



Project No. 018385
INTARESE
Integrated Assessment of Health Risks of Environmental Stressors in Europe

Integrated Project
Thematic Priority

D15 Health Effects Methodology - Protocol and Guidelines

Due date of deliverable: April 2007
Actual submission date: April 2007

Start Date of Project: 1 November 2005	Duration: 60 Months
Organisation name of lead contractor for this deliverable: UU (IRAS), the Netherlands	Revision: Final

Project co-funded by the European Commission with the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	x
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contents

Summary

1 Background on the INTARESE project	1
1.1 Introduction	1
1.2 Rationale	2
1.3 Overall aims	2
1.4 Objectives	3
1.5 Structure of INTARESE	5
1.6 Aims of WP 1.3	5
1.7 Links with SP1	7
2 Background	8
2.1 Health impact assessment	8
2.2 Terminology of exposure response functions	10
2.2.1 What is an exposure response function?	10
2.2.2 Exposure assessment terminology and approaches	10
2.2.3 Different types of health effects	14
2.3 Final remarks and outline of the protocol	16
3 Selection of exposure response functions for quantification	17
3.1 Strength of evidence, severity and number of people affected	19
3.2 Location / population selection	24
3.3 Combined exposures	25
4 Review of methods to derive ERFs	29
4.1 Systematic review	30
4.1.1 Systematic review to derive ERFs	31
4.1.2 Meta-regression methods	39
4.1.3 Meta-regression methods to derive ERFs	39
4.1.4 Bayesian analysis	42
4.1.5 Bayesian meta-analysis	43
4.2 Expert panel	45
4.2.1 Expert panel to derive ERFs	46
4.2.2 Role within the INTARESE project	53

4.3 Comparison of methods	55
4.4 Extrapolating from animal to human	56
4.4.1 Qualitative aspects of animal studies	56
4.4.2 Which animal models are relevant for humans?	58
4.4.3 Quantitative aspects of non-cancer endpoints in animal studies	59
4.4.4 Guidelines for carcinogen dose-response risk assessment	61
4.5 Informal methods	65
5 Review of methods for characterizing uncertainty in ERFs	67
5.1 Introduction	68
5.2 Ordinary sensitivity analysis	68
5.3 Bayesian methods	69
5.4 Monte Carlo sensitivity analysis	69
5.5 Multiple-bias modelling	69
5.5.1 Examples of Monte Carlo sensitivity analysis and Bayesian analysis	70
5.6 The NUSAP approach	71
6 Combining animal and human studies in exposure-response assessment	73
6.1 Uncertainties in the different study types	74
6.2 Recent developments in quantitatively combining human studies	76
6.3 Recent developments in quantitatively combining animal studies	76
6.4 Combining human and animal studies	77
7 Suggested methodology	80
8 Further work	84
References	85
Appendices:	96
1 Review of health impact assessment issues	96
2 Hazard identification	105
3 WHO meta-analysis	108
4 Possible sources of uncertainty within individual studies	109

Summary

This protocol aims to give a general methodology to derive a certain exposure-response function (ERF). It is meant to assist the SP3 teams in their policy assessments. The current text describes methods for ERFs that can be used in the first assessment. The methodology will be further developed in the next phase to refine and improve the suggested methodology.

Methods developed in health impact assessment (HIA) should be the basis of the SP-3 policy assessment. One key issue is transparency, that is, assumptions should be made explicit. Another key issue is integration of methods used in different parts of the chain, e.g. the level of detail of exposure information should agree with that for which exposure response functions are available. An assessment should start with listing potential relationships between stressors and health, not all of which need to be quantified. Criteria for quantifying relationships are the assumed causality of the association; the severity of the health response and the number of people affected.

We recommend that if there is already a published and up-to-date ERF available, preferably from an authoritative and influential institute or organisation, like for example the World Health Organization, one should use that in the SP-3 assessments. If not available, we recommend using the frequentist systematic review (including if appropriate a meta-analysis) to derive an ERF for key impact pathways. An alternative is to make use of the formal methods of an expert panel. Both methods are time consuming and therefore modified less-time consuming versions are discussed. Options include any of the following: making use of the ERF used in previous HIAs; utilizing the results of a previously published good-quality meta-analysis; using a key multi-centre study or a core (non-exhaustive) set of studies. If you make use of these informal methods it is recommended that you consult expert(s) in your policy field (thus not a real expert panel!).

These methods can be applied to animal and human studies. However, if applied to animal studies, extrapolation to humans is necessary. First, a qualitative assessment is necessary to evaluate whether an observed effect in animal studies applies to humans.

In addition to the recommended systematic review, one should consider some important sources of uncertainty by filling in two tables for human and animal studies separately (qualitative and quantitative assessment).

Box 1 below presents a checklist of factors that should be considered in deriving an ERF. This serves as a summary of this protocol.

Over the next few years we will work on the refinement of ERFs using both animal and human data. This is strongly linked to the topic of taking into account uncertainty more systematically. Proposed methods included forming a formal expert panel, more informal methods like using expert judgements/opinions and guidance, multiple-bias (Bayesian) modelling and the NUSAP approach. To test and illustrate the suggested methods, a set of case studies of selected exposure-health effects relationships will be performed. We will focus on the following exposures: black smoke, ultrafine particles, dioxins, disinfection by-products and traffic noise.

Starting point is a certain chosen exposure

1. Which health endpoints have been associated with the exposure?
2. Which health endpoints are likely causally associated with the exposure?
3. Which ERF will be further quantified?
4. Is there a recent authoritative (systematic) review that can be used to characterize the ERF?
 - Evidence for a threshold?
 - Evidence for another mathematical function?
 - Quantitative relationship
5. Is it likely given the exposure levels and the ERF that a health response is non-zero?

THE NEXT QUESTIONS APPLY IF AN ERF NEEDS TO BE DEVELOPED

6. For which exposure metrics (e.g. external, internal dose) are studies available?
7. Is the linear mathematical model reasonable for the ERF?
8. Representation of exposure data in the individual studies
 - Choice of taking a specific averaging time
 - Which lag time?
9. For which population group is an ERF developed (general population or subgroups such as different age groups, children, elderly)
10. Are there differences in ERF related to geographical location?

1 Background on the INTARESE project

1.1 Introduction

INTARESE brings together a team of internationally lead scientists in the areas of epidemiology, environmental science and biosciences to collaborate on developing and applying new, integrated approaches to the assessment of environmental health risks and consequences, in support of European policy on environmental health. This project is designed to support implementation of the European Environment and Health Action Plan, by providing the methods and tools that are essential to enable integrated assessment of environment and health risks. Drawing upon the large range of studies carried out in Europe over recent years (many led by the project partners) and the advances made in specific areas of toxicology and epidemiology (especially air pollution), and in close collaboration with users, it will develop a methodological framework and a set of tools and indicators for integrated assessment that can be applied across different environmental stressors (including pollutants and physical hazards), exposure pathways (air, water, soil, food) and policy areas. It will review, bring together and enhance the monitoring systems needed to support such analyses, including routine environmental monitoring (ground-based and Earth observation), biomonitoring and health surveillance. The framework, tools and data will be tested and demonstrated through integrated assessments of exposures and health risks in a number of specific policy areas, including transport, housing, agriculture, water, wastes, household chemicals and climate.

Results from these will be used both to refine the assessment methods and to provide specific information on health implications of current and potential future, policies. Based on the results, a toolbox for integrated environment and health risk assessment will be developed, which will be further tested and demonstrated through a series of higher level policy analyses. Particular attention will be given throughout to issues of uncertainty, sensitive or susceptible groups, and possible interactive and cumulative effects of different stressors. Deliverables will include new, integrated methods and indicators for environment and health risk assessment and monitoring, an operational assessment toolbox, and a set of validated assessments that can directly inform policy.

1.2 Rationale

This project is designed to make a major and practical contribution to implementation of the European Action Plan on Environment and Health, the European Environment and Health Strategy and the European Integrated Environment and Health Monitoring and Response System, and, more generally, to assist in achieving the goals of the Environmental Action Plan. It is based on a simple but powerful precept: one that has in recent years repeatedly emerged from initiatives such as the SCALE procedure and the formative stages of GMES. This is that, if scientific support for policy on environment and health is to be improved, the immediate priority is not so much for further detailed studies, investigating individual causal associations between environment and health, but rather to improve the use made of the data and knowledge that we already have in order to obtain more integrated assessments of risks and impacts. In this context, three key gaps need to be addressed:

1. data and knowledge are spread across disciplines, through different networks and in different databases – tools, methods and collaborative research are needed to bring together and link these different areas of data and knowledge more effectively to inform integrated assessments;
2. in many areas, a large gap between science and policy remains – methods are needed to bridge this gap by translating the science that exists into information that is of direct relevance to policy;
3. in specific contexts, there are key gaps in data or knowledge that break the continuity of our current understanding and undermine its utility for policy support–targeted research

1.3 Overall aims

The overall aims of the project are thus:

1. to develop a conceptual framework to bring together the latest scientific evidence across all the relevant environmental sectors and disciplines as a basis for integrated assessment of both environmental and health impacts and risks;
2. to address the scientific and information deficits that currently exist by:

- specifying the information and knowledge requirements needed to implement this framework, identifying the key gaps in existing monitoring and analytical capacity in Europe, and
 - developing and testing new systems for monitoring and modelling (e.g. biomonitoring, exposure assessment).
3. informed by this work, to build an operational toolbox for integrated assessment that can be applied to different stressors and environmental media (air pollution, water pollution, climate change etc), settings (ambient, domestic, occupational) and agents (chemicals, solid wastes, natural hazards, noise etc), in order:
- to provide early warning of environment and health problems;
 - to help quantify and compare environment and health risks and thereby establish policy priorities;
 - to help set policy objectives and targets;
 - to make international comparisons and assess progress towards these policy targets;
 - to inform the public and other stakeholders; and
 - to enhance scientific support to policy.
4. to apply this approach to undertake integrated assessments for a range of key policy areas, including transport, housing, agricultural land use, water management, household chemicals, waste management and climate, in order:
- to help develop and test the methodology and customise it to different policy needs; and
 - to provide information of direct and immediate relevance to current and emerging policy.

1.4 Objectives

The specific objectives and associated outcomes and deliverables of this project are as follows:

To develop a coherent, conceptual framework for integrated environment and health risk assessment. To compare, evaluate and develop a set of methods and indicators to represent the links between source and exposure, for use in the assessment procedure.

1. to compare, evaluate and develop a set of methods and indicators to represent links between exposure and health effect.
2. to develop a set of methods and indicators to characterise these risks and impacts in a form of direct relevance to policy, including economic measures such as the cost-benefit ratio and societal measures such as those based on multi-criteria assessment.
3. to analyse and assess the various generic and cross-cutting issues that need to be considered in applying methods of integrated assessment to environment and health issues.
4. to define and assess the environmental information needed to support routine implementation of these integrated assessment methods in support of policy, and to develop and evaluate new data sources, monitoring techniques and models.
5. to define and assess the capability of bio monitoring and associated modelling techniques to meet the information needs of integrated assessment, and to develop and evaluate new data sources, biomarkers, monitoring techniques and models.
6. to identify needs for health information in support of integrated assessment methods, to assess the ability of existing surveillance systems to provide this information, and to analyse the potential to establish more effective and concordant health surveillance systems across the EU.
7. to assess the capability, and where appropriate develop the methods, to combine these various monitoring and analysis systems into an integrated monitoring system, covering different environmental agents, media and pathways, and different population groups.
8. to apply the integrated assessment framework and methods developed within the project to assess health risks and impacts associated with a range of different policy issues, under different policy scenarios.
9. to translate these methods and indicators into an operational, computer-based toolbox (decision support and analysis system) for integrated assessment for policy support in Europe.
10. to demonstrate the application and utility of this toolbox by applying it to a range of policy issues at the European level.
11. to manage and support the project effectively, in order to ensure its successful conclusion.

