into account the increased uncertainty due to the questionable/poor quality.

A number of organisations have established their own frameworks for assessing/evaluating evidence. Wherever appropriate these have been drawn on in the development of this memorandum. Some are focussed principally on hazard e.g. IARC (see the preamble to the IARC Monograph Series IARC 2006). It is worth noting that the result of this process is not an assessment that a specific study is unequivocally negative or positive or whether it is accepted or rejected. Rather, the assessment will result in a weight that is given to the findings of a study. In this process, risk assessors need to assess uncertainties in the underlying data as well as in their own interpretations of these data (Levin *et al.,* 2004). Unfortunately, formal procedures and consistent terminology for weight of evidence processes are lacking. They are often claimed to have been applied, but without adequate documentation of the methods used (Levin, 2004; Weed, 2005; Hart *et al.*, 2007; Gee, 2008).

An important issue in weighing of evidence is the influence of values on expert judgment due to differences in ideological views. This needs to be recognized and made explicit as far as possible. It may introduce a further degree of subjectivity that is difficult to quantify but which can impact on the uncertainties in the risk assessment. Another important manifestation of uncertainty is the qualification of the data base (Van der Sluijs, 2003 and 2005).

5.1. Weighing of positive and negative findings

Epidemiological (field) and experimental studies should be subject to similar treatment in this evaluation process. Positive and negative studies should be evaluated using similar procedures and criteria and considered of similar importance if the quality is judged to be comparable. In positive studies the evaluation needs to consider both causal and non-causal explanations of the results. For example, a key question is, with what degree of certainty one can rule out the possibility that the observed positive result is produced by

bias, e.g. confounding or selection bias, or chance. In the case of negative studies it is necessary to assess the certainty with which it can be ruled out that the lack of an observed effect constitutes evidence against a hazard or whether it could result from (masking) bias, e.g., too small exposure contrasts, too crude exposure measurements, too small exposure groups/populations, or chance. Consideration should also be given to the possibility of publication bias i.e. that positive findings are more likely to be published than negative findings.

In risk assessment, it is preferable to get insight into the full distribution of risk, i.e. the balance between false positives and false negatives. Regulatory risk assessors, however, often claim to err on the side of caution, i.e. aim at a reduction of false negatives to avoid advice for accepting a harmful substance. This is not appropriate as it blurs the line between risk assessment and policy matters. If it occurs it needs to be documented and explained.

It is noted that the main direction of most errors in experimental animal studies in fact increases the chances of detecting a false negative: few dose levels, short exposure periods, low genetic variability, low statistical power, or use of a conservative significance level. This is normally countered by using high doses (Gee, 2008).

5.2. Statistical significance

It is a possibility that the lack of an observed effect is due to chance i.e. random variability hiding a possible effect due to inadequate statistical power (i.e. findings are compatible with both the study hypothesis and the null hypothesis). This is a particular problem in human studies with small sample sizes, low exposure levels and small risks. The presence or absence of statistical significance is only one factor in the evaluation. More importantly, other factors are the strength of the association (effect size) with its related statistical uncertainty (e.g. confidence intervals of the effect estimates) and the internal consistency of the results. This includes aspects such as comparability of effect across sub-groups, the form of the dose-response relationship and the methods used for the assessment of exposure, as well as biological or health endpoint, the relevance of any experimental biological model used and other aspects. Methods used in control of confounding including assessment of confounding factors, study design and statistical analysis are also a crucial aspect of study quality. Interpretation of the statistical significance also depends on the purpose of the study, use of predefined analysis plan, and number of hypotheses evaluated.

Regarding experimental studies, other important characteristics that are taken into consideration are the types of controls that have been used, blinding to assure comparability of information and to what degree replication studies have been performed.

Similarly, mere presence of a statistically significant difference between the groups does not alone constitute sufficient evidence for causality. Bias can produce a spurious association and in very large studies even non-relevant differences can reach statistical significance. In an exploratory (or a hypothesis screening) study, the number of tests will also need to be considered. Again, effect size and its uncertainty will primarily be evaluated, especially for epidemiological studies, and interpreted in respect to the possibility of upwards or downwards bias in the effect estimation.

5.3. Assessment of each publication/set of data

A major task of the committee/working group in conducting a risk assessment is to evaluate and assess the published articles and to judge their quality (validity, reliability) and relevance.

Key issues to be evaluated are:

- The characterization of the stressor.
- Soundness and appropriateness of the methodology used
- The reproducibility of findings between experiments
- The extent to which the full details of methodology are provided.
- The relevance of the set of data for a particular endpoint.

Quality and relevance need to be assessed independently. The following categorization applies:

Quality

- *Good Scientific quality.* The study is considered to be appropriately designed, conducted and reported, and to have used valid methodology.
- *Adequate/utilizable scientific quality but with significant limitations*. The study is scientifically acceptable but there are some important deficiencies in the design and/or conduct and /or the reporting of the experimental findings
- *Inadequate scientific quality*. There are serious concerns about the design and/or conduct of the study
- *Not assignable*. The study is lacking insufficient detail to make an evaluation

See Klimisch et al. (1997).

Relevance

- *Direct relevance,* The study addresses the specific agent (stressor), model and outcome of interest
- *Indirect relevance*, The study concerns a related agent (stressor), model or outcome of interest
- *Insufficient relevance.* The study cannot be used for the purposes of the risk assessment.

Table 1 Matrix to assess individual publications

	Good Scientific quality	**Adequate/utilizable scientific quality**	**Inadequate scientific quality**	**Not assignable**
Direct relevance	X	X		
Indirect relevance	X	X		
Insufficient relevance				

Those areas with a cross are preferably considered in the subsequent assessment. Based on the use of the above matrix, the papers identified as not useful should be listed in the opinion as:

'Publications noted but not considered suitable for the purposes of developing the opinion'.

6. WEIGHING OF THE INDIVIDUAL LINES OF EVIDENCE

Publications on the same topic/exposure estimate/ endpoint/effect may vary in their findings for a number of reasons including:

- Differences in the physical or chemical properties or purity of the agent being investigated.
- Variations in the study design, e.g. routes and form of exposure, sample size, dose selection, choice of endpoints, species, strain, experimental conditions, and statistical methods.
- Differences in the methods used to ascertain exposure, and/or toxicokinetics.
- Interlaboratory differences. Studies carried out according to the same guidelines with the same cell types, species, strain etc. in different laboratories may result in variations in results due to variations in experimental conditions, methods of analysis, statistical procedures and subjective judgements*, e.g.* in the evaluation of clinical and (histo)pathological effects.
- In epidemiological studies, comprehensive identification of the source population, non-selective recruitment/participation of the target population, validity and accuracy of the exposure assessment and outcome data, sufficient control of potential confounding factors, adequate statistical power and appropriate statistical methods are among the key considerations

Weighing evidence thus involves identification of the usefulness of the set of data on a particular endpoint in answering the question

6.1. Stressor identification and characterisation

There is the potential for a mismatch between the stressor that is identified and characterised and the stressor(s) to which humans and/or other species are exposed. This can arise because the stressor identified is a pure material while the form of exposure is modified, either because it is a component of a formulated product and /or because the nature of the stressor is modified as a result of environmental influences (adsorption onto particles, chemical or microbial transformations).

Important considerations therefore are:

- The stressor to which man/other species including environmental systems may be exposed is well defined as well as the identity of the stressor in the experiment.
- The proper characterisation of all relevant physical /chemical/biological properties of the stressor e.g. stability, volatility
- The identification of the matrix and any co-stressors, including impurities.

6.2. Exposure to the stressor

6.2.1. External exposure

Important considerations for exposure data are the nature, route and duration of exposure and whether co-exposure to other important stressors is likely. Key elements of exposure assessment are (IPCS, 2008):

- The exposure scenario: defined as a combination of facts, assumptions, and inferences that define a discrete situation where potential exposures may occur.

These may include the source, the exposed population, the time frame of exposure, microenvironment(s) and the activities of the exposed subjects. Under REACH (ECHA, 2008), an exposure scenario is more specifically defined as a set of information describing the conditions under which the risks associated with the identified use(s) of a substance can be controlled. It includes operational conditions and necessary risk management measures.