1.5 Structure of INTARESE

INTARESE is planned as a 5-year integrated project, comprising four main phases. Phase 1 (to month 18) focuses on development and testing of the assessment methods, and assembly of the information needed to apply these through a series of policy-related case studies. Phase 2 (months 18-36) uses the methods to carry out integrated assessments for these policy issues, and to refine and revise the assessment methodology in the light of that experience. Stage 3 (months 36-48) constructs and tests an operational assessment toolbox; phase 4 (months 48-60) applies this toolbox to a set of high-level case studies. The end of each of these phases represents a key milestone and review point (see also figure 1).



Figure 1 INTARESE comprises 4 main phases

As this indicates, it comprises seven major components (sub-projects). Five of these focuses on scientific research, the sixth on user consultation and dissemination, and the seventh on consortium management (see figure 2).

1.6 Aims of WP 1.3

The overall aim of this WP Exposure-health effect workpackage is to develop methods, protocols and an information base for converting information of exposures into quantitative estimates of health effects. Specific objectives are thus:

1. to develop a systematic methodology to assess ERFs for diverse environmental exposures such that the output can be linked to the valuation of health effects (WP 1.4) and used as a basis for the policy assessments in SP-3

2. to apply the developed methodology to selected environmental exposures, relevant for the policy studies identified in SP-3
3. to identify gaps in the selected exposure-response relationships
4. to resolve these gaps using existing data on exposure response relationships

The specific tasks for Phase 2 month 12-30 are to:

1. complete the development of the methodology (review)
2. apply the basic methodology to a limited number of selected exposure response functions
3. guide SP-3 in applying the methodology appropriately
4. refine the methodology by developing methods to integrate information from epidemiological and toxicological (animal – human) studies
5. refine the methodology by developing methods to better characterize uncertainty
6. identify gaps in the selected exposure-response relationships
7. resolve some of these gaps using existing data on exposure response relationships

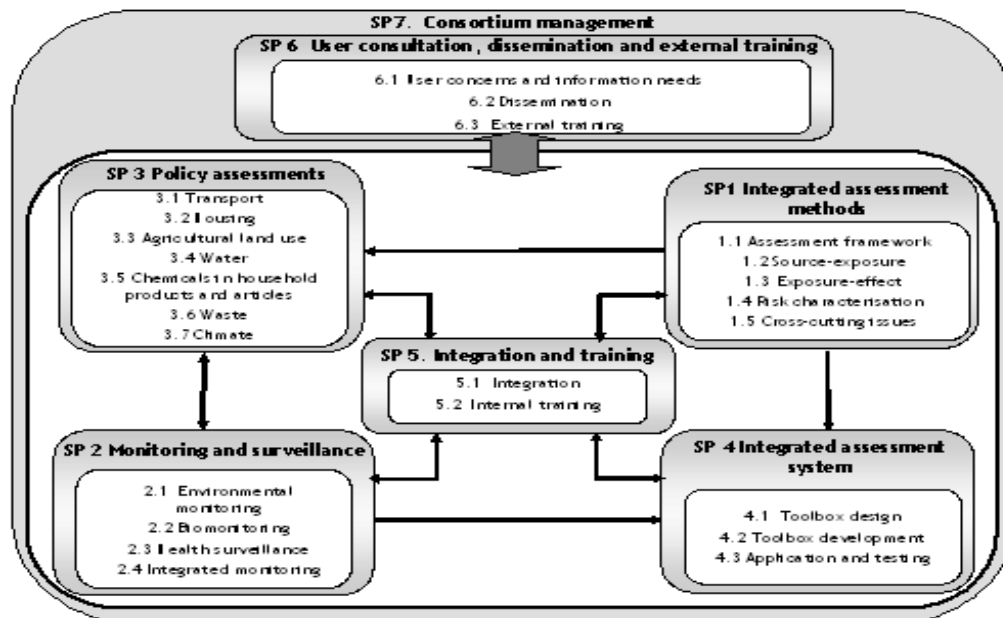


Figure 2 The overall structure of INTARESE

1.7 Links with SP1

WP 1.1 was requested to provide a conceptual framework by month 12. They adopted a full chain approach, as an extension of earlier DPSEAA and MEME frameworks. They developed and distributed a questionnaire to SP3 to identify the core indicators used in the common policy assessments in their domains, associated models and main issues of uncertainty and debate.

Methods will be developed to translate this framework into an operational system for integrated assessment. This will comprise three key sets of analytical techniques:

1. methods to link sources to exposures, including (WP 1.2)
 - improvements in exposure modelling
 - development of intake fraction and source apportionment techniques
2. methods to link exposure to health outcome, including (WP 1.3)
 - dose-response and exposure-effect relationships
3. methods to characterise the resulting risks and impacts e.g. through the development and application of: (WP 1.4)
 - policy-relevant indicators (e.g. DALYs)
 - techniques such as health impact analysis and cost-benefit analysis
 - measures and techniques for weighting of information (e.g. using Bayesian belief networks or value-of-information methods)
4. cross-cutting issues which involves:(WP 1.5)
 - uncertainties
 - variability in susceptibility
 - multiple exposures
 - environmental justice
 - scale of the analysis
 - value systems

2 Background

In paragraph 2.1 a short review is given of issues in health impact assessment (HIA) in order to provide a framework for the exposure-response assessment work. The methodology to assess ERFs for diverse environmental exposures must fit into the overall HIA methodology, as we ultimately need to estimate the burden of disease attributed to a certain exposure. This does not only involve an ERF, but also exposure estimation and the assumed link between them. Problems identified in HIA are important not only for WP 1.3 but for the INTARESE SP-3 assessments in general. Paragraph 2.2 describes what an exposure-response function is and the different exposures and health effects. Paragraph 2.3 discusses shortly the issue of combined exposures.

2.1 Health impact assessment

A WHO working party defined a HIA as a combination of procedures, methods and tools by which a policy, program or project may be judged as to its potential effects on the health of a population, and the distribution of those effects within the population (1). Specifically, the purpose of a HIA is (2):

- to assess the potential health impacts, both positive and negative, of projects, programmes and policies
- to improve the quality of public decision making through recommendations to enhance predicted positive health impacts and minimise negative ones

The major steps that play a role in a HIA are the following (3, 4):

1. specify the purpose and framework of the HIA
2. decide which exposure-effect pathways will be quantified
3. identify and characterise the population at risk
4. select or develop a suitable set of exposure-response functions (ERFs) that link (individual) pollutants with specific health endpoints, i.e. % increase in morbidity per $\mu\text{g}/\text{m}^3$ of a pollutant
5. derive the population exposure distribution

6. estimate the background rates of the relevant health endpoints in the population at risk
7. calculate the burden of disease or death in the population at risk
8. value the burden of disease or death in the population at risk
9. assess and quantify the uncertainty of the HIA

There has been much debate about the HIA methodology; different approaches have been proposed (2, 5-7). It has been concluded that there is no single ‘blueprint’ for HIA that will be appropriate for all circumstances (6). For a more detailed description of HIA methods, see the Merseyside Guidelines for health impact assessment (2) or check www.who.int/hia/en/ for more (country specific) guidelines.

Table 1 Overview of health impact assessment (HIA) issues

Steps in HIA	HIA issues
1. Specify the purpose and framework of the HIA	Transparency
2. Decide which exposure-effect pathways will be quantified	Systematic approach Judgements
3. Identify and characterise the population at risk	Integration several disciplines Mixture of pollutants Combined effects Vulnerable individuals Environmental justice
4. Select or develop a suitable set of exposure-response functions (ERFs)	Limited data Transferability Mathematical choices Limited exposure assessment Integration several disciplines Precautionary principle Methodology for developing ERFs
5. Derive the population exposure distribution	Linkage with ERFs
6. Estimate the background rates of the relevant health endpoints in the population at risk	Limited data
7. Calculate the burden of disease or death in the population at risk	Non-comparability Use of RR of RD Competing risks
8. Value the burden of disease or death in the population at risk	Discount rates Severity/quality weightings Double counting
9. Assess and quantify the uncertainty of the HIA	Measurement error modelling

Identified problems in HIA are the lack of transparency and the lack of a systematic approach. Typically there is limited data, so that a lot of choices are made that are not often made explicit. Another issue in HIA is the poor integration between various disciplines like exposure and health effect assessment or epidemiology/toxicology. Another important topic is how to deal adequately with uncertainty. In table 1 an overview is given of all the identified HIA issues.

The issues marked in bold are possible innovation ambitions and are important not only for WP 1.3 but for the INTARESE SP-3 assessments in general. Appendix 1 contains a more complete review of main HIA issues.

2.2 Terminology of exposure response functions

There are many different levels of exposures, different types of health effects and relations possible. Different terms are used in practice. The definitions to be used in this document are given below.

2.2.1 What is an exposure response function?

According to WHO guidelines (3) ERFs may be reported as a slope of a regression line with the health response as the dependent variable and the stressor as the independent variable. Alternatively an ERF may be reported as a relative risk of a certain health response for a given change in exposure. ERFs may be derived from studies in the field of epidemiology and/or toxicology. In addition to the central estimate, the uncertainty of the central estimate should be available as well, e.g. as a confidence interval.

2.2.2 Exposure assessment terminology and approaches

We make use of the recently published adoption of the ISEA glossary for the exposure assessment terminology (8). Exposure is defined as contact between an agent and a target. Contact takes place at an exposure surface over an exposure period. An agent is defined as a chemical, biological or physical entity that contacts a target. A target can be defined as any biological entity that receives an exposure or a dose of an agent.

Figure 3 shows the several steps between emissions of a pollutant in the environment and a potential health effect.

Exposure is sometimes also called external dose. The internal dose is the amount of a substance penetrating across the absorption barriers or exchange boundaries of an organism, via either physical or biological process, sometimes termed absorbed dose. The biologically effective dose is the amount of an agent that reaches the cells or target site where an adverse occurs, or where that agent interacts with a membrane surface.

Exposure can be characterised in different ways, see figure 4. Often in the literature it is stated that ideally, the exact amount of a certain agent that an individual comes into contact with over his/her lifetime have to be estimated. This would include estimating the exposure of the agent of concern via different environmental media to include air, water, soil, food and house dust and exposure routes (9). However, this depends critically on the design and purpose of a study which exposure information is necessary e.g. for a time-series study only exposure data of the past days is necessary.

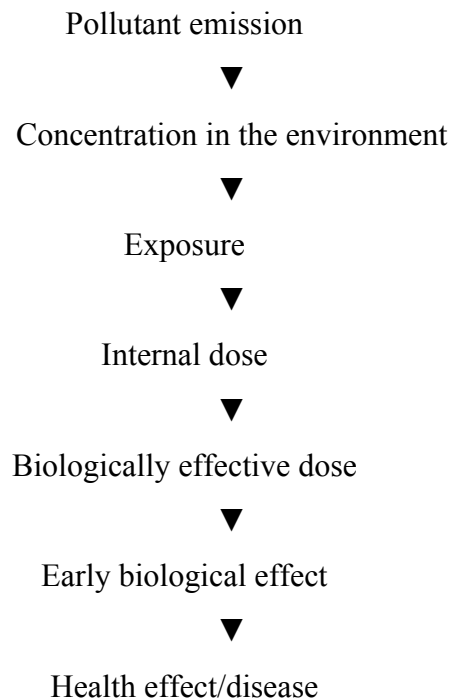


Figure 3 Steps between emissions of a pollutant in the environment and a potential health effect

Complete information on exposure is often lacking, a number of exposure assessment approaches are available for estimating personal exposure. In the case of airborne particles, direct measurements in the subjects' breathing zone, using personal monitors, are often considered the most accurate estimate of the subject's true exposure. In practise, however, it is not possible to measure personal exposure for large numbers of people. Epidemiological studies have generally relied on environmental measurements and modelling rather than on personal measurement and modelling. As environmental measurements and modelling do not take into account differences in activities, physiologies and routes of subjects, they may under- or overestimate personal exposure or uptake (10). Sometimes these environmental

measurements have been combined with personal (time-activity) data obtained with questionnaires to obtain more specific exposure metrics.