- The exposure model: a conceptual or mathematical representation of the exposure process.

- Sample collection and analytical methods.

The aim of the evaluation of exposure data is to conclude on the relevance and accuracy of the data and the integrity and transparency of the data collection and documentation. Key criteria for weighting an exposure assessment with regard to the relevance and accuracy of measured or modelled exposure are (based on IPCS 2008):

- Representativeness of the exposure scenario investigated for the situation at hand, *e.g.*:
  - Relevance of the route(s) of exposure for which hazard and exposure level data is available for the target population.
  - Relevance of the dose metrics for the type of effects being investigated.
  - Appropriate spatial and time scales.
  - If estimates of exposure were made by modelling, the validity, reliability, relevance of the model and method of extrapolation (see also 6.4.4).
- Suitability of the measuring device(s) and method employed. E.g. were the devices calibrated, did the method comply with GLP, was the mathematical method used appropriate?
- Nature of the sampling regimen e.g. continuous or intermittent, personal or area sampling, timing of assessment, selection of subjects/sites.
- Number of measurement data available.
- Availability of data on background exposure.
- Assumptions made in any form of data extrapolation or interpolation.
- Variability and uncertainty associated with the exposure data.
- In cases where biological monitoring is involved an additional criteria is the extent to which all the relevant metabolites were estimated.

Key criteria for weighting an exposure assessment with regard to the integrity and transparency of the data collection and documentation are (based on IPCS 2008):

- Quality assurance/quality control programmes in place.
- Collection, storage and analysis of samples for chemical analysis
- Controls on data entry and data transfer.
- Full documentation of study design, methods, model information, key determinants of exposure, findings, uncertainties and limitations.
- Clear rationales for interpretations made and conclusions drawn.

6.2.2. Internal exposure (toxicokinetics)

The first aspect to be considered is what aspect(s) of ADME (absorption, metabolism, distribution and excretion) are covered, whether the data is qualitative or quantitative and whether the stressor used and the dosages employed are relevant to either the animal hazard studies or to the likely exposure levels etc. of man and or environmental species.

Important criteria in evaluating such studies are:

- The suitability of the methodology for the purpose,
- Whether the data is relevant to critical target organs for toxicity,
- The quality control and/or reference stressors used,
- Assumptions made in any form of data extrapolation,
- If estimates of exposure were made the validity, reliability, relevance of the model /method of extrapolation.

6.3. Human studies

Only studies that meet relevant ethical criteria can be considered.

6.3.1. Epidemiologic studies

Epidemiology deals with the occurrence and distribution of diseases in populations. Most epidemiological studies are non-experimental i.e. exposure is not assigned by the researchers. This makes it very important to assess the potential for bias and confounding. Thus, critical judgment is needed to assess whether observed empirical exposure-disease associations are possibly causal, or more likely result from play of chance or methodological shortcomings/interpolation.

Making sense of results from epidemiological studies is particularly challenging when they are conflicting, or when there is a discrepancy between epidemiological findings and other domains of evidence.

Epidemiological studies range from descriptive studies and surveillance statistics to analytical studies and randomised trials. In evidence-based decision making, different study types contribute with different weights. More emphasis is given to results derived from prospective cohort studies followed by case-control studies, whereas firm conclusions can rarely be drawn from cross-sectional studies or ecological and descriptive studies, as they can provide only indirect, circumstantial evidence. Nevertheless, there is a considerable range of quality within study types. This applies especially to case-control studies, which are the most commonly used in investigations of risk of major chronic diseases such as cancer. Case-control studies are prone to selection bias (in case exposure distribution among cases or controls does not represent that of the target population, or cases and controls being derived from different source populations) and recall bias (lack of comparability of exposure information between cases and controls). As in all studies, comprehensive and detailed description of the study material and procedures is a necessary requirement for evaluating study quality.

Key features considered in evaluation of epidemiological evidence (adapted from the Bradford-Hill criteria) are:

- Temporality,
- Strength of the observed association,
- A dose-response pattern,
- Internal and external consistency of results,

- The specificity of the association,
- The absence of bias and control of confounding factors

Further opportunity for evaluation of whether the research results represent a genuine causal effect is provided by comparison of the findings of analytical studies with the trends in the population disease rates over time, particularly if analytical studies estimate a risk having a large population effect expected to alter the incidence rates.

Meta-analyses are a useful tool to numerically summarise the evidence, but if substantial heterogeneity is identified, a structured approach trying to clarify the source of such heterogeneity is more important than the calculation of pooled estimates. A good meta-analysis or review can be seen as a study of studies; hence, like original studies, they vary considerably in quality.

A common challenge is to distinguish between effects that are the direct result of the chemical acting on the body and psychological effects resulting from a perceived risk. Ill health (especially headache and tiredness) is very common. In many publications symptoms of ill health are ascribed to environmental stressors although there are other causes and/or contributory factors. A surprisingly high proportion of the population consider themselves to be particularly sensitive to chemicals.

The understanding of the relationship between the perception of risk and ill health (including mental health) is far from clear despite many publications on the subject.

6.3.2. Human volunteer studies

Human volunteer studies are used to evaluate whether effects can be observed during or shortly after exposure to an exposure (stressor). These studies are often termed provocation studies, as they are used to find out whether an agent will trigger (provoke) a certain effect, e.g. a measurable physiological response or symptoms. The quality of experimental studies on humans varies depending on their design and protocol. Human volunteer studies, as compared to epidemiological studies, have the advantage of providing better possibilities to control the exposure(s) under study, as well as possible confounding factors. On the other hand, the relevance of experimental laboratory studies to the real life situation may be less clear. For example, the absence in laboratory settings of contributing factors present in everyday life may influence the results and possibly reduce the chance to discover an effect. Moreover such studies are inevitably of short duration and extrapolation to long-term exposure may be problematic.

A double blind experimental laboratory study where subjects are randomly allocated to two or more exposure conditions is considered the strongest design to study acute effects. The goal is to create contrasting exposure under otherwise as similar conditions (and groups) as possible to discover possible effects. Subjects should be randomly allocated to the different exposure conditions. A cross-over design, where the same individuals are exposed to both (or several) conditions in a random order, is preferred. In the cross-over design, the subjects serve as their own controls in the comparisons between e.g. sham and the exposure under study. This approach avoids the possible error arising if two separate groups of participants are assigned either to sham or the exposure under study, and other possible differences between the groups than the exposure conditions may influence the results.

However, the cross-over design may be biased by carry-over effects if the time between the two (or more) conditions is not long enough for possible effects to wash out. If that is the case, a true effect may be hidden. Effects due to the order in which exposure conditions are applied may also obscure the results if the numbers of subjects that begin with the separate conditions are not balanced. For example, unfamiliar routines and environments may produce different reactions during the first experiment as compared to the later sessions. In order to prevent expectations of participants or researchers to

distort the results it is important that the study is performed double blinded, i.e. neither the researchers that lead the experiments nor the participants are aware of the true state of the exposure conditions during the study and with adequate allocation concealment so that there are no features in the experimental setting that would allow the subjects to identify the exposure status at a given time.

The choice of study group impacts the external validity of the study (generalizability of the results) because a very homogenous group (age, gender or symptom profile) may limit the population that the results are applicable to. On the other hand, a more heterogeneous study group may risk missing an effect present only in one or few sub-groups.

The outcomes that are assessed in a study may be more or less robust. If possible, objectively measured (e.g. heart rate, blood chemistry etc.) data are desired. Self-reported effects are more difficult to assess. The choice of scales for self-reported effects or interpretations of open questions may also influence the results.

Criteria to be considering for human volunteer studies:

- selection of the volunteers and the extent to which they may be considered as representative,
- relevance of the exposure conditions (including duration),
- degree of control over confounding factors,
- type of effects investigated and the degree of objectivity involved,
- Assumptions made in any form of data extrapolation.

6.3.3. Biomarker studies in humans

Biomarker studies in humans are used to evaluate whether biological indicators of a certain type of exposure (biomarkers of exposure) or an effect associated with an exposure such as an increased risk of a disease (biomarkers of effect) can be observed at an altered level or type in subjects who have been exposed to a stressor in the workplace or in other environments. These studies may also involve phenotypic or genotypic features that are linked with susceptibility to an exposure or to an increased risk of disease (biomarkers of susceptibility).