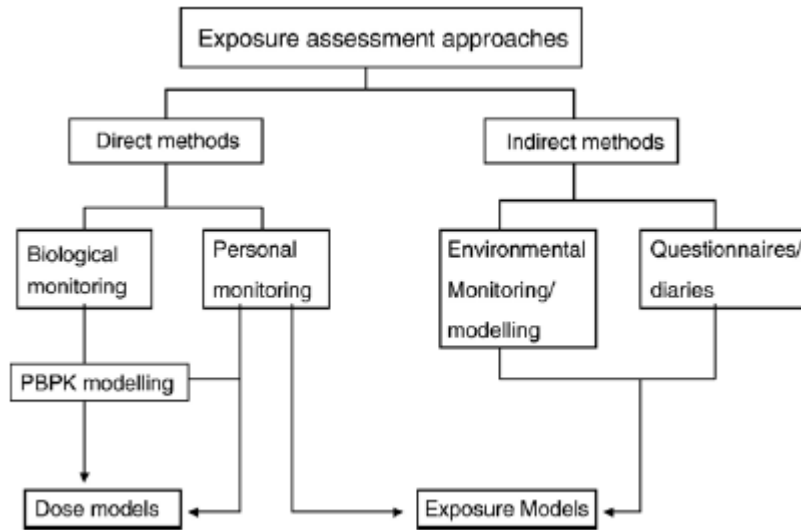


Figure 4 Exposure assessment approaches (10)

Most environmental measurements are generally obtained from stationary ambient monitoring stations that routinely measure background concentrations at a number of points in the area where the subjects live. Other environmental epidemiological studies have used simple proxies such as distance from a point source or emissions from roads to estimate exposure.


Certain areas within environmental epidemiology rely less on routinely collected data. For example, epidemiological studies of the effects of environmental tobacco smoking have often used questionnaires or biomonitoring to obtain exposure estimates. Biological monitoring is an important direct method of monitoring the internal dose. The magnitude of the biomarker accounts for bioavailability and is influenced by numerous parameters such as route of exposure, physiological characteristics of the receptor and the characteristics of the agent. Biomonitoring techniques can also be used to assess early biological or physiological changes which are correlated with the uptake of the agent. Biomonitoring of blood lead has been used frequently in studies on the effects of exposure to lead (9). Biomarkers can be linked to external exposures through physiologically based pharmacokinetics models, requiring an understanding of absorption, distribution, metabolism and excretion of a given substance (11).

In HIA, the same exposure metric should be used for the population exposure distribution as that used to derive the ERF to calculate e.g. the burden of a certain disease. It may however not

be possible to find or model data expressed in the same exposure metric. An example of this includes the use of different exposure metrics in epidemiological and toxicological studies. Epidemiological studies assess the exposure often as the concentration in the environment whereas in toxicological studies often the exposure is assessed as the internal dose.

Another example is the measurement of particulate matter air pollution as they can be expressed as PM₁₀, PM_{2.5} or BS. In some cases it may be possible to use conversion factors. Using assumptions about inhalation rates it may be possible to convert exposure to inhaled doses. However, generally these conversion factors are not constant across different populations, so it depends on the question whether conversion is appropriate. See WP 1.2 Source-exposure for more information on conversion of metrics.

Table 2 List of exposure metrics with respect to the true exposure of a fixed source agent (12)

Type of data	Approximation to actual exposure
1. Quantified personal exposure	Best type of data for estimating actual exposure
2. Quantified area measurements in the vicinity of the residence or sites of activity	
3. Quantified surrogates of exposure	
4. Distance from the site and duration of exposure	
5. Distance or duration of residence	
6. Residence or employment in the geographical area in reasonable proximity to the site where exposure can be assumed	
7. Residence or employment in a defined geographical area of the site	Worst type of data for estimating actual exposure

When there is enough appropriate literature available to derive an ERF for different exposure metrics assessing the same stressor, decisions have to be made about which one to choose. In table 2 a ranking is given of exposure metrics with respect to their ability to predict actual exposure. In principle one should use the ‘best approximation to the actual exposure’ available. However, this also depends on 1) the amount of individual studies, 2) the quality of the individual studies and 3) the representation of a mixture. Point 3 needs to be considered as in some particular cases the quantified proxy of a certain stressor is more informative for the expected health response than the measured estimate. Imagine the example of the proxy ‘traffic intensity’ compared to measured NO₂ personal exposure. There is some evidence that other components other than NO₂ in the emissions from motorized traffic are actually responsible for the observed health effects. Thus all the specific physical removal processes and time activity

patterns affecting the personal exposure to NO₂ may actually create more noise than make the exposure more specific and accurate. Traffic intensity captures not only the effects of air pollution and noise but also the possible combined exposure of air pollution and health effects. The key question is whether the ERF is of interest in your policy assessment, is it the ERF of NO₂, of proximity to traffic, or of traffic pollution using NO₂ as an indicator. The three ERFs are different, and in the last e.g. indoor or occupational sources of NO₂ just add noise to the assessment.

2.2.3 Different types of health effects

The impact of environmental exposures on human health can take numerous shapes of various severity and clinical significance. Often a hazardous exposure is associated with a spectrum of responses. This spectrum has been often characterized as a pyramid, based in the most common consequence—exposure—and having mortality, the least common and most severe consequence, at its tip (see figure 5).

While we will mostly deal with negative health effects, some stressors may be associated with positive effects on health. An example is the positive impact on many diseases including cardiovascular health of physical activity related to cycling or walking. Methods for deriving an ERF do not differ for positive and negative impacts.

Some effects occur soon after the onset of exposure; others emerge after long-term cumulative exposure, including a latency period. The public health significance of any response depends on many endogenous and exogenous factors. Whether or not an environmentally induced change affects individual health may be a function of its reversibility, individual possibilities of compensation or level of resilience (13).

Not all responses are adverse health effects. The American Thoracic Society provided some guidelines to make the distinction between adverse and non-adverse health effects clear. Although they stated that the actual decision on ‘where to draw the line’ to categorize a response as an adverse health effect or an action level between pathophysiology or physiologic change is probably best left to the individual or the community’ (14).

The US EPA (15) considered adverse health effect to be functional impairments or pathological lesions that may affect the performance of the whole organism or that reduce an organism’s ability to cope with an additional challenge, as has been defined by the Federal Register (16).

To deal with the diverging health impacts of various types of environmental exposures Murray and Lopez (17, 18) have developed an aggregated health impact indicator called disability

adjusted life years (DALYs). This health impact measure combines years of life lost and years lived with disability that are standardized by means of severity of the response. See WP 1.4 risk characterisation for more information about DALYs.

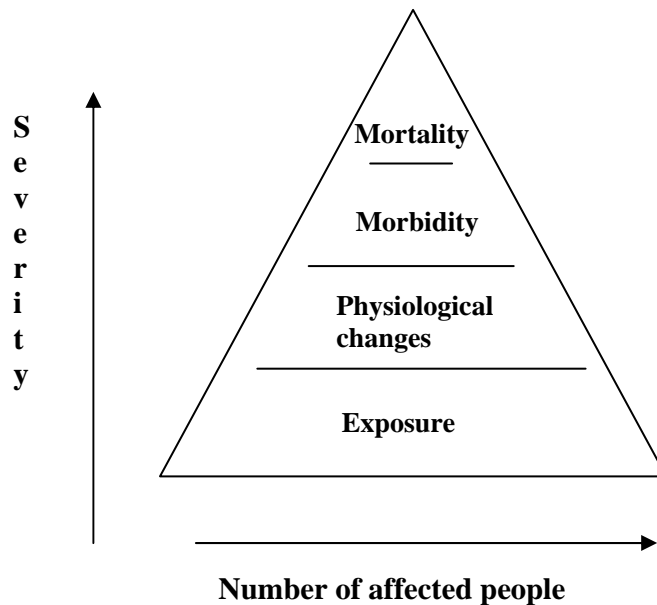


Figure 5 Health effect pyramid

Health effects may vary from acute and short-term effects to long-term chronic effects. In between, there is a whole range of sub-acute effects. On a cellular level, effects may already be noticeable, but it may take a certain degree of repetitive damage/accumulation of effects to induce observable health effects at the organism level. Studies which are focused on long-term health effects, such as cancer, must bridge the latency period that is incorporated in the pathologic pathway (19).

Different types of health effect include direct effects of an exposure and effects relayed through an intermediate variable, thus indirect effects. Next, health effects can be specific or non-specific for a given exposure. Mesothelioma is an example of a disease that is thought to be specific for exposure to crocidolite asbestos. These monocausal diseases are relatively simple to investigate. Most diseases, however, are of a multi-causal origin (19).

Another distinction is between reversible versus irreversible health effects. Sometimes health effects are classified based on the severity of the response; light, moderate and severe.

Often in HIA only those health effects are selected which have been studied comprehensively in epidemiological studies and for which clearly defined health outcomes have been established. Those health effects that are recorded in epidemiological studies, surveys and health care registration may only be the ‘tip of the iceberg’, with many minor impairments remaining beneath the surface undetected (13).

2.3 Final remarks and outline of the protocol

There is a large diversity of levels of exposures, different types of responses and relations possible. Can we make use of one general methodology to derive a certain ERF in spite of all these diversities? We regard this mostly as a problem of ‘issue framing’. Specifically, it is a problem of defining the ‘best’ exposure metric and response variables available. Consider as an example the issue of an exposure that acts through several media and pathways simultaneously, like the exposure of lead; most human exposure to lead occurs through ingestion or inhalation. To derive an ERF from the exposure to lead in air, and to use that to calculate numbers of people affected, makes certainly no sense. For this example it is more accurate to measure the concentration of lead in the blood, and use that as the exposure to derive a certain ERF.

In chapter 3 we deal with selection of exposure-response functions for further quantification. We make use of a short checklist that will guide you systematically through questions that need to be considered in order to derive an ERF. Chapter 4 reviews methods to derive a certain ERF including systematic review, meta-regression methods, Bayesian analysis and expert panel. These methods can be applied equally well to animal and human data, but if applied to animal data an additional step is necessary. This step is the extrapolation from animals to humans and is described in paragraph 4.4.

As those methods are time consuming, modified less-time consuming versions are discussed in paragraph 4.5. Chapter 5 reviews the methods which can be used to take into account uncertainty in ERFs (sensitivity analysis, multiple-bias modelling and the NUSAP approach). Chapter 6 discusses the integration of animal and human data in dose-response assessment (integration toxicology and epidemiology). Chapter 7 describes the general suggested methodology to be used in the SP3 first pass assessments. Chapter 8 describes the planned further work in our WP 1.3.

3 Selection of exposure response functions for quantification

The starting point is that an analysis has been made of which stressors potentially are important in a specific policy assessment. It is important that this analysis does not exclude implicitly stressors of which you believe few data exist. For a systematic approach it is best to start with a long list of potential exposures and associated health effects and then select a few that will be further quantified. This process is referred to as ‘issue framing’, an important step in an assessment. The first step is to describe which health effects have been associated with a selected stressor. Information can be collected both from epidemiological studies and toxicological studies (alternatively called human and animal studies). The second step involves a decision of which ERF will be quantified. This includes the following aspects: a) strength of evidence b) severity of the response c) number of affected subjects and d) logistic aspects. The last aspect deals with the resources available to select or develop ERFs. Paragraph 3.1 contains methods and an example for the first three aspects. Paragraph 3.2 discusses another aspect of selection, namely the issue of transferability of ERFs between different locations and populations. Paragraph 3.3 deals with the issue of combined exposures in the selection of ERFs.

For ERFs that will be quantified, we recommend that if there is already an existing ERF available, preferably from an authoritative and influential institute or organisation, like for example the WHO, one can better use that in the SP-3 policy assessments. Especially when there is much controversy on a certain topic, this is highly recommended. Of course the existing and available ERF have to be derived properly and in a systematic and transparent way and it has to be up-to-date as well. Thus, always check the literature to avoid missing an important and recently published study.

As transparency is identified as one of the main important things in HIA, box 1 below presents a checklist that needs to be considered in order to derive an ERF. To derive an ERF you can make use of different types of studies including epidemiological studies, human / volunteers experiments and toxicological studies. You can use this framework for each type of study separately.