Biomarker studies typically utilize biological samples - urine, blood or its components, tissues, or cells which are collected from the exposed subjects and analyzed for the biomarker. As with human experimental studies, the quality of human biomarker studies varies depending on design and protocol. Biomarker studies provide possibilities to correlate the biomarker with external exposure measure and to control factors that could confound or modify the outcome. While conditions cannot be as well controlled as in laboratory studies, biomarker studies represent the real life situation. The studies are usually cross-sectional, but may involve repeated samplings, follow-up, and re-analyses after intervention. Depending on the exposure and endpoint studied, the findings may depict events that occurred recently or that accumulated during a longer time.

Biomarker studies often include one or several exposure groups and one or several control groups that are not exposed to the agent under study (or other agents with similar outcome) or are exposed at a clearly lower level. Otherwise, the control group(s) is (are) as similar as possible to the exposure group(s) as concerns sex, age, and social group and other variables that may affect the biomarker. These variables may include lifestyle factors, other exposures and lifestyle factors. Often the control subjects are matched pair-wise with the exposed for the critical variables. If the biomarker is a specific indicator that is not observed in unexposed subjects or whose level in the unexposed population is not overlapping with that found in the exposed, control groups are not always utilized; this is the situation with established methods of biological

monitoring where biological guidance values are available. Control groups are compulsory with less specific or less explored endpoints where the distributions of the exposed (or highly exposed) and unexposed (or less exposed) are overlapping or poorly known. Control group(s) should be described in similar detail as the exposed group(s). Depending on the endpoint and interval between sampling, the same person may serve as his/her own control before exposure or after a period of no/low exposure

To avoid possible bias due to researchers/analysts knowing the source of the samples, it is important that the study is performed in a blind manner, so that the staff involved does not know which study group each sample represents. This is usually achieved by coding the samples at the earliest possible phase of the study in a way that exposure status of the samples cannot be identified.

Criteria to be considered in human biomarker studies:

- numbers of exposed and control subjects
- matching of the exposed and controls
- information on exposure type, levels, duration and history
- adequate description of the control group(s)
- degree of control over confounding and modifying factors
- type of biomarkers investigated and the degree of objectivity involved
- specificity of biomarkers with respect to an exposure or effect such as disease
- relevance and correctness of analysis methods
- relationship between microflora and specific disease predisposition

6.3.4. Clinical studies

The concept of hierarchy of evidence has been developed as part of the evidence-based medicine approach. It classifies various study designs in accordance with the applicability of the results for clinical decision making, primarily those concerning effectiveness of treatment. It is therefore not strictly a weight of evidence methodology nor a strength of evidence classification, but is structured based on relevance for clinical medicine. The methodological quality or validity indicates how robust and generalizable results are likely to be. Sometimes hierarchy of evidence is used to categorise study designs, but often it also encompasses the findings of the studies.

A pyramid showing the relative validity of various methods for obtaining evidence usually has clinical experience as the bottom level, followed by case reports and case series and culminates in randomised trials (or meta-analyses and systematic reviews of randomised trials). Laboratory studies and expert opinion are also listed as low-level evidence in some formulations. Sometimes case-control studies are included, though they are not readily suited for evaluating the effects of treatment (but used sometimes in evaluation of e.g. effects of screening). One should also note that the quality of a study depends not only on the chosen study format, but a number of decision regarding e.g. sample size and other methods and approaches such as randomisation procedure, blinding, allocation concealment, use of placebo, and choice of end-point are crucial for determining the quality of the evidence produced.

The term ‘level of evidence’ is used when the findings of the studies are also incorporated. Again, there are various classifications and criteria. Typically, the strongest knowledge base is characterised as several randomised trials showing consistent results with a narrow confidence interval on a patient-relevant outcome (possibly with a meta-analysis to quantify a pooled effect). Less firmly established knowledge can be based on several high-quality trials with partly conflicting results or on an end-point without direct

counterpart in patient outcome (e.g. biochemical or radiologic indicator of disease process) US PSTF, UK NHS, GRADE, (Guyatt et al 1995, Cook et al 1992).

The criteria for 'level of evidence' include:

- study design: well conducted randomised trials providing most valid results
- outcome: clinical relevance of the main endpoint, should be similar to the treatment goals in patient care
- Effect size: the benefit or adverse effect evaluated should be sufficiently large to be meaningful, absolute effects expressed as number needed to treat/harm.
- consistency of results: agreement between studies, meta-analysis showing no substantial heterogeneity or publication bias

6.3.5. Other human data sources

For perceived health and self-reported health (soft end-points), a common challenge is to distinguish between effects that are the direct result of the physiological effects of the agent and psychological effects resulting from a perceived risk. Ill health (especially headache and tiredness) is very common. In many publications symptoms of ill health are ascribed to environmental stressors although there are other causes and/or contributory factors. A surprisingly high proportion of the population consider themselves to be particularly sensitive to chemicals. Also, changes in practices such as diagnostic criteria, disease classification, or compensation of occupational diseases can substantially affect disease rates even for apparently objective (hard).

A classification of evidence has been developed by the Cochrane Collaboration primarily related to therapeutic interventions. It emphasizes the importance of study design and stresses the importance of randomised trials, but also end-points relevant for patient outcomes (clinically important indicators of effect). Preliminary evidence mainly sufficient to justify further research with more rigorous methods can be obtained from studies lacking a proper comparative design (contrasting an outcome between groups with different exposures). Such human data unsuitable for risk assessment purposes can be obtained from the following:

-Observations by a health professional in the relevant area e.g. cases series, case reports on individuals;

-Experiences by individuals reported by others;

-Experiences described by individuals but not confirmed by others.

Criteria that may be used in judging the utility of these are:

- Have the same effects been reported in other human studies?
- Are the effects described only self-reported symptoms or is there additional evidence for the effects?

6.4. Hazard assessment

6.4.1. Animal studies

Usually, laboratory strains of mice or rats are used. The advantage of animal studies is that they provide information about effects on a whole living organism that displays the full repertoire of body structures and functions, such as nervous system, endocrine system, and immune responses. In this respect, animal studies are usually a more

powerful experimental tool than cellular studies for assessing health risks to humans. If animal studies are to be used to anticipate potential effects in man then extrapolation of the date is needed. Too often this is simply done on the basis of applying an arbitrary conservative, default factor. This blurs the line between science and policy since it is not the role of a risk assessment to apply conservatism. Rather this should be done by the risk managers based on the uncertainties in the assessment and any additional precautionary measures they deem to be needed.

In the scientific extrapolation of animal findings to man, attention needs to be paid to obvious differences in, e.g., body mass, life expectancy, physiology and metabolism between species. Rodent carcinogenicity studies, for example, have been criticised because many agents that are carcinogenic to rodents (often only at very high doses) are not carcinogenic to humans, and some human carcinogens do not affect rodents in standard carcinogenicity tests. Extrapolation from animal experiments to humans should always include consideration of the validity of the animal model used – good animal models do not exist at present for all human diseases. Nevertheless, at a molecular level, many basic processes, such as DNA damage and repair, are similar in animals and humans, and animal studies have remained a cornerstone in evaluating toxicity of chemical and physical agents. In the evaluations of IARC, for example, agents for which there is sufficient evidence of carcinogenicity in animals are considered to pose carcinogenic hazard to humans, unless there is scientific evidence that the agent causes cancer through a species-specific mechanism that does not operate in humans (IARC 2006).

Genotoxicity assays in experimental animals are usually applied after in vitro studies to see if a positive effect seen in vitro could be ascertained in vivo. They can also be used to check the correctness of negative results obtained in vitro, especially if it is suspected that in vitro conditions may not have been able to detect the activity, e.g. due to a lack of a crucial metabolic route. In vivo results are considered to have more relevance than in vitro results in the overall assessment of a genotoxic hazard. The in vivo micronucleus assay is most often applied, and it may be complemented by the liver unscheduled DNA synthesis (UDS) assay. Both assays belong to the OECD test battery, and there is much experience especially on the in vivo micronucleus assay. The in vivo micronucleus assay detects genotoxicants (clastogens and aneugens) that have a systemic genotoxic effect or that target on the bone marrow. Genotoxicants with a local effect e.g. at the site of first contact or in another target organ than bone marrow may not be detected. The liver UDS test detects DNA repair synthesis and is usually considered rather insensitive. These assays are normally performed after short-term exposure, and may not detect agents that become genotoxic only after prolonged exposure.

Other tissues and genotoxicity endpoints may be examined, e.g., by the comet assay (measuring DNA damage) and gene mutation assays using transgenic animals (neither of these is presently included in OECD test battery). The comet assay is relatively simple to perform on various tissues but it depicts transient DNA damage and may easily give positive results. Transgenic animal assays reflect true gene mutations in a transgene, but are relatively expensive. The size of the transgene may limit the size of deletions that can be seen. Micronucleus assays have also been described for other tissues than bone marrow, but there is presently little information on their performance.

Criteria for evaluating individual animal studies include the following questions:

- Was the number of animals per group adequate?
- Were animals of both sexes used (if relevant)?
- Were animals randomly allocated to groups?
- Was sham exposure used with similar procedures as for the experimental group but without the active agent?

- Was a positive control exposure included and did it give the correct result?
- Were exposure levels and treatment durations appropriate and fully characterised?
- Were the exposure system and measurement of exposure adequate?
- Was the duration of observation adequate with respect to the health endpoint addressed (for example, lifetime observation in carcinogenicity studies)?
- Apart from the exposure of interest, was treatment of exposed and control groups identical? Was there possibility of bias related to differences in survival between groups?
- Was blinding of samples used during the administration and evaluation of data?
- Was the endpoint measured adequately?
- Did the test agent reach the target organ or have toxic effects there (e.g. bone marrow in the in vivo micronucleus assay)?
- Were data reported adequately?
- Was a dose-response relationship observed?
- Availability of historic data on the occurrence of the adverse effects of interest in the animal strain studies (important for chronic studies).
- Were the methods of analysis performed correctly?
- Was a statistical analysis performed and was it done correctly?
- Were assumptions made in any form of data extrapolation?

With some variations these criteria can also be applied to species used in ecotoxicological studies.

6.4.2. *In vitro* studies

In vitro studies (using animal or human tissues) investigate toxicological, mechanistic and other relevant effects, which can provide evidence for and possible understanding of the development of cancer and other diseases. In vitro assays can show potential effects of various agents on a wide variety of biological endpoints in a manner which is rapid and cost-effective.

Genotoxicity studies include assays showing the interaction of the possible risk factor with DNA, the mitotic apparatus, or other cellular targets that can result in DNA damage, gene mutations, or structural or numerical chromosomal alterations. These assays are used for revealing carcinogenic potential that is based on genotoxic mechanisms, but do not identify non-genotoxic carcinogens. A battery of techniques is available for this purpose. Ideally, the used methods should confirm or compensate each other. For genotoxic agents capable of inducing various types of genotoxic damage, positive findings can be shown by using different techniques (see below). For genotoxicants with a narrow mode of action, the effect may be restricted to (or preferentially expressed as) gene mutations (gene mutagens), chromosome aberrations (clastogens), or numerical chromosome alterations (aneugens). In general, the reproducibility of positive findings and negative has to be shown by independent laboratories.

Also other *in vitro* toxicity studies usually aim at mechanistic understanding by using a wide variety of endpoints. This can elucidate the machinery of action on the cellular level which can also be predictive to a certain extent for some hazardous effects.

In vitro studies contribute to acute toxicity testing and can provide information relevant regarding carcinogenesis and other physiological or pathological processes but cannot

replace *in vivo* conditions or long term exposure conditions. Therefore information about e.g. genotoxic capacity can only be indicative of a potentially serious public health risk.

Criteria for evaluating in vitro studies include:

- Were there suitable positive and negative controls?
- Was the cell type used relevant?
- Was metabolic activation available (when relevant for the test agent)?
- Was the dose dependency of the effect investigated?
- Were the levels used appropriate considering the assay used (e.g., was dosing extended up to toxic levels in genotoxicity assays)?
- How did the levels used compare with those likely to be experienced by man?
- Were the exposure levels maintained throughout the test?
- Was blinding of samples used during the administration and evaluation of data?
- Was a threshold for the effect observed?
- Were the methods of analysis performed correctly?
- Was a statistical analysis performed and was it done correctly?
- Were the findings statistically significant?
- Assumptions made in any form of data extrapolation

6.5. Mathematical models, structure activity and other *in silico* data

There is likely to be a continual increase in modelling and QSAR data available for risk assessment purposes. The evaluation of such data again relate to validity, reliability and relevance and transparency. Criteria that may be applied in the weighting process include (based on Nendza *et al*., 2010 and OECD, 2007):

With regard to validity:

- Has the model been verified or validated for issues similar to those relating to the commissions question(s)?
- Does the validation follow the OECD principles (defined endpoint, unambiguous algorithm, defined domain of applicability, described with sufficient statistical characteristics, mechanistic interpretation)?
- Is the training set of high quality?
- Is the stressor within the domain of the model?
- Have appropriate measures of goodness-of-fit, robustness and productivity been applied?
- Is the model validation adequately described, e.g. by using QMRFs ((Q)SAR Model Reporting Formats?

With regard to reliability:

- Has the model been utilised appropriately?
- Was the statistical evaluation method adequate and performed correctly?

With regard to relevance:

- Suitability of the model for the situation at hand, e.g. with regard to species or population, route of exposure, endpoint considered, environmental conditions.

With regard to transparency:

- Is the model adequately described/referenced, e.g. by using QPRFs ((Q)SAR Prediction Reporting Format)?
- Assumptions made in any form of extrapolation or interpolation.

In the absence of data, grouping and read across methods may also be applied. These methods are based on grouping of chemicals on structural and/or mechanistic ground. The robustness of a chemical category can be evaluated on the basis of the following considerations (ECHA Guidance R.6, 2008):

- The membership of the category characterised by the number of members in a category and the available data
- The density and distribution of the category both in terms of the chemicals present and the available data
- The quality of the underlying experimental data for each of the endpoints covered
- The presumed mechanistic basis underpinning the category for a particular point
- The quality of the data estimated by the external computational approaches.

6.6. Studies on Modes/Mechanisms of action

IPCS and ECETOC have published guidelines on the assessment of MoA studies for both carcinogens and non-carcinogens (Boobis et al 2006, Boobis et al 2008). The main criteria used can be summarised as:

- Is sufficient data available to make a judgement?
- Have likely key events been identified?
- Is the proposed MoA biologically plausible? For example are there parallels for the same or a similar MoA?
- Is there good concordance between the various published MoA studies?
- Have alternative hypotheses been properly considered?
- Has the MoA been shown in the species and organ(s) in which the adverse effects have been shown to occur?
- Are the exposure conditions comparable?
- Is there a reasonable scientific basis for extrapolation of the MoA to other species and or other affected organs?

6.7. Omics

Omics is not much used in risk assessment at the present time but it is anticipated to make an increasingly important contribution both to hazard and exposure assessment.

In human hazard assessment a particularly important role is in the elucidation of modes of adverse action. Omics is not anticipated to make a significant contribution to environmental risk assessment. Outline criteria for evaluating omics data include:

- Was the biological test system and the exposure conditions used relevant to the question under consideration?
- Was the methodology used for obtaining the omics data well described and weel, conducted?

- Can the omics changes observed be linked with confidence to identifiable health effects?
- Are any extrapolations of the data justified?

6.8. Developing methods and their assessment

A number of other promising techniques are emerging that are likely in the future to make a significant contribution to aspects of risk assessment (see SCENIHR opinion on future challenges in risk assessment).

Since their precise contribution to risk assessment cannot be evaluated at the current state of their development it is not appropriate to set out clear guidelines on how the data generated by them should be evaluated or weighed against other data sources. Instead a case by case approach is required. Nonetheless there are some general parameters that may be considered:

- Were the exposure conditions to the chemical relevant (e.g. route, dose, duration)?
- Was the biological system involved relevant?
- Was the data reproducible in more than one test run?
- Is the methodology clearly described and does it follow established protocols?
- Where appropriate controls run?