Box 1 Checklist of factors that should be considered in deriving an ERF

Starting point is a certain chosen exposure

1. Which health endpoints have been associated with the exposure?
2. Which health endpoints are likely causally associated with the exposure?
3. Which ERF will be further quantified?
4. Is there a recent authoritative (systematic) review that can be used to characterize the ERF?
 - Evidence for a threshold?
 - Evidence for another mathematical function?
 - Quantitative relationship
5. Is it likely given the exposure levels and the ERF that a health response is non-zero?

THE NEXT QUESTIONS APPLY IF AN ERF NEEDS TO BE DEVELOPED

6. For which exposure metrics (e.g. external, internal dose) are studies available?
7. Is the linear mathematical model reasonable for the ERF?
8. Representation of exposure data in the individual studies
 - Choice of taking a specific averaging time
 - Which lag time?
9. For which population group is an ERF developed (general population or subgroups such as different age groups, children, elderly)
10. Are there differences in ERF related to geographical location?

3.1 Strength of evidence, severity and number of affected people

This second step involves a decision of which ERF will be quantified. This includes the following aspects described below; 1) strength of evidence (causality) and 2) severity of the response and 3) number of people affected. An example is given of the evaluation of the health impact of the extension of a large airport in the Netherlands including the health related responses to community noise exposure in box 2 and table 4.

Before considering these three aspects one has to think about the following as well:

1. Are people actually exposed to the agent of concern? (difference between hazard and risk; an agent can be dangerous but if nobody is exposed, there is no additional risk)
2. Is the additional exposure to humans relevant in comparison with for example the background exposure levels? Additional exposure can occur through:
 - different media
 - different pathways
 - from different sources

Appendix 2 provides a more detailed treatment of the issue of hazard identification.

Judgements about causality and to a lesser extent severity and number of affected people require a detailed assessment of the literature. This is likely not feasible in the selection. Therefore in practice one will make use of existing classifications such as the IARC classification of carcinogens, existing review articles or individual key epidemiological and toxicological studies of the topic supplemented with some expert judgements. This method of eliciting expert judgement is similar to the IARC methods, where more details can be found.

1. Strength of evidence

To assess the strength of evidence from both epidemiological and toxicological studies an often used method is the classification scheme of the International Agency for Research on Cancer (IARC). IARC classified potential carcinogens into the following categories:

Sufficient evidence: a causal relationship has been established between exposure and the health effect. That is a positive relationship has been observed between the exposure and cancer in studies in which chance, bias and confounding could be ruled out with reasonable confidence.

Limited evidence: a positive association has been observed between exposure and the health effect for which a causal interpretation is considered to be credible, but chance, bias, or confounding could not be ruled out with reasonable confidence.

Inadequate evidence: the available studies are of insufficient quality, consistency or statistically power to permit a conclusion regarding the presence or absence of a causal association between exposure and the health effect.

Evidence suggesting a lack of causality: several adequate studies covering the full range of levels of exposure that human beings are known to encounter, are mutually consistent in not showing a positive association between exposure and any studied endpoint at any observed level of exposure. Of course, the possibility of a very small risk at the levels of exposure studied can never be excluded.

Table 3 The Bradford Hill's causality criteria (20)

Criteria	Short explanation
1. Strength	A strong association is more likely to have a causal component than is a modest association
2. Consistency	A relationship is observed repeatedly
3. Specificity	A factor influences specifically a particular outcome or population
4. Temporality	The factor must precede the outcome it is assumed to affect
5. Dose-effect relationship	A dose-effect relationship is established (whether an increasing dosage or a longer duration of exposure is associated with more frequent or more serious health effects)
6. Plausibility	The observed association can be plausibly explained by substantive matter (e.g. biological) explanations
7. Coherence	A causal conclusion should not fundamentally contradict present substantive knowledge
8. Experimentation	Causation is more likely if evidence is based on randomised experiments
9. Analogy	For analogous exposures and outcomes an effect has already been shown

An important consideration is whether an association is causal or not in the decision process of which ERF will be quantified. Only for exposures and responses where sufficient evidence is available for a causal relationship an ERF can be quantified. However this importance of causality is debatable in the light of a HIA. The best working answer is that the entire set of ERFs to be used, combined with the rules for aggregation across pollutants (WP1.4 Risk

characterisation), should represent as well as possible the causal effect of the entire pollution mixture. In other words, causality is important, but it applies to the set as a whole and not necessarily to each individual pathway that will be quantified and included in the HIA (4).

The most widely recognized are the Hill's epidemiological causality criteria, appeared in a 1965 article on the causes of occupational diseases written by Austin Bradford Hill (20). Hill has provided nine considerations for assessing whether an observed association involved a causal component or not, see table 3. These criteria are generally not considered as checklist criteria that can be entered to give a firm yes / no response. Actually, most criteria are controversial, with the exception of temporality.

2. Severity of health responses

Sometimes health effects are classified based on the severity of the response; light, moderate and severe. A light response falls within normal biological variation, does not affect normal functioning or if it does, only temporarily. A severe response represents a serious handicap in everyday functioning, and is in general clinically significant. Moderate responses are somewhere in between, e.g. because they may be adverse to very susceptible individuals (13).

3. Number of people affected

The number of affected people within the exposed population is relevant as well. This is often a function of the distribution of the exposure and the susceptibility of the exposed. One can use three categories: a) highly exposed and/or susceptible individuals, b) specific sub-groups, such as patients, elder people with frail health, people with certain professions, inhabitants of deprived and heavily exposed areas and c) substantial part the total population (13).

Items number 2 and 3 are also key components of the calculation of DALYs, see WP1.4 Risk characterisation.

Box 2 Example of the evaluation of the health impact of the extension of a large airport (13)

Noise, smell, risk

It is well established that as levels of noise exposure increase a higher percentage of any representative population will report to be severely annoyed. No doubt this presents an important social problem. Whether this is a health problem as well is less obvious. It is a fact that several studies have found an association between serious annoyance and highly correlated social responses, such as perceived stress, anxiety, and risk perception on the one hand and reported (psychological) symptoms, cognitive complaints, (self-) medication and use of health services on the other. The causality of this association however is far from obvious. It would potentially be confounded by many interacting factors, which may explain the contradictory results seen over all studies. Furthermore, the attitude towards the source of noise, and sensitivity to noise accounts for more variation than the level of noise exposure by itself. People who report themselves seriously annoyed might just tend to report symptoms, use self-medication or even visit a GP, irrespectively of actual noise levels.

The case for stress-related responses, such as elevated diastolic blood pressure, stress hormones, causing prolonged hypertension or hypercholesterolemia, attributing to the prevalence of cardiovascular disease at the level of populations is far from strong. However, it is conceivable these types of acute haemodynamic responses may contribute in one way or another to the onset of acute cardiac infarctions (as many other stimuli may).

There is not much doubt that nocturnal noise may reduce the quality of sleep, change sleep patterns (including awakenings, EEG's) and may lead to chronic loss of sleep in the end. Poor quality of sleep affects daily mood, and performance, among other things. There is much uncertainty on the ability of residents to habituate to noise and the long term consequences of any particular degree of noise induced sleep disturbance. Excessive sleep disturbance will eventually compromise social-psychological well being, but until now the amount of sleep loss required to adversely affect health cannot be defined properly.

In a recent study in the Netherlands indications were found that the use of psychotropic and cardiovascular medication was associated with the level of noise exposure. The results are consistent with those of other cross-sectional studies (and one cohort-study). The impact on medication use is consistent with observed social responses, such as increased levels of stress, concern about health and mortality risks, and noise and smell annoyance.

There is no convincing evidence for a direct effect of exposure to noise, smell, vibrations or risk on health outcomes such as congenital abnormalities, birth weight, or disorders related to the immune system (infectious or auto-immune disease). So far plausible mechanisms of action for these disorders are lacking.

The indications that noise would contribute to the prevalence of cardiovascular and psychiatric diseases are inconclusive. However, with respect to these disorders the evidence is not strong enough to reject the hypothesis noise is in some way involved in the multi-causal process leading to disease.

Table 4 Overview of reported health related responses to community noise exposure (13)

Response	Number affected¹	Clinical significance²	Evidence³
<i>Social responses</i>			
Annoyance	***	*	***
Performance	**	*	**
Symptoms (anxiety, depression)	**	*	**
Perceived health	**	*	**
<i>Sleep</i>			
Sleep pattern	**	*	**
Disturbance, awakenings, loss	**	*	**
Reported quality	**	*	***
Mood, performance	*	**	*
<i>Stress related responses</i>			
Blood pressure	**	*	**
Pulse rate, arrhythmia	**	**	**
Hypertension	**	**	**
Serum lipoprotein composition	**	**	*
Cardiovascular disorders	*	***	*
Ischaemic heart events	*	***	**
Psychiatric disease	*	***	*
Immune system	**	*	*
Stress-hormones	**	*	*
Birth weight	*	**	*
Congenital abnormalities	*	***	-

1. * = highly exposed and/or susceptible individuals, ** = specific sub-groups, *** = substantial part of the total population

2. * = clearly within normal biological variation, ** possibly adverse to susceptible individuals, *** clinically relevant

3. * = inadequate, inconclusive evidence, ** = limited evidence, *** = sufficient evidence, - = lack of evidence

3.2 Location / population selection

In principal the ERFs to be used should be representative of conditions in the target population of the HIA. The issue of transferability of the ERF should be considered. Basically transferability contains two aspects: 1) locations and 2) populations.

Probably the most important issue is the application of certain ERFs derived from typically multiple studies to specific locations. In general it is not recommended to use exclusively results from one study conducted in the local area for which you want to assess the health responses. Between different locations, systematic reviews and meta-analyses have the greatest credibility and give more precise uncertainty estimates than individual studies. The trade-off between meta-analyses and individual studies arises when the HIA is being undertaken in a local area where relevant studies are available. For example: you want to perform a HIA on the health effects of a certain exposure in children in Rome, Italy and one of the many studies have been conducted in Rome. We know from multi-centre studies that different cities can give different estimated ERFs. This may partly reflect real differences between cities. It may also reflect how chance variations affect results in individual studies. They affect meta-analyses too, but less so, because of their greater statistical power. Thus even when local results are available, the weight of evidence from the international literature should have a strong influence on the selection and use of an ERF. This suggests a weighted mean of the local-city value and for example the general meta-analysis value. Such an approach has been recently developed by APHEIS-3 (21) using the statistical concept of a shrunken estimate (4).

Alternatively, if a meta-analysis has revealed specific factors that explain why ERFs vary between locations, this information could be used to assign a more specific ERF rather than the overall average. For example the APHEA-2 study has discovered several factors affecting the acute effects of PM₁₀ on mortality (temperature, NO₂ concentration). It is important to realize that a meta-analysis should not be considered as a set of rules to come up with always one overall average, but rather as a tool to study between individual studies what characterizes the response to a certain exposure.

Finally, it is useful to put the issue of transferability between locations in perspective. If we accept that it is possible to extrapolate from animal data to human data, the uncertainty of extrapolating to another location is likely much smaller.

While substantial differences between different countries in terms of susceptibility to a certain stressor like air pollution seem unlikely, there are factors that may affect transferability such as:

differences in daily patterns of activity, climatic conditions, housing etc, that would result in different exposures/doses from the same ambient concentration; differences in the pollution mixture; different importance of confounding factors that might not have been properly controlled for in the epidemiological studies; different techniques in air pollution concentration measurement, and others (4, 22, 23).

The issue of the application of ERFs to populations other than the study population from which the ERFs have been derived, is in most cases of little concern; if there is an indication of a different distribution of effects between subpopulations then there is often enough international literature available that investigates specific subgroups and age groups. Then it is possible to derive different ERFs for different subgroups and age groups in a population (4).

3.3 Combined exposures

In general, methods which have been used to assess a certain ERF are based on the selection of only one exposure and one associated health effect. However, in the real world, people are often exposed to multiple exposures simultaneously rather than one single component.