6.9. Use of ecotoxicicity data for human risk assessment purposes

This approach which has been described as integrated risk assessment has limited application at present. It use may become clarified as a result of the framework 7 project named Heroic. It is not appropriate at the present time to propose how such data should be assessed apart from the principles set out in this memorandum.

6.10 Scoring system for individual lines/domains of evidence

The same criteria should be applied as in the assessment of individual publications (see section 5.3). However, two additional parameters need to be addressed:

- Utility, and
- Consistency.

6.10.1 Utility

Utility combines quality and relevance. Tabulation is a useful means of characterising utility into: high utility, medium utility and low utility.

Table 2 The proportion of each publication in each box and utility ranking.

	Good Scientific quality	***Adequate/utilizable scientific quality***
Direct relevance	high	medium
Indirect relevance	medium	low

It is important in the tabulation to consider whether differentiation is needed according to study size, exposure levels, and number of endpoints.

6.10.2 Consistency

Studies may be classified into those that:

- Indicate the presence of an effect;
- Indicate the absence of an effect;
- Are consistent with either the presence or absence of an effect.

The assessment of consistency is only meaningful for comparable studies.

Consistency is defined as the agreement in the results of the analysis between all the individual publications/data sets. Different studies can be classified as:

- **HIGH** – most studies show findings in the same direction;
- **MEDIUM** – the studies show a mixture of findings in the same direction and those consistent with either outcome;
- **LOW** – little agreement between studies. This may be due to heterogeneity of results because of particular features of the studies considered or to effect modification, e.g. because of the presence of susceptible subgroups in the study.

The effects of a given exposure to an agent/stressor may vary between sub-groups within a population, which is called susceptibility, vulnerability or sensitivity (indicating a higher risk at a given exposure level). In epidemiology, this phenomenon is called heterogeneity of effect (effect modification). Determinants of response may be biological (e.g. genetic), medical (co-morbidity), or social and behavioural (lifestyle, working and living conditions). Joint effects with other agents can also be considered as a determinant of susceptibility.

Subgroup analyses are commonly carried out in a variety of studies, but their interpretation requires caution in particular if they are not pre-specified in the study protocol (but carried out ad hoc), or based on adequate statistical power. As for data analysis, a significant effect in a subgroup is not sufficient evidence for demonstrating the existence of a sensitive subgroup, but an interaction term (indicating the joint effect of the exposure and the modifying factor) should be evaluated to properly assess effect modification (as a secondary hypothesis). In the event of negative overall result, there is a temptation to explore the data and a larger effect in a subset of the data is sometimes even reported as the main finding, which is prone to type I error i.e. false positive result due to selective reporting. In general, consistency of results across sub-groups increases the credibility of the findings and a well justified hypothesis and/or similar results from several studies are required for establishing a (identification of a susceptible sub-group).

The following table should also be produced for each line of evidence:

Table 3 Matrix to weigh individual lines of evidence (scoring indicated by crosses)

		Consistency		
		High	***Medium***	***Low***
Utility	***High***			
	Medium			
	Low			

6.11 Citing papers examined

As a consequence of in-depth evaluation, publications, and any other sources of data used will be cited in the reference list in the opinion or in a separate document in one of three categories:

1. *'Publications that are relevant and of sufficient/suitable quality and were important for the development of the opinion'*
2. *'Publications that are relevant and of sufficient/suitable quality but were not judged to be necessary for the development of the opinion'*
3. *'Publications noted but not considered adequate (relevant or of sufficient quality) for the purposes of developing the opinion'; this group might be listed in an annex*

6.12 Identification of critical data gaps

Data gaps appear when the risk assessor cannot come to a firm conclusion on one individual line of evidence because:

- data are inconsistent without a valid explanation,
- the available data are consistent but highly uncertain (low utility due to low relevance and/or low quality),
- data are lacking and not fulfilling regulatory or scientific requirements.

Data gaps can only be identified after all available testing, non-testing and exposure information have been considered. The data gap identified will be carried forward to the final assessment (section 7). At that stage it will be decided whether the data gaps prevent an overall conclusion on the risk. It is noted that data gaps may arise because:

- The data is old and derived using techniques that are no longer considered suitable for the purpose
- It is based on exposure patterns that are no longer valid

6.13 Noting outliers and genuine data variability

For each line of evidence it is important to identify studies that appear to have been well conducted but generate findings that are very different (outliers) from those of other studies in the same line of evidence. Differences between apparently very similar, good quality studies (genuine variability) also need to be addressed in the final risk

assessment along with other issues such as data gaps and comments on possible unknown unknowns.

7. WEIGHING THE TOTALITY OF EVIDENCE

7.1. General approach

Integrative risk assessment means that the results from all relevant individual studies are compiled into an overall assessment. In respect of the assessment of human health risk priority is given to studies on health outcomes where good quality human data is available. The step that follows the evaluation of the individual studies within a particular area (e.g. the properties of a particular chemical of interest) is the assessment of the overall evidence for a given outcome. This involves several steps:

- Check that all the relevant available data, including endpoints has been considered

- Reaching conclusions on exposure. Combining modelling and monitoring data, Identification of critical data gaps.

- Reaching conclusions on hazard data. Combining in vivo, *in vitro* and *in silico* data also combining animal and human data. This stage also includes dose-response and internal exposure modelling and extrapolation based on the critical study or studies. Identification of critical data gaps.

- Deciding on the applicability of mode(s)/mechanisms of action. The plausibility of the observed or hypothetical mode(s) of action and its validity for extrapolation purposes particularly between species.

- Quantifying the risks (including the statistical analysis) using the hazard and exposure data. Overall impact on man and on the environment**.**

At each stage, a narrative justification should be provided for the final conclusions. It should highlight possible knowledge gaps and other uncertainties.

7.2. Conclusions on exposure

Exposure assessment is commonly the weakest point in risk assessment. It is crucial to first identify the extent to which the overall exposure findings relate to those covered by the question. If this is not entirely the case, the confidence in the extrapolation needs to be specified.

The *aims* of exposure assessment include:

- Identifying exposure sources and activities resulting in exposure (or opportunity for exposure);
- Estimating the exposure levels and frequency;
- Defining the key populations at risk (e.g. those with the highest exposure or special vulnerable groups such as pregnant women);
- Identifying critical data gaps.

A number of technical questions need to be addressed to reach scientifically valid conclusions on exposure based on the available data (IPCS, 2008):

- Does the exposure data capture the important exposure pathways and routes, and does it quantify these reliably?

- Are the measurement data and/or the modeling estimates relevant to the target group or ecosystem — either the general population or a selected sensitive subgroup

- Is there sufficient measurement/modeling data to represent vulnerable populations, such as children or the elderly, or an ecosystem?

- What is the scientific basis for any extrapolation from a relatively small sample to the larger group?

- Do the exposure assessments describe the variability and uncertainty associated with the exposure scenarios or processes? Are these expressed in a form that users of the risk assessment understand any limitations of exposures and risk estimates?

In principle, measured exposure data, representative of the situation to be assessed should be given a higher weighting than modelled data. However this requires that they are representative of the exposure scenario and have been adequately measured (Van de Meent and De Bruijn, 2007). Empirical, measurement data often have deficiencies and modelling or extrapolation may be needed to reach sound conclusions on the external exposure. Therefore, comparison of the estimated and measured concentrations in order to select the "right" data for use in the risk-characterization phase may be required. This *comparison* should be done in a systematic way, recognising the following:

- The appropriateness of the exposure data depend on analytical techniques and time scale of measurements (e.g. spot measurements or long-term monitoring).
- The sampling, processing and detection techniques have to be evaluated in the light of the physicochemical properties of the chemical.
- In correlating these data to the appropriate emission and modelling scenarios. The measured data must be allocated to a certain spatial scale in order to be able to compare specific modelling scenarios.
- The need to compare representative measurement data with corresponding estimations and undertaking a critical analysis of the differences between the two.
- Characterisation of the genuine data variability and uncertainties involved in the above. The assessment of variability and uncertainty in exposure assessments needs to include the identification of sources and nature of uncertainty (adapted from IPCS 2008) i.e.:

  - Scenario variability and uncertainty: sources of release and the chemicals considered, exposure pathways, exposure events, exposure routes, exposed populations and ecosystems, spatial and temporal information, microenvironments, population activities, environmental variability, potential risk management options.

  - Model uncertainty: sources are the link between conceptual model and adopted scenario, model dependencies, model assumptions, model detail, model extrapolation, model implementation

  - Parameter uncertainty: sources are measurement errors, sample uncertainty, data type (e.g. surrogate data, expert judgement, defaults, modelling data, and measurement data), extrapolation uncertainty, uncertainty in statistical distribution.

Exposure conclusions may need to consider exposure to one chemical involving one or more routes of exposure or exposure to multiple chemicals via one or more routes (IPCS, 2009). Combined exposure to multiple chemicals can be evaluated in the context of whether or not the components act by similar or different modes of action (i.e. "single mode of action" or "multiple modes of action"). Chemicals that act by the same mode of action and/or at the same target cell or tissue often act in a potency-corrected "dose additive" manner. Where chemicals act independently, by discrete modes of action or at different target cells or tissues, the effects may be additive ("effects additive" or "response additive"). Alternatively, chemicals may interact to produce an effect, such that their combined effect "departs from dose additivity". Such departures comprise synergy, where the effect is greater than that predicted on the basis of additivity, and antagonism, where the effect is less than that predicted on the basis of additivity. *Relevant questions* for the assessment of combined exposure are:

- What is the nature of exposure? Are the key components known? Are there data available on the hazard of the mixture itself?

- Is exposure unlikely or very low, taking into account the context?

- Is there a likelihood of co-exposure within a relevant time-frame?

- What is the rationale for grouping particular chemicals in a specific assessment group?

7.3. Conclusions on the hazard data

The data base for hazard assessment is anticipated to change considerably over time for a number of reasons. In the light of current knowledge, for human risk assessments, for existing chemicals, in principle, human evidence merits a higher weight than animal studies which are preferable to *in vitro* and *in silico* data. However this assumes that the human data covers the range of exposure situations covered by the question(s). In practice it is usually necessary to use all domains of evidence. IPCS have recently produced a report setting out a strategy for the use of human and animal data.

Risk assessment integrates evidence from human population studies together with mechanistic, cellular and laboratory studies covering animal, cellular and genetic research, to assess disturbances of physiological processes related to reproduction and development as well as cancer and other major diseases. Mechanistic, cellular and laboratory studies are part of the overall criteria used to determine causality in interpreting epidemiological studies.

Animal studies more and more will be supplemented or replaced by alternative approaches such as (Q)SARs, read-across, *in vitro* data, exposure-based waiving, supplemented by MoA-information.

The *aims* of the hazard assessment include:

- to utilise data from all relevant lines of evidence
- to check that in whole or in part the hazard studies reflect the actual/likely human exposure situations
- to consider whether non-human derived data can be legitimately extrapolated to humans (e.g. is there valid mode of action data)
- To assess whether the hazard studies allow an indication of potential susceptible groups of the population.
- To distinguish between genuine variability and uncertainties relating to methodological aspects etc.

A number of technical questions need to be addressed to reach scientifically valid conclusions on the hazard based available data:

- Are the studies properly validated e.g. performed to GLP, GCP etc.?

-Is the exposure route used in the hazard testing relevant to human exposure scenarios?

-Are there significant variations in findings between apparently similar studies and if so are there explanations of these?

Is the information sufficient to characterise a mode of action that is likely to be relevant to man?

Comparison between different lines of evidence on the hazardous properties and dose response relationships may be required. This comparison should be done in a systematic way, recognising the following:

- Has all the relevant available data been properly considered (testing and non-testing information)?

- What basis has been used for the ranking of testing and non-testing data?
- What are the critical data gaps and why are they critical?

7.4. Conclusions on the overall risks

Prior to the development of the conclusions it is important to reconsider all aspects of the risk(s) and risk factors under consideration, particularly what may be missing since failure to do so may result in inappropriate conclusions. Dimensions of risk include the severity of the effect/outcome (nature of the adverse effect) and the likelihood of its occurrence at individual or population level. Both of these aspects should be addressed in depth in risk characterisation.

It is not recommended to simply add together weighting from individual lines of evidence to reach a final conclusion. Rather the combining of conclusions from different lines of evidence demands an element of expert judgement. In making the final weight of evidence assignment the basis for the judgement based decisions should as far as practicable is summarised. To identify strong weighting profiles need to be identified for different types of questions. For example:

* For a new chemical there is inevitably a lack of human data. A strong profile in terms of priority lines of evidence would be:

- Consistent animal data in more than one species;
- Indicators of mode of action along with consideration of its relevance to humans.

Currently, the results from *in silico* and specific *in vitro* tests tend in general to have a lesser contribution to the overall weighting. However this may change with time based on growing experience on the utility of such information for risk assessment purposes.

* For a chemical in widespread use on the other hand human data should in principle receive the highest weighting.

The key issue in evaluation of human evidence is to assess whether the results demonstrate a true causal effect, what is the affected population and to what extent the adverse effects of the exposure might be avoidable.

This involves:

- Estimation of incidence and severity of adverse effects likely to occur in a population/ecosystem due to exposure to a substance
- Addressing several potential toxic effects and human (sub)populations, and considering each (sub)population's exposure by relevant exposure routes
- Focus on most critical effect(s) (with consideration of population, route, and time scale)
- Provide quantitative (or if not possible, qualitative) assessment of risk, and
- Characterization of the sources and magnitude of uncertainties

Crucial in the determination of the critical effect is:

- Differentiation between non-adverse and adverse effects
- Ensuring that the adverse effect is related to exposure (substance-related)
- Assessment of biological significance not simply statistical significance
- Presence of dose/time-effect relationship

- Data on the reversibility of effect
- Information on normal variation in the incidence of the disease (effect0 of interest (e.g. consideration of historic controls)

A well-defined and consistent framework can aid reaching conclusions in risk assessment and indicate confidence in the findings. One such a framework is set out in table 1. If the risk assessment covers both human and environmental risks, separate tables should be constructed. Yet such scheme provides only as a framework for the process. With the lack of experience it is not appropriate to define scores for acceptability.

7.5. Weighing of the total evidence

The weighing of the total evidence can be presented as in the following table:

Table 4 Contribution of the different lines of evidence to the opinion

Factor	Strong	Moderate	Weak	Uncertain	Not possible
A. Weight of evidence from the following lines of evidence: Exposure measurement Exposure modelling Epidemiologic studies Human volunteer studies Other human data sources Animal studies In vitro studies Mathematical models, structure activity and other *in silico* data Studies on Mechanisms					
Conclusion from the totality of evidence (short description)					

Human, animal and mechanistic studies comprise the primary line of evidence along with exposure.

Strong overall weight of evidence: Coherent evidence from human and one or more other lines of evidence (in particular mode/ mechanistic studies) in the absence of conflicting evidence from one of the other lines of evidence (no important data gaps).

Moderate overall weight of evidence: good evidence from a primary line of evidence but evidence from several other lines is missing (important data gaps).

Weak overall weight of evidence: weak evidence from the primary lines of evidence (severe data gaps).

Uncertain overall weight of evidence: due to conflicting information from different lines of evidence that cannot be explained in scientific terms.

Weighing of evidence not possible

No suitable evidence available.

7.6. Additional explanatory information

In each case free text is required to explain the assignment. It is important to identify studies that appear to have been well conducted but generate findings that are very different (outliers) from those of other studies in the same line of evidence. Differences between apparently very similar, good quality studies (genuine variability) also need to be addressed in the final risk assessment along with comments on possible unknown unknowns.

8. EXPRESSION OF UNCERTAINTY

8.1. The need for expression of uncertainty

The strength of evidence is inversely related to the degree of uncertainty. In principle, therefore, completion of Table 5 and the supporting rubric is adequate to indicate the level of uncertainty in the use of the data for a specific risk assessment. Nonetheless it is proposed that uncertainty is specifically described in the SCENIHR opinions, since uncertainty is an essential part of the weighting in each rubric for each line of evidence.

Characterization of the uncertainties in a risk assessment is important for transparency and should also be a valuable aid to risk managers in determining how to respond to risk management advice. In addition it is a useful way of indicating priorities for further work to improve the robustness of risk assessments. However, if not clearly and suitably described, the expression of uncertainty may result in inappropriate concerns and/or actions. The degree to which characterisation of uncertainty (and variability) is needed will depend on the risk assessment and risk management contexts as determined in the questions asked, i.e. problem formulation.