Multiple exposures include exposure to 1) different chemicals and 2) chemicals and non-chemical factors, simultaneously (air pollution and noise) or not. Exposure to one pollutant can affect the response of another exposure later in time as well.

It is not easy to describe effects of mixtures, as the data are often lacking for adequate description. In this section we give a short overview of both toxicological and epidemiological approaches to combined exposures. WP 1.5 Cross cutting issues has already made a draft on the issue of combined exposure, and will work on the issue of combined exposures further (24).

Although toxicological research on the effects of chemical mixtures has been initiated, the combination of chemicals and non-chemical factors has received little attention. The simplest way to study the effect of mixtures is to compare the effect of a mixture with the effects of all its constituents at comparable concentrations and duration of exposure at one dose level without testing all possible combinations of chemicals. This strategy has been used to assess the combined effects of mixtures like for example pesticides and cigarette smoking (25). However in this approach it is not able to test all potential combinations for every pair of chemicals. Any time a mixture consists of more than two compounds many two or three factor interactions are

likely. In a complete experimental design the number of possible tests combinations increases exponentially with increasing numbers of compounds in a mixture. Likewise, the number of experimental groups will also increase with the number of doses of each compound. Therefore it is virtually impossible to perform these kinds of testing strategies both from a economic, pragmatic as well as an ethical point of view (26).

Bliss et al. (27) identified already in 1939, three basic concepts of joint action or interaction of combinations of chemicals:

I Simple similar action (concentration/dose addition)

Simple similar action or concentration/dose addition of chemicals is a non-interactive process, because each of the chemicals in the mixture acts in the same way, by the same mechanisms, and differ only in their potencies. Similar joint action allows the additive effect to be described mathematically, using the summation of the doses. The use of the relative potency factors of which toxicity equivalence factors (TEF) are a special case is based on dose addition. This approach can be used for a well-defined class of agents that operate through a common mode of action for the same health outcome. Examples of this approach are the relative potency factors for some carcinogenic polycyclic aromatic hydrocarbons and the TEF for dioxin-like compounds.

II Simple dissimilar action (effect/response addition)

The modes of action and possibly the nature and site of action differ among the chemicals in the mixture which exert their individual effects, but do not modulate the effect of other compounds in the mixture. Effect addition is the additive effect determined by the summation of the effects to each compound in the mixture.

III Interactions

It describes the combined effect between two chemicals resulting in a stronger effect or weaker effect than expected on the basis of either dose or response additivity. A famous example of a

synergistic effect is asbestos and cigarette smoking on lung cancer. An example of an antagonistic effect is the exposure to dioxins and flavonoids and cancer.

These basic concepts are reflected in epidemiological approaches to mixtures and to combined effects of environmental agents as well, see table 5.

Table 5 Epidemiological approaches to mixtures and to combined effects of environmental agents (28)

Treating the mixture as a single agent

Select an indicator compound

Create a summary index

Separating effects of the mixture's components

Characterize the independent and joint actions of the components (interactions)

Treating the mixture as a single agent is common practise in epidemiology. Example exposures include cigarette smoking and diesel exhaust. For example, the acute and chronic diseases were associated with active cigarette smoking using questionnaires that characterized the duration of smoking and the usual number of cigarettes smoked on a daily basis. Causal links of cigarette smoking and several adverse health effects were made long before the toxicology of individual components of cigarette smoking was fully understood and before relevant animal studies have been developed (28).

Another method is to select an indicator compound, which characterise a certain mixture. Examples are the use of ozone as an indicator for photochemical smog; the use of PM₁₀ for the wide variety of particulate matter air pollution; the use of black smoke (BS) to assess coal combustion or more recently diesel vehicles; the use of NO₂ to assess traffic-related air pollution. The indicator is then used to represent the mixture. It is probably fair to state that in many cases the exposure variable is an indicator of a mixture rather than the causal component itself in epidemiological studies. An important issue with this method is whether the indicator component represents the mixture equally well in different settings. This is a topic that is addressed in an analysis of heterogeneity in systematic reviews (paragraph 4.1).

This approach can be refined by taking into account two or more indicators of the mixture. An example would be to use PM₁₀, NO₂ and O₃ as indicators of general air pollution instead of PM₁₀ only. When you take this approach, then the problem of ‘double-counting’ can play a role, which should of course be avoided or accounted for.

The concentration of multiple components may be measured and combined into a summary index. This approach might be employed if measurement of multiple components is feasible and a single component reflective of the mixture’s toxicity cannot be identified. Some investigators of the health effects of volatile organic compounds (VOCs) have followed this approach. This approach is not widely used in epidemiology.

Some mixtures have been approached as if the components had independent and separable effects. This approach has been applied to mixtures for which individual components can be measured and then regression methods applied in an attempt to characterize ‘independent’ effects’ of the components. It is widely applied, particularly in investigating the health effects of air pollution mixtures generated by combustion of fossil fuels.

Another method is to account for important known interaction effects within a certain mixture and to allow for that in the ERF. Interaction can be tested using statistical methods. In the analysis of studies, the method of meta-regression has been used extensively, see paragraph 4.1.

Finally, biomarkers of (early, physiological) effect have been used to integrate effects of different components in a mixture.

4. Review of methods to derive ERFs

In this chapter we briefly describe the two main methods to derive ERFs including the methods of systematic review (including if appropriate meta-analysis) and the methods of an expert panel. Some subparagraphs are added about the use of meta-regression methods and Bayesian meta-analysis. In paragraph 4.3 a comparison between these methods has been made.

Upon request we have some tools/programs available for an easy use of these systematic review methods.

We focus on the use of epidemiological studies for the systematic review. These methods will generally be applicable to animal studies as well. We know that systematic reviews and meta-analyses of animal studies are uncommon (29) and that the current toxicological risk assessment bases its judgment often on one or two studies, which are selected to be the most relevant or with the ‘worst’ estimate, discarding the evidence from other studies (30). Although uncommon, the use of these methods to evaluate animal evidence has increased over time (31) and we think that systematic review is the most appropriate method for the evaluation of the literature for exposure-response assessment.

When there are only animal studies available for the exposure-response assessment, an extra step is necessary; the extrapolation from animal to human evidence (paragraph 4.4).

In paragraph 4.5 some informal methods to derive an ERF has been discussed. These could be used if resources do not allow the above mentioned methods. These could be used to get a first estimate of the expected health responses. If it turns out that the exact value of the ERF is very consequential for the final assessment, one may still decide to use the more formal methods of systematic review / expert panel.

4.1 Systematic review

A number of terms are used concurrently to describe the process of systematically reviewing and integrating research evidence, including systematic review, meta-analysis and pooled analysis.

A systematic review is a review that has been prepared using a systematic approach to minimise biases and random errors which is documented in a method section (32). A systematic review may include a meta-analysis: a statistical analysis of the results from independent studies, which generally aims to produce a single summary estimate (32). Actually, a meta-analysis is a systematic review including a quantitative summary estimate. This distinction between a systematic review and meta-analysis is important because it is always appropriate to systematically review individual studies, but it may sometimes be inappropriate, or even misleading, to statistically pool results from separate studies (32).

Originally, systematic review was developed for analysis of controlled clinical trials and several authors have expressed major concerns in applying meta-analysis to observational studies. One reason for this is that potential biases make the calculation of a single summary estimate of effect of exposure potentially misleading. Similarly, the extreme diversity of study designs and populations in epidemiological studies may make the interpretation of summary estimates problematic. In addition, methodological issues related specifically to meta-analysis, such as publication bias, could have particular impact when combining results of different studies. Criticism refer also to the subjectivity involved in selecting only studies which are of a certain quality (33, 34).

A pooled analysis is a re-analysis of individual data based on primary studies instead of a quantitative summary of published data (meta-analysis). Thus publication bias may be reduced in comparison with a meta-analysis. Harmonization of study methods is another major advantage. However, a major obstacle for the use of pooling, is the fact that it is more expensive, time-consuming and requires close cooperation between study coordinators (32).

In this protocol we concentrate on the use of a systematic review to derive a certain ERF for the policy assessments as pooling of individual data is in principal preferable but not feasible in the scope of INTARESE.

It is very important to conduct the systematic review in an as transparent and reproducible way as possible. However, there is only limited methodology on these reviews available, especially

for observational studies (33). In table 6 the general steps in conducting a systematic review are described. They focus on systematic reviews of controlled trials but they are also applicable to reviews of any type of study, including toxicological studies (32).

4.1.1 Systematic review to derive ERFs

For examining the scientific evidence and selecting valid studies for the review, an advisory report of the WHO on evaluating epidemiological evidence provides useful guidance for this paper (3). These WHO guidelines were also used as a basis for the WHO project ‘Systematic review of health aspects of air pollution in Europe’. One of the products of this project was a quantitative meta-analysis of peer reviewed studies to obtain summary estimates for certain short-term health effects linked to the exposure to PM and O₃, see also appendix 3 (35). We can use this as an example of how to conduct a systematic review to derive ERFs in an appropriate way.

In the section below we summarize the main issues related to the steps above. For a more detailed description of systematic reviews and meta-analysis we refer to the literature.

Again note that we only want to make use of this method when there is sufficient evidence for causal relationship.

Table 6 Steps in conducting a systematic review, partly based on Egger et al. (32)

Steps	Issues
1. Formulate review question	
2. Definition of inclusion / exclusion criteria	<ul style="list-style-type: none"> • Population of interest • Exposure • Health outcomes • Study design and methodological quality
3. Location of studies	<ul style="list-style-type: none"> • Bibliographic databases like Medline, Pubmed, Embase and Web of Science • Reference tracking • Hand searching of key journals • Personal communication with experts in the field
4. Selection of studies	<ul style="list-style-type: none"> • Have eligibility checked by more than one observer • Develop strategy to resolve disagreement • Keep log of excluded studies, with reasons for exclusion
5. Quality assessment of studies	<ul style="list-style-type: none"> • Consider assessment by more than one observer • Use simple checklists rather than quality scales • Consider blinding of observers to authors, institutions and journals
6. Data extraction	<ul style="list-style-type: none"> • Design and pilot data extraction form • Consider data extraction by more than one observer • Consider blinding of observers to authors, institutions and journals
7. Analyse results (summary estimate)	<ul style="list-style-type: none"> • Tabulate results from individual studies • Examine forest plot • Explore possible sources of heterogeneity • Consider a meta-analysis of all studies/subgroups • Perform sensitivity analyses
8. Interpret results	<ul style="list-style-type: none"> • Consider limitations, including publication and related biases • Consider strength of evidence • Consider applicability and recommendations

1. Formulate review question

At the beginning of every study, formulation of detailed objectives is very important.

These objectives should include the definition of population of interest, health outcomes and exposures.

A systematic review needs a detailed study protocol. The protocol should include a clear description of the aim of study, the questions to be addressed, the methods and criteria to be employed for identifying and selecting relevant studies. Also define already potential sources of

heterogeneity and appropriate subgroup analyses. This protocol should be written in advance (32).

2. Definition of inclusion/ exclusion criteria

To increase the uniformity and homogeneity of the epidemiological studies and to ensure comparability, decisions have to be made which in/exclusion criteria should be used.

Decisions have to be made about the type of studies to include, the quality of the studies and to the combinability of participants etc.

Although some individual epidemiological studies may investigate the same exposure and the same health outcome, the results can be rather different from each other. This variability in results could be a function of site-specific differences, analytical decisions or simply random variation. Studies are often set in different geographical regions. These study sites are different in their demographic compositions, and baseline mortality rates and patterns. Often variability in results might be the result of differences in the pollution mixture or the use of different measurement methods. Variability between studies could occur due to the use of different statistical models including the use of a one pollutant model or a multiple pollutant model etc.

For the meta-analysis, you want to calculate from each individual study an estimated percentage of change in a predefined health outcome per a certain increase in a pollutant, together with a confidence interval. In general, you should exclude studies if they lack this kind of data.

The most appropriate way of handling the selection of studies is to include all studies that meet the basic entry criteria; thus using rather broad inclusion criteria and then perform sensitivity analysis with regard to the possible other in- and exclusion criteria. Any conclusions from a meta-analysis that are highly sensitive to altering the entry criteria should be treated with caution (36).

3. Location of studies

In your study protocol it is important to think about an appropriate search strategy, including the use of key terms. At least the following sources needs to be considered:

a) electronic databases, b) check reference lists of located studies and c) perform some hand searching of key journals.

The inclusion of unpublished results in the systematic review is debatable. On the one hand unpublished results can be of poor quality and you do not want to include them. On the other hand if the reasons that studies remain unpublished are associated with their outcome then the results of the meta-analysis could be seriously biased and you want to include those studies as well (36).

For practical purposes, often the meta-analysis is restricted to the English language. Authors might be more likely to report positive findings in an international English language journal and negative findings in a local journal. Bias could thus be introduced (37). Also consider the possibility of database bias. Studies that are published in journals not indexed in one of the major databases are often hidden from the reviewers. Studies with significant results might be more likely to be published in an indexed journal, whereas negative studies are published in non-indexed journals (36).

4. Selection of studies

It is important to be transparent in which studies are excluded and which studies are selected for further analyses. Make a list of the excluded studies together with the reason for exclusion (e.g. insufficient control for major confounders).

Decisions regarding the inclusion or exclusion of individual studies often involve some degree of subjectivity. It is therefore useful to have two observers checking eligibility of candidate studies. Disagreements must be resolved by discussion or with a third reviewer (32).

5. Quality assessment of studies

A large number of quality assessment scales is available. These quality scores vary considerably in terms of dimensions covered, size and complexity. Use of quality scoring in systematic reviews of observational studies is controversial, as it is for clinical trials, because scores constructed in an ad hoc fashion may lack demonstrated validity, and various scores may give different results in the meta-analysis. Reducing the features of a set of studies to a single measure of quality is not recommended (32, 38-40). It is preferable to assess the

methodological characteristics of the individual studies case-by-case, by the use of a simple quality checklist (32).

Of course it makes sense to take into account the quality of studies. Studies should be excluded when they have serious deficiencies in methodology, design or analysis. The possible influence of study quality on the summary effect estimates can be examined by sensitivity analysis as well (32).

6. Data extraction

Data to be extracted from individual studies for a meta-analysis include the effect estimate, an estimate of its variance and descriptive information about the study, such as country, design, population, sample size, exposure assessment, lag times, valid range, time frame and control for confounding.

Preferably, extract the effect estimates adjusted for confounder variables from the individual studies instead of the unadjusted effect estimates (41).

Studies presented results in several forms such as a relative risk, odds ratio, the percent increase in a health outcome each corresponding to a specified increase in a pollutant concentration. Alternatively the slope from the regression model is reported. Studies might differ in the way they express their exposure, for example they have considered numerous exposure periods, including 1-hr maximum, 8-hr average, or a 24-hr average concentration. In order to make results comparable these estimates have to be converted into a standard metric: percentage of change in a predefined health outcome per (e.g. $10 \mu\text{g}/\text{m}^3$) of a (24-hr, or 8-hr) average increase of a pollutant, together with a confidence interval.

Time series studies often report results for a number of pollutant lags or day prior to the health event. For the meta-analysis, when estimates for multiple lags are provided for an individual study, the estimate for lag 0 can be used. This approach minimizes the bias of choosing the lag with the largest effect, although some studies only presented results for a single often most significant lag (35, 42). Also for other kind of studies, lag time and time frame in general is an important issue to consider when extracting certain estimates for the meta-analysis.

Special attention should be given to the use of single pollutant or multi-pollutant models in the individual studies. In the WHO meta-analysis of Anderson et al. (35) their initial analysis was

focused upon single-pollutant models. Individual studies which only presented results from multi-pollutant models the results were only extracted from the model with the most pollutants in it.

A number of studies have been published more than once. Studies may be updated or reanalyzed several times. For the meta-analysis it is preferable to extract effect estimates only from the most recent study published.

7. Analyze results (summary estimate)

If a meta-analysis is appropriate there are two different models which have been used to obtain a summary estimate, a fixed and a random effect model, respectively. In the fixed effect model it is assumed that the ‘true’ effect is the same in each study, and that the results of the individual studies differ only because of chance. In random effect meta-analysis the effects for the individual studies are assumed to vary around some overall average effect, which has a normal distribution, characterized by an overall mean and a standard deviation (32).

In standard meta-analysis effect estimates of individual trials are weighted by the inverse of their variance. The larger the study, the smaller the variance of the effect estimate, and the greater the weight the study receives in meta-analysis. For the simple formulas see below, based on the article of DerSimonian (43).

Fixed effect summary:

$$B_{\text{summary}} = \sum(w_i * B_i) / \sum w_i$$

- w_i = weight for study i

- B_i = effect estimate study i

$$w_i = 1 / se_i^2$$

- se = standard error

$$SE_{\text{summary}} = 1 / \text{sqrt}(\sum w_i)$$

Test for significance: $B_{\text{summary}} / SE_{\text{summary}}$

Has t distribution

Random effect summary:

$$B_{\text{summary}} = \sum(w_i * B_i) / \sum w_i$$

- w_i = weight for study i

- B_i = effect estimate study i

$$w_i = 1 / (se_i^2 + \sigma^2)$$

- se = standard error

- σ^2 = between study variance

Test for homogeneity:

$$Q = \sum(w_i * (B_i - B_{\text{summary}})^2)$$

Has a Chi-square distribution with n-1 degrees of freedom

n = number of studies

A formal statistical test of homogeneity, usually called a test for heterogeneity (see above), is often used to decide which of the two models is more appropriate for a particular meta-analysis. Use a fixed effect model when no heterogeneity could be detected; otherwise use a random effect model. However, this test is generally thought to be of low sensitivity and has low statistical power to detect heterogeneity (32, 44). Heterogeneity should also be assessed graphically by the use of for example a forest plot. A forest plot is a graphical display of results from individual studies on a common scale, which allows a visual examination of the degree of heterogeneity between studies (45).

Meta-analysis should incorporate a thorough consideration of possible sources of heterogeneity between individual studies. For this purpose, meta-regression can be used, which investigate whether a particular covariate or characteristic of the individual studies, is associated with the sizes of effect observed in the studies (see paragraph 4.1.2). Another way to investigate possible sources of heterogeneity is to make use of subgroup analyses. In these analyses, you stratify the individual studies with respect to study characteristics and calculate summary estimates per subgroup. Subgroup analyses are, in other words, meta-analyses on subgroups of the studies (32).

Influence and robustness can be assessed in sensitivity analyses by repeating the meta-analysis on subsets of the original dataset. The influence of each study can be estimated by deleting each individual study from the analysis and noting the change in the summary estimate and the confidence interval. Other sensitivity analyses can assess robustness of summary estimates by varying for example some in- and exclusion criteria (32).

The approach to obtain summary estimates, described above, is a frequentists approach. Another approach uses Bayes's theorem, named after an 18th century English clergyman (46). (see paragraph 4.1.3).

Excel sheets for an easy use of these methods are available upon request, mail to j.m.c.boogaard@iras.uu.nl.

A meta-analysis is only appropriate when the results of the individual studies be expressed as an “effect measure”, which has a numerical value and can be combined into a single estimate. More importantly, it should make sense to combine the results of the different studies into a single estimate. The individual studies have to be sufficiently similar in respect with the population studied, the measured exposure and the measured health effect.

8. Interpret results

There are numerous ways in which bias can be introduced in reviews and meta-analysis. Of course if the methodological quality of the individual studies is inadequate then the findings of the systematic review may also be compromised (garbage in garbage out).

The presence of publication bias is one of the major concerns in conducting a systematic review. Publication bias occurs when studies showing evidence for associations in a particular direction are selectively published. If present, publication bias might lead to the adoption of a false hypothesis, or to an estimate of a true effect that is biased away from the null (47, 48).

To detect publication bias in the systematic review, a funnel plot is often used. It is a simple scatterplot of the study effect against the sample size. The funnel plot is based on the fact that precision in estimating the effect will increase as the sample size of component studies increases. Results from small studies will scatter widely at the bottom of the graph. The spread will narrow as precision increases among larger studies. In the absence of bias, the plot should thus resemble a symmetrical inverted funnel. If the plot shows an asymmetrical and skewed shape, bias may be present. This usually takes the form of a gap in the wide part of the funnel, which indicates the absence of small studies showing no effect (49).

In principal, publication bias should be:

- minimized, by doing a comprehensive literature search (e.g. including, if possible, unpublished results)
- detected, by funnel plots and statistical tests like the Egger’s linear regression test (49) or the Begg’s test (50)
- corrected, by statistical models in which missing data are imputed, for example by use of the trim-and-fill method (51)

- assessed by sensitivity analysis

Other reporting biases can play a role in affecting summary estimates as well. One could think about time lag bias, multiple publication bias, citation bias, language bias, outcome reporting bias and biased inclusion criteria (32).

4.1.2 Meta-regression methods

Meta-regression is an extension to meta-analysis, and thus, is also a meta-analytic method. It is a generalization of subgroup analyses, which can be used to investigate heterogeneity of effects between individual studies. It examines the relationship between one or more study-level characteristics and the sizes of effect observed in the studies. Characteristics of studies might be, for example, a particular study design, size of the study, certain exposure characteristics, control for a specific confounder, geographic region etc. (52).

Understanding the possible causes of any heterogeneity (even if an initial overall test for heterogeneity is non-significant) is maybe as important as the meta-analysis itself, and therefore we included this as a separate paragraph in the protocol.

The use of meta-regression has certain drawbacks. First, ecological fallacies ensue when summary data for a group misrepresent the individual participants. This may be a problem for covariates with different values for each participant within a group (53).

Secondly, exploring sources of heterogeneity by the use of meta-regression may result in false positive conclusions through ‘data dredging’. The only way to protect against this is to prespecify which covariates (including a scientific rationale!) are going to be investigated by meta-regression in order to minimise spurious findings and restrict the number of covariates to a minimum. Also meta-regression, as other observational studies, can be prone to bias, notably bias by confounding; an association identified with one covariate may in reality reflect a true association with other correlated characteristics (54).

4.1.3 Meta-regression methods to derive ERFs

To provide a quantitative summary of results and to investigate potential effect modifiers, it is possible to apply univariate or multivariate meta-regression models. In such models, fixed-

effects pooled regression coefficients are estimated by weighted regression of individual estimates on potential effect modifiers with weights inversely proportional to their individual variances. If significant heterogeneity among estimates remains beyond the variation associated with fixed effects, random-effects meta-regression models can be applied. A number of variables characterizing the individual estimates may be considered as potential effect modifiers.

In univariate meta-regression, the outcomes, for example pollutants, are treated as being independent and study (city) specific effect estimates β^c are pooled in a common estimate under the assumption that they are normally distributed around an overall mean. More specifically, fixed effects pooled regression coefficients are estimated by weighted regression of study-specific estimates on potential effect modifiers (at study level) with weights inversely proportional to the study-specific variances (43). If substantial heterogeneity remains among study-specific estimates beyond the variation associated with the effect modifiers, then random effects regression models are applied. In these models, it is assumed that the individual coefficients are a sample of independent observations from the normal distribution with mean equal to the random effects pooled estimate and variance equal to the between-studies variance. The random variance component may be estimated by reweighted least squares (55). Models for multivariate meta-regression as proposed by Berkey et al (56) and implemented in the APHEA project (57) are of the form:

$$\beta^c = X_i^c a_i + \delta^c + \varepsilon^c$$

where β^c is the vector of the estimates of interest that measures the log-relative rate of health outcome for a unit of increase in exposure in study/dataset (eg one city) c for the p environmental factors; X_i^c is a matrix containing the observed study-level covariates for study c ; its pattern of entries indicates what exposure's effect appears in each row of β^c and to which exposure effects each covariates relates, a_i is the vector of regression coefficients to estimate-it may include a separate intercept and slope for each exposure against each corresponding covariate; δ^c is a vector of p random effects associated with study/dataset c representing, for each environmental factor, the study's deviation from the average having the same values of covariates, and ε^c (assumed independent from δ^c) is the vector of random sampling errors within each study.