Uncertainty analysis should be incorporated during the weighing of evidence rather than added after this process is completed. Integration of the uncertainty analysis with the other parts of the risk assessment process should be carried out wherever possible. In the unusual case that uncertainty analysis has to be carried out after an opinion is completed, a top down approach may be appropriate. Procedures that can be applied in a top down approach include a critical path analysis along the lines of HAZOP or HACCP (e.g. Sperber 2005, Dunjo et al 2010).

To date, there has been no formal system in regular use by the SCENIHR or other committees of the EU concerned with risk assessment to express uncertainty. Instead various terms such as likely, probably, etc, have been used.

Most of the currently conducted risk assessments are deterministic rather than probabilistic. In deterministic regulatory hazard assessment although the uncertainty is

not specifically stated, standard default values are often used to allow for identifiable uncertainties. The use of standard defaults often incorporates two separate elements, namely: the actual uncertainty (which often includes an allowance for data variability) and the achievement of a desired level of protection (the so-called expanded uncertainty (see Sassi and Ruggeri 2008). It should not be the role of risk assessors to set a desired level of precaution; this is the responsibility of risk managers and other stakeholders. Where probabilistic risk assessments are conducted, worst case scenarios are often built in and, in order to identify the uncertainty, these conservative assumptions need to be properly characterised.

8.2. Weight of evidence and uncertainty

In addition to considering the uncertainties involved in the weighing of evidence it is important to identify significant uncertainties in the judgement used. This may include:

- extrapolation from *in vitro* to *in vivo;*
- extrapolation from animals to man /test species to field situations (if defaults are used their suitability);
- extrapolation from inter-individual differences (if defaults are used their suitability);
- extrapolation from acute or sub acute studies to (semi-)chronic;
- extrapolation between routes of exposure;
- use of computer/mathematical modelling;
- applications of putative mechanisms;
- other assumptions made.

8.3. Expression of the uncertainties

A simple scheme is required that is readily understood by both risk assessors and risk managers. Uncertainty may be expressed in several ways namely using:

a) Standardised terms or phrases. Various terms are used by the EU scientific committees. However as noted in the SSC opinion on harmonisation of risk assessment (2000, 2003), there is no consistency in how different terms are used.

b) Tabular forms. This must be linked to the weight of evidence summary table (table 5).

c) Quantitative expression. This is only apropriate if the risk assessment is expressed in probablistic terms.

These three ways of expressing uncertainty may be regarded as a tiered approach. If there is limited data, the use of standardised terms may be the only one suitable.

8.3.1. Standardised terms

The expression of the significance of the uncertainties associated with a particular risk assessment taking into account both weighting of evidence and judgemental factors is proposed as:

1 Certain (i.e. very little doubt, around 1in 100 or greater chance of being wrong);

2 Probable (i.e. reasonable confidence, of the order of 1in 10 of being wrong);

3 May (i.e. some confidence, of the order of 1in 3 to 1 in 5 of being wrong);

4 Possible (i.e. rather limited confidence);

5 Very uncertain (i.e. no confidence).

Other terms to express certainty and uncertainty should not be used without a supporting text.

8.3.2. Tabular expression of uncertainties

Various systems have been proposed to set out the uncertainties in a risk assessment including tables (Hart 2010) an uncertainty matrix (Walker *et al.,* 2003; Van der Sluijs *et al*., 2003) and source listing (EFSA, 2006; IPCS, 2008; REACH Guidance 2008b).

The tabular presentation from table 5 is proposed to be used as the basis for proposed tabular presentation of uncertainties given in Table 6. In the rubric linked to the table the symbols used are identified.

Table 6: Expression of uncertainty for individual lines of evidence

Aspect	Nature of the uncertainty	Magnitude and direction of the uncertainty	Importance of the uncertainty to the risk assessment
Quality of the data **Key aspects:** * *			
Comprehensiveness of the data **Key aspects:** * *			
Judgements and assumptions made **Key aspects:** * *			

The table should indicate:

- The direction of any uncertainties, i.e. are they equally distributed or are they most likely to be over or underestimates of the risk. This requires consideration of the degree of conservatism used in modelling, etc;
- The magnitude of any uncertainties, i.e. are they likely to be small or large;
- Any allowance already made for each uncertainty;
- The importance of each uncertainty in the overall level of confidence in the conclusions of the risk assessment.

It may be helpful to use the following symbols to simplify the expression of the analysis:

Direction of uncertainties

The direction of uncertainties (i.e. whether there is a trend towards an over or underestimation of the risks) should be expressed by the use of + and – values as follows:

+ The risk could be higher due to the uncertainty;

- The risk could be lower due to the uncertainty;

+/- There is an equal chance of the uncertainty producing a risk estimate that is either too high or too low.

Magnitude of the uncertainties

For each assignment a degree of uncertainty can also be assigned according to the IPCC (2005) categorisation:

1. Virtually certain (99% probability that the risk assessment is accurate i.e. insignificant uncertainty);
2. Very likely (uncertainty is around or less than 1 in 10), i.e. rather limited uncertainty);
3. Likely (significant uncertainty of around 1 in 3);
4. About as likely as not (high level of uncertainty, around 50:50).

It is likely that some uncertainties cannot be readily assigned a magnitude. However this must not mean that they are not considered. These should be listed in the nature of uncertainty column and the third column of impact on the risk assessment completed.

Allowance for uncertainty already made

Where significant uncertainty in the risk assessment is identified it is important to consider first whether some allowance for uncertainty ha already been included in the risk assessment by the use for example of conservative modelling or default factors. If this is the case, the following needs to be addressed:

a) is there is an allowance for overestimate of the risk estimate only?

b) is there is an allowance for underestimate of the risk estimate only?

c) is the allowance for over and under estimate of the risk estimate comparable?

Importance of each uncertainty to confidence in the risk assessment conclusions

If significant uncertainty is identified in the overall risk assessment, then it is appropriate to consider where the uncertainties lie in the risk assessment and how important they are in determining the accuracy of the risk assessment.

The importance may be classified as:

i) Insignificant

ii) Rather limited

iii) Considerable

iv) Large

If the impact of the uncertainty is estimated to be considerable or large, an explanatory text on the implications for the risk assessment conclusions will be necessary.

8.3.3. Quantitative expression of uncertainty

A probabilistic approach to the risk assessment is adopted then quantification of the uncertainties should be carried out where possible. Probabilistic risk assessment (PRA) potentially gives more information to both risk assessors and risk managers, because it gives more quantitative insight into the range of possible outcomes in a risk assessment and the degree of cumulated conservatism in the deterministic risk assessment. In a deterministic assessment the main concern usually is to avoid acceptance of harmful substances: i.e. the errors made decrease chances of detecting false negatives or Type II errors. This is fine if the point estimate clearly shows insignificant risks, erring on the side of safety is an acceptable policy, one does not worry about costly risk reduction measures or the final decision is based on socio-economic factors (e.g. very high, low benefits versus costs). However, if the policy is to reach an optimum decision on prioritisation, acceptance or rejection of harmful substances and cost-effective risk reduction measures, one would like to know the full distribution of the risk to know the balance between false positives (Type I errors) and false negatives.

PRA can use all information about quantifiable variability and uncertainty in both the exposure and the effects assessment and forces experts to reveal the nature and extent of their judgment, e.g., on types of uncertainty, distributions, the shape of the dose-response curve and the nature of the critical effect. Sensitivity analysis is able to reveal the relative impact of uncertainties in parameters on the final result and can reveal where the risk assessment can be improved in the most time- and cost-efficient manner and whether it is necessary and achievable to reduce the uncertainty further.

The probabilistic approach in exposure assessment needs the following steps:

- A clear separation needs to be made between the uncertainty due to lack of knowledge and that due to variability to be able to answer different risk questions. These two types of uncertainty could also be treated separately in a two-dimensional probabilistic analysis which would result in probability curves showing risk levels for different percentiles of the population, together with confidence bounds showing the combined effect of those uncertainties that have been quantified (e.g. Hoffman and Hammonds, 1994; Van der Voet and Slob, 2007; IPCS, 2008).
- It should be made very clear which uncertainties due to lack of knowledge and which due to variability are included in the risk assessment and which not. Non-quantifiable uncertainties such as poor data quality or model uncertainty cannot be easily addressed in the probabilistic approach.