The $p \times p$ matrix $\text{cov}(\delta^c) = D$ (needed to be estimated) represents the between-study covariance that is unexplained by the fixed effects (*i.e.* the regression). It is assumed that:

$$\begin{aligned}\delta^c &\sim \text{MVN}(0, D) \\ e^c &\sim \text{MVN}(0, S^c) \\ \beta^c &\sim \text{MVN}(X^c \alpha, D + S^c)\end{aligned}$$

where S^c is the estimated variance-covariance matrix in study c for the p exposures. When $D \approx 0$ we have the corresponding fixed effects estimates while when $D \neq 0$ we have the random-effects estimates. Ignoring the correlation among pollutants in each study (city) implies that the off-diagonal elements of D and S^c are set equal to zero. That is equivalent to treating results from multi-pollutant models as if they were produced from single-pollutant models.

To estimate the model parameters, the method described by Berkey et al. (56) may be applied. In contrary to the usual meta-regression in which results from each environmental factor are analyzed separately, the multivariate meta-regression provides more accurate estimates by incorporating the correlation among different exposures within each study.

Summarizing out, univariate meta-regression models represent a specific case of the multivariate meta-regression models where $p=1$ (single exposure study-specific models). In such cases, the between variance is estimated from the data using the maximum likelihood method described by Berkey et al. (55) and added to the specific variances. Such univariate models can be easily fitted in any standard statistical software by weighted least squares regression. Multivariate meta-regression models can be fitted using programs that have been developed in S-Plus and R. These programs are available upon request (mail to xpedeli@med.uoa.gr).

Results from simulation studies (56), submitted APHEA paper) have shown that, in the meta-analysis of multiple correlated factors, the application of multiple univariate models (*i.e.* one for each factor) or that of a single multivariate meta-regression model (simultaneous meta-analysis of all factors) provide very similar results. Hence, the extra effort to fit the multivariate model may not be worthwhile when pooling effects of two pollutants. However, it may be worthwhile when pooling effects of more than two pollutants (57).

4.1.4 Bayesian analysis

Most researchers first meet concepts of statistics through the frequentists way of thinking.

Bayesian statistics offers us an alternative to the frequentists methods. In this chapter we will present to reader basic Bayesian ideas.

Bayesian thinking and modelling is based on probability distributions. Very basic concepts in Bayesian analysis are prior and posterior distribution. A prior distribution $p(\theta)$ summarizes our existing knowledge on θ (before data is seen). It can for example describe an opinion of a specialist and therefore Bayesian modelling requires that we accept the concept of subjective probability. A posterior distribution $p(\theta|data)$ describes our updated knowledge after we have seen data. A posterior is formed by combining a prior and likelihood $p(data|\theta)$ (derived using same techniques as in the frequentists statistic) using Bayes' formula,

$$p(\theta|data) = p(data|\theta) p(\theta) / p(data)$$

$p(data)$ is the prior or marginal probability of B, and acts as a normalizing constant.

As we see this is a natural mechanism for learning, it gives a direct answer to the question: "How does data change our belief of matter we are studying?"

From above we also see that one of the main differences between the frequentists and Bayesian analysis lies in whether we use only likelihood or whether we also use prior distribution. Prior distribution allows us to make use of information from earlier studies. We summarize this information with our prior distribution and then use Bayes' formula and our own data to update our knowledge. If we do not have any previous information on issue we are studying we may use so called uninformative prior meaning that our prior distribution does not contain much information, for example normal distribution with large variance. As we see from Bayes' formula use of uninformative prior lets data define our posterior distribution.

Of course there are also differences between Bayesian and frequentists statistics.

One is the way of thinking. In frequentists analysis parameter θ is taken to be fixed (albeit unknown) and data is considered to be random whereas Bayesian statistician says that θ is uncertain and follows a probability distribution while data is taken to be fixed.

Also it is important to notice that Bayesian analysis carefully distinguishes between $p(\theta|data)$ and $p(data|\theta)$ and all inference from Bayesian analysis is based on a posterior distribution, which is a true probability distribution. Thus Bayesian analysis ensures natural interpretations for our estimators and probability intervals. More on the basics of Bayesian analysis can be found for example in (58) and (59).

In practice it is not straight forward to compute an arbitrary posterior distribution but we can sample from it. For sampling we may use of the Markov chain Monte Carlo idea which is often abbreviated MCMC. Simply the idea is to construct a Markov chain such that it has desired posterior distribution as its limiting distribution. Then we simulate this chain and get sample from desired distribution. Maybe the most used software written for MCMC is called BUGS which is abbreviation for Bayesian statistics using Gibbs sampling.

4.1.5 Bayesian meta-analysis

A meta-analysis is a statistical procedure in which the results of several independent studies are integrated. There are always some basic issues to be considered in meta-analysis such as the choice between fixed-effects model and random-effects model, treatment of small studies and incorporation of study-specific covariates. For example in a clinical trial with many health care centres involved. The centres may differ in their patient pool e.g. number of patients, overall health level and age or the quality of the health care they provide. It is widely recognized that it is important to explicitly deal with the heterogeneity of the studies through random-effects models, in which for each centre there is a centre-specific "true effect" included.

Bayesian methods allows us nicely to deal with these problems within an unified framework (60).

To give more light on Bayesian meta-analysis, let us let us now give a relatively simple example. Assume that we have n studies. We are interested in average μ of the parameters θ_j , $j = 1, \dots, n$. Using information available from these studies we calculate a point estimators y_j for parameters θ_j . The first stage of the hierarchical Bayesian model assumes that point estimators conditioned on parameters are e.g. normally distributed i.e.

$$y_j | \theta_j, \sigma_j \sim N(\theta_j, \sigma_j)$$

We can simplify this model by assuming that σ_j is known. This simplification does not have much effect if sample sizes for each study are large enough. As an estimator of σ_j we can take for example sampling variance of point estimator y_j . The second stage of our model assumes normality for θ_j conditioned on hyperparameters μ and τ ,

$$\theta_j | \mu, \tau \sim N(\mu, \tau)$$

Finally, we assume a noninformative hyperpriors for μ and τ . The analysis of our meta-analysis model follows now normal Bayesian procedure, the inference is again based on posterior distribution $p(\mu|data)$.

More reading where Bayesian meta-analysis is applied can be found in Dominici et al. (61) and Samet et al. (62).

Let us now consider an example on air pollution (fine particles) and its health effects measured by a health test. We are interested in the relation of personal exposure to fine particle matter and the health effect. For health effect we have binary data y , $y \sim Ber(p)$, given by a health test (e.g. ST-segment depression in ECG) where one indicates a health problem and zero stands for no problem. We also have data on ambient concentration of fine particle matter, denoted by variable z_1 , for each day of health test. What we do not have is data on personal exposure for health measurement days which we denote by x_1 . So there is missing information between personal exposure and health test. However we have another data set that connects ambient concentration to personal exposure. In this second data set we denote ambient concentration by z_2 and personal exposure by x_2 . Solution to our problem now is a model consisting of two parts. First part is logistic health effect model which assumes that

$$p = \text{logit} (a + \beta_1 x_1 + \beta_c c)$$

where c stands for all confounding variables required in model. Second part is a linear regression model

$$x_1/a, b \sim N (a + b z_1, \sigma^2_{x1}) \tag{1.1}$$

where we obtain estimates to personal exposure on health test days. Parameters a and b are estimated from

$$x_2/a, b \sim N (a + b z_2, \sigma^2_{x1})$$

using our second data set. One of advantage using Bayesian statistics here to analyze relation on personal exposure and health effect is that the model takes into account our uncertainty of x_1 in formula (1.1).

To summarize above we have

$$\begin{aligned}
 y &\sim \text{Ber}(p) \\
 p &= \text{logit}(a + \beta_1 x_1 + \beta_2 c) \\
 x_1 | a, b &\sim N(a + b z_1, \sigma_{x1}^2) \\
 x_2 | a, b &\sim N(a + b z_2, \sigma_{x1}^2)
 \end{aligned}$$

To complete our model we set some prior distributions for parameters a , β_1 and β_2 and for parameters a and b . The posterior distribution of this model can be simulated using MCMC methods described above. This kind of idea is applied for example in Dominici et al. (61).

4.2 Expert panel

Expert elicitation is a structured process to elicit subjective judgements from experts. An expert is a person who has special skills or knowledge in a particular field. A judgement is the forming of an estimate or conclusion from information presented to or available to the expert (63).

Quantitative data are usually derived from several studies (e.g. epidemiological, toxicological). However, these data are often not (yet) available or, if available, are incomprehensive, unreliable and/or only indirectly or not at all applicable, resulting in knowledge that is incomplete. In such cases, expert judgement is a good way to complete the required knowledge (64). An expert panel has been used in several scientific fields, including chemistry, nuclear sciences and environmental sciences (65-69).

Traditional scientific methodology does not explicitly recognize the use of experts' opinions as scientific data; it remains rather controversial. One reason for that is the subjectivity involved; a subjective probability is 'just someone's opinion'. Another reason is the several (cognitive) biases which can affect the results of an expert panel (see section 9). Also there is not much uniformity in the methodology of an expert panel (64).

Therefore, it is very important to use the expert judgement methods in an as transparent and reproducible way as possible. The following principles can be used as a guideline for using expert panel in science (64):

- **Reproducibility.** It must be possible for scientific peers to review and if necessary reproduce all calculations. The whole expert judgement exercise should be transparent.

- **Empirical control.** A methodology for using expert judgement should incorporate some form of empirical quality control of final results.
- **Neutrality.** The method for combining/evaluating expert judgements should encourage experts to state their true opinions.
- **Fairness.** All experts are treated equally, prior to processing the results of their assessments.

While researchers have been studying the process of eliciting and interpreting expert judgements for several decades, and some protocols for the elicitation of experts have been described previously (69-72), no single accepted, standardized elicitation method has emerged (73).

As an outcome of a joint project of the European Union and the US Nuclear Regulatory Commission, the Delft University of Technology, the Netherlands, has developed a procedure guide document (70) which provides details for a structured expert judgement (see also table 7). This guide document refers in particular to expert judgements with the aim of achieving uncertainty distributions for uncertainty analyses in the nuclear sector. Of course, the method itself is applicable for other purposes as well.

Table 7 Steps in expert elicitation, partly based on Cooke et al. (70)

Different Steps in expert elicitation

1. Formulate the aim of the expert panel
 2. Developing questions of interest
 3. Identification/selection of experts
 4. Developing calibration questions
 5. Development of the elicitation protocol
 6. Expert training session
 7. Expert elicitation session
 8. Analysis of expert data
 9. Robustness and discrepancy analysis
-

4.2.1 Expert panel to derive ERFs

Since 2003, Industrial Economics conducted two expert judgement studies/assessments of the ERF between fine particulate matter (PM_{2.5}) exposure and mortality for the US EPA.

The first was a ‘pilot’ expert judgement study of five experts conducted in 2003 and 2004 aimed at exploring and refining the application of expert elicitation methods in the context of air pollution policy, as well as to evaluate the use of expert panel for providing a more complete uncertainty characterization in the ERF of PM_{2.5} and mortality (73). The second, which began in late 2004, was a full-scale study of twelve experts that built on the experience gained from the pilot and incorporated numerous refinements, and was reported in September 2006 (74).

The current EPA approach to characterize uncertainty in the PM_{2.5} / mortality ERF relies primarily on the statistical error reported in the selected epidemiological studies. In these expert elicitation studies, the experts were asked to consider and reflect in their judgements the impact of several other sources of uncertainty. Examples of these other uncertainties include the shape of the ERF, the strength of the causal relationship, uncontrolled confounding, effect modification, errors in exposure measurement, the role of co-pollutants etc.

Tuomisto et al. (68) have performed recently a structured expert judgement study of the population mortality effects of PM_{2.5} air pollution as well. This expert elicitation exercise was part of the Harvard Kuwait public health project which assessed the health impacts of the 1991 Kuwait oil fires. Six European air pollution experts were elicited, responding to an extensive series of questions including the effects of PM_{2.5} in general, in specific locations, among particular populations, the time lag between exposure and the relative toxicity of particles from various sources.

In the section below we summarize some main issues related to the steps in expert elicitation as named above. For a more detailed description of expert judgement see Hogarth (75), Morgan et al. (71), and Cooke (64).

1. Formulate the aim of the expert panel

At the beginning of every study, formulation of detailed objectives is very important.

2. Developing questions of interest

In the expert panel study of Tuomisto et al. (68) 12 questions of interests and 12 calibration questions were asked. To gain insight in the experts' reasoning, they also asked the reasons underlying their judgements.

The expert judgement studies of Industrial Economics (73, 74) included both qualitative and quantitative questions. The qualitative questions probed experts' belief concerning key evidence and critical sources of uncertainty when characterizing the ERF of PM_{2.5} and mortality, and were intended to make the conceptual basis for their quantitative judgements explicit. These questions covered topics such as potential biological mechanisms linking PM_{2.5} exposures with mortality, key evidence on the magnitude of the PM mortality relationship, sources of potential error and bias in epidemiological results, the likelihood of a causal relationship between PM_{2.5} and mortality, and the shape of the concentration-response function. A standard format of a quantitative question about the ERF of PM_{2.5} and mortality is for example: what is your estimate of the true, but unknown, percent change in non-accidental mortality in the total European population over the one week following a 10 µg/m³ increase in a pollutant (PM_{2.5}) level on a single day throughout the EU?

Given the amount of uncertainty inherent in predictions of this sort of questions, the experts may feel uncomfortable about giving point values. Therefore, the judgement is usually given as probability density function (PDF) reflecting the expert's degree of believe (63). The top of a PDF represents the most likely value of the variable, whereas its range is reflected in the lower and upper bounds (76).

For each questions, the experts' are often asked to specify their 5th, 25th, 50th, 75th and 95th percentile. Experts are also asked to provide a minimum and a maximum value to bind the distribution. Problems can arise here when experts relied much on the heuristic of 'anchoring and adjustment' to develop their distributions.

In the full-scale expert judgement study, the experts first specified a functional form for the PM_{2.5} mortality ERF and then developed an uncertainty distribution for the slope of that function, taking into account the evidence and judgements discussed during the qualitative questions (74).

3. Identification/selection of experts

Experts are often selected using peer nomination. The top researchers in a particular area are asked to request nominations. Then these nominations are used to rank and select experts for the expert panel. Selection criteria for the nominees ranking has to be made. Of course they have to be at a certain level of expertise within a particular area. Criteria about the origin of the nominees' expertise should be formulated as well. Nominees may include primary scientific researchers as well as prominent individuals from scientific panels, institutions, journal editorial boards etc.

A number of experts with the highest numbers of nominations are asked to participate. If an invited expert is unwilling or unable to participate, the expert with the next most highly nominated candidate will be asked to participate.

Ideally in the end, the experts have to be a balanced group that reflects the full range of respected scientific opinions concerning a certain topic.

In the report of this type of study, it is quite normal to describe the expertise of the experts in the expert panel shortly. Also the names and functions should be reported in the paper. Of course the judgements will be not associated with the names of the experts individually, so results are treated anonymously.

4. Developing calibration questions

Techniques exist to objectively evaluate the experts' performance. The 'gold standard' for judging their performance requires that the 'truth' is known which is of course clearly beyond reach in exposure-response assessment. Other investigators have used additional sets of questions, for which the truth can subsequently be known, to assess the calibration of experts.

Calibration questions are often developed to actually assess the accuracy and precision of the experts' judgements about the questions of interests. These calibration or 'seed' questions must be drawn from the experts' area of expertise, but need not pertain to the actual questions of interests. The 'answers' of the calibration questions are not known to the experts at the time of the expert elicitation session.

Different weights for all the experts are then determined on the basis of these calibration questions and used to aggregate the individual PDFs in one combined PDF for each of the questions of interests.

In the expert panel of Tuomisto et al. (68) their calibration questions were based on PM₁₀ monitoring data and mortality data from London and Athens for the period 1997-2001. Very specific questions were asked, for example ‘on how many days in 2001 did the daily average PM₁₀ concentration exceed 50 µg/m³ at least at one of the London monitoring stations?’

The choice of meaningful calibration questions is often difficult, critical and time consuming (77). One can wonder whether the performance on the calibration questions is representative for the performance on the questions of interests. In the pilot expert panel of the Industrial Economics (73) and in the full-scale study (74), no calibration questions were developed at all due to concern about feasibility of devising calibration questions that would be equally applicable to experts with the variability in technical background represented by the expert panel. In the absence of calibration questions they have at least tried to design an elicitation procedure to help experts avoid some of the common biases and errors of judgement that can lead to poor calibration, see also section 9.

5. Development of the elicitation protocol

It is very important to standardize the whole expert process. An elicitation protocol include an introduction and of course the questions of interests and the calibration questions.

This elicitation protocol will be used for the expert elicitation session as well as for the elicitation training session.

Often the elicitation protocol will be sent to the experts’ a few weeks before the expert training session together with a ‘briefing book’. A briefing book contain scientific articles, summaries etc., thus providing relevant documents on a particular area.

Pilot testing is critical to develop a well-functioning elicitation protocol. It has to be tried out on experts to find out where ambiguities and flaws need to be repaired, and whether all relevant information and questions are provided (63).

6. Expert training session

It is most practical that the elicitation session is preceded by a training session (workshop) in which all experts are gathered together, discuss the issues and undergo a practice elicitation exercise. At first, the researcher generally explains to the experts the nature of the problem at hand and the analysis being conducted. The purpose of this discussion is to give the experts some context regarding how their judgements will be used.

They need to be trained also in providing subjective assessments in probabilistic terms and in understanding subjective probability related issues. Also the workshop can be used to clarify the questions and underlying assumptions. The evidence can be discussed with respect to key issues. The session should also include an explanation of the types of heuristics that experts may use when making judgements.

The goal of such a workshop would not be to force a consensus but rather to help experts avoid errors in their judgements that might arise from uneven preparation and understanding (73).

7. Expert elicitation session

Ideally, in the expert elicitation session, the experts have to be interviewed individually by a normative analyst and a substantive analyst. A normative analyst is somebody who has experiences with probability issues and a substantive analyst is someone who is experienced in the expert's field of interest. However, there may be cases, for practical and financial reasons, in which the researcher may have to work with a group of experts' simultaneously. This is not recommended, because group interactions may interfere with the purpose of obtaining subjective probability distributions and some research does suggest that interactions between experts can increase rather than decrease the problem of overconfidence (71).

At the other hand, a key benefit of a group of experts in one elicitation session is the sharing of knowledge. One approach to deal with groups is to use them for the motivating / structuring phase of an elicitation, and then separate the group and ask individuals to make the judgements (78).

When experts are dispersed it may be difficult and expensive to bring them together. For these reasons the whole expert elicitation exercise could be done by mail, but this has serious drawbacks. An actual meeting for the elicitation training session and the formal elicitation session is by far preferable. To find a place where the experts are already gathered, and arrange a pre-meeting to do the expert elicitations training and session is a good option which reduces the costs considerably.

An elicitation may take anywhere from half an hour to a day. Cooke (77) argued that the elicitation session should not exceed a half day. The elicitation process is far more taxing for the expert than for the researcher; fatigue sets in after two hours. The full-scale study of the Industrial Economics session takes approximately eight hours.

8. Analysis of expert data

Morgan et al. (71) have emphasized that the full range of expert judgements should be presented in the results. Individual distributions of a question of interests can be depicted as boxplots with a diamond symbol show the median, the box defining the interquartile range and the whiskers defining each expert's 90 percentiles.

These individual distributions should be analysed separately, but it is also useful to aggregate the individual PDFs in one combined PDF for each of the questions of interests. Often weighted averages of the expert's distributions have been used, where the weights are non-negative and sum to one. The simplest method is to take all weights to be equal as was done in the Industrial Economic expert panel study (73).

Other methods to derive weights can be used included assigning global or item weights to experts (Cooke's Classical Model) (64, 70). These weights are based on two performance measures, calibration scores and information scores, which are assessed on the calibration questions whose true value are known after the elicitation session. An expert is well-calibrated if, for example, when asked to give his 90 percent confidence intervals for 100 predictions, his intervals contain the true value 90 percent of the time. Information is about the wide of his confidence intervals. Two experts, one giving very broad intervals and the other very narrow, can both be well calibrated, but the latter is more informative (73, 74).

The global weight is defined, per expert, by the product of expert's calibration score and his overall information score on the calibration questions. The item weight method is constructed using weights for each question of interest separately, using the experts' information scores for each specific question, rather than the use of one average information score for all questions of interests. For more information see Cooke (64, 77).

In principal, global weights are used unless item weights perform markedly better. Global weights are theoretically preferable to equal weighted combination. However, in practise it is obviously easier and much cheaper to use equal weighted combination.

9. Robustness and discrepancy analysis

Robustness analysis addresses the question, to what extent the results of the study would be affected by loss of a single expert or calibration question. Discrepancy analysis identifies the probability distributions of the questions of interests on which the experts differ most (63).

Conscious or subconscious discrepancies between the experts' answers and an accurate description of his underlying knowledge are termed biases. Experts are subject to a variety of biases, including motivational and cognitive biases. All these biases should be taken into account, or at least should be considered when analyzing the results.

Motivational biases occur if an expert wants to influence the decision in his favour by given a particular set of responses. Or the expert wants to bias his response because he believes that his performance will be evaluated by the outcome. An expert may have cognitive biases as well; it depends on the experts' modes of judgment. Some of these common pitfalls in expert judgement include overconfidence, adjustment/anchoring, availability, representativeness and coherence (63, 64, 72, 79).

Experts tend to *overestimate* their ability to make quantitative judgments. This is difficult for an individual to guard against; but a general awareness of the tendency can be important.

Most of the experts relied in some form or another on the heuristic of *adjustment and anchoring* to develop their uncertainty distributions. In this process, the expert begins his estimates with a particular study or set of studies, and then develops confidence intervals to account for various factors that influence his judgments. A technique to decrease the impact of this kind of bias the experts can be asked to begin development of their distributions by discussing the maximum and minimum values they believed possible.

Availability refers to the tendency to give too much weight to readily available data or recent experience in making assessments, which may not be representative of the required data.

Representativeness is the tendency to place more confidence in a single piece of information that is considered representative of a process than in a larger body of more generalized information.

Events are considered more likely when many scenarios can be created that lead to the event, or if some scenarios are particularly coherent. Conversely, events are considered unlikely when scenarios can not be imagined. Thus, probabilities are assigned more on the basis of ones ability to tell *coherent* stories than on the basis of intrinsic probability of occurrence.

For more information about biases and heuristics, see Cooke (64).

4.2.2 Role within the INTARESE project

Within the INTARESE project, we can make use of an expert panel for the following purposes:

1. ERFs + confidence limits
2. sources of uncertainties in ERFs in a more quantitative way

3. application of ERFs

1. ERFs + uncertainty

We can use the expert panel to assess ERFs. Relevant questions are for example ‘what is your estimate of the true, but unknown, percent change in a certain health effect in the total European population over a certain time period following a certain increase in a pollutant level throughout the EU?’

The experts will give their responses in probability distributions, which reflect the expert’s uncertainty/degree of believe. For the ERF, the experts’ are asked to specify their 5th, 25th, 50th, 75th and 95th percentile. Experts are also asked to provide a minimum and a maximum value to bind the distribution.

2. Sources of uncertainties in ERFs in a more quantitative way

The use of a confidence interval as a reflection of the overall uncertainty is rather limited. Actually, the underlying views and uncertainties and the assumptions/judgements that the experts’ make in order to assess a certain ERF are as important as the estimate of the ERF itself. The uncertainties involved have to make transparent and explicit as well. Possible sources of uncertainties include:

- Uncertainties in the problem framing
- Potential