8.3.4. Explanation of the implications of the uncertainties to risk managers

In situations where a considerable uncertainty is identified, it is proposed that guidance is provided by the SCENIHR where appropriate on actions that might be taken to reduce the uncertainty. Possibilities include:

- Relevant data expected to be imminent,
- Specific research recommended to substantially reduce the uncertainty,
- Unlikely that the uncertainties can be significantly reduced in the foreseeable future,
- Options for precautionary risk management measures recommended to avoid further exposure.

In the first two cases the time involved and the nature of the likely reduction in uncertainty should be indicated.

9. COMPLETION OF THE PROCESS

The weighing of evidence and the assignment of uncertainties for questions raised by the Commission is normally conducted by working groups that may comprise a majority of non SCENIHR members. Inevitably each working group involves a number of independent scientists in different disciplines a number of whom may have had no previous experience of working with the SCENIHR. It is vital for the purposes of transparency, consistency and overall scientific quality that once the weighing of evidence process is complete a final check is made on the following:

- The questions asked by the risk manager(s) have each been addressed in a way that is clear, transparent and readily understood. This is a role for both the working group and the SCENIHR, with the support of the Secretariat.
- All the references are appropriately cited and categorised. This aspect must be the primary responsibility of the Working Group with the support of the Secretariat
- The various lines of evidence are combined to provide an overall assessment in a manner that is both transparent and consistent with other SCENIHR opinions. This is a task for the full SCENIHR.
- There is a clear linkage between the identification and nature of important data gaps and other statements in the opinion on uncertainty. This is a task for the full SCENIHR committee.

10.CONCLUSIONS

This memorandum is intended to make explicit the approach used by the SCENIHR for determining the weight of evidence and the uncertainties involved in the development of its opinions. It involves a staged approach. The approach draws on a number of schemes that have been developed by various national and international bodies. However it introduces a number of additional elements that are considered to benefit both transparency and consistency.

Particular attention has been paid to ensuring that the format can be applied to a wide range of lines of evidence and types of publication.

11.RECOMMENDATIONS

It is recommended that the procedures set out in this memorandum are used by the SCENIHR where appropriate and that the methodology is readily available to stakeholders. It would be helpful to have the opinion from ecological risk assessors on the extent to which this scheme could be applied in their domain. To ensure transparency and to enable input from others the memorandum should be publically available.

12.REFERENCES

Boobis AR, Cohen SM, Dellarco V, McGregor D, Meek ME VickersC, Willcocks D and Farland W (2006) IPCS framework for analysing the relevance of a cancer mode of action for humans Crit Rev Toxicol 36 781-792,

Boobis AR, Doe JE, Heinrich-Hirsch B, Meek ME, Munn S, Ruchirawat M, Schlatter J, Seed J and Vickers C, (2008) IPCS framework for analysing the relevance of a noncancer mode of action for humans Crit Rev Toxicol 38, 87-96

DFG (2010), List of MAK and BAT values. Wiley.VCH, Weinheim, Germany

Dunjo J, Fthenakis V, Vilchez JA and Arnaldos J (2010) Hazard and operability (HAZOP) analysis. A literature review J Hazard Mats 173, 19-32

ECHA. 2008. Guidance on information requirements and chemical safety assessment. Part D: Exposure cenario Building. ECHA, Helsinki, Finland. www.echa.europa.eu

EFSA. 2006. Guidance of the Scientific Committee on a request from EFSA related to uncertainties in dietary exposure assessment. The EFSA Journal 438:1-54.

Gee D. 2008. Establishing evidence for early action: the prevention of reproductive and developmental harm. *Basic Clin Pharmacol Toxicol* 102:71-72.

Hart A, Roelofs W, Hardy AR, Macleod A. 2007. Comparative review of risk terminology. A comparative review of terminology and expressions used by the three scientific committees established by Commission Decision 2004/210/EC and by their predecessors established by Commission Decision 97/579/EC (repealed by Commission Decision 2004/210/EC). Report No. S12.454739. The Central Science Laboratory, DEFRA, Sand Hutton, York, UK.

IARC. 2006. The Preamble to the IARC Monographs. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Lyon, France.

IPCS. 2008. Uncertainty and data quality in exposure assessment. Part 2: Hallmarks of data quality in chemical exposure assessment. WHO/IPCS, Geneva, Switzerland. IPCS Harmonization Project Document No. 6. www.who.int/ipcs

IPCS, 2009. Risk assessment of combined exposures to multiple chemicals: a WHO/IPCS framework. WHO/IPCS, Geneva, Switzerland. Harmonization Project DRAFT Document for Public and Peer Review

Kalberlah F, Schneider K and Schuhmacher-Wolt U (2003) Uncertainty in toxicological risk assessment for non-carcinogenic health effects Regulat Toxicol Pharmacol 37, 92-104

Klimisch HJ, Andreae M and Tilmann U (1997) A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data Regul Tox and Pharmacl 25 1-5

Levin R, Jansson SO, Rudén C. 2004. Indicators of uncertainty in chemical risk assessments. *Regul Toxicol Pharm* 39:33-43.

Linkov I, Loney D, Cormier S, Satterstrom FK and Bridges T (2009) Weight of evidence evaluation in environmental assessment: review of qualitative and quantitative approaches Sci Total environ 407, 5199-5205

Nendza, M., Aldenberg., Benfenati, E., Begnini, R., Cronin, M.T.D., Escher, S., Fernandez, A., Gabbert, S., Giralt, F., Hewitt, M., Hrovat, M., Jeram, S., Kroese, D., Madden, J.C., Mangelsdorf, I., Rallo, R., Roncaglioni, A., Rorije, E., Segner, H., Simon-Hettich, B., Vermeire, T. 2010. Data quality assessment for *in silico* methods: a survey of approaches and needs. Chapter 4 in: Cronin, M.T.D., Madden, J.C. (eds.) In silico toxicology, principles and applications. RSC Publishing, Cambridge, UK. ISBN 978-1-84973-004-4

OECD, 2007. Guidance document on the validation of (Q)SARs. Organisation of Economic Cooperation and Development, Paris, France. Series of Testing and Assessment 69

Sassi G and Ruggeri B (2008) Uncertainty evaluation of human risk analysis (HRA) oc chemicals by multiple exposure routes Risk Analysis 28 1343-1356

Sperber WH (2005) HACCP and transparency Food Control, 16, 505-509

Van der Sluijs JP, Craye M, Funtowicz S, Kloprogge P, Ravetz J, Risbey J. 2005. Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: the NUSAP system. *Risk Anal* 25:481-492.

Van der Sluijs JP, Risbey JS, Kloprogge P, Ravetz JR, Funtowicz SO, Quintana SC, Pereira AG, De Marchi B, Petersen A, Janssen PHM, Hoppe R, Huijs SWF. 2003. RIVM/MNP guidance for uncertainty assessment and communication. Detailed Guidance. Utrecht University, Utrecht, The Netherlands, ISBN 90-393-3536-2.

Van de Meent D. De Bruijn JHM.2007. Environmental exposure assessment. In: Van Leeuwen CJ, Vermeire, TG., Risk assessment of chemicals: an introduction. Springer, ISBN 978-1-4020-6101-1

Weed DL. 2005. Weight of Evidence: a review of concept and methods. *Risk Anal* 25:1545-1557.

Other relevant references not specifically cited so far

Evans JS and Baird SJS (1998) Accounting for missing data in non-cancer risk assessment Human and Ecological Risk Assessment 4, 291-317

Gamble C and Hollis S (2005) Uncertainty method improvement on best-worst case analysis in a binary meta-analysis J Clin Epidemiol 58, 579-588

13.LIST OF ABBREVIATIONS

CA	chromosome aberration test
MN	micronucleus test
SCE	sister chromatid exchange
SCGE	single cell gel electrophoresis

14.ANNEX

The framework for risk assessments by scientific committees and potential factors that may limit their value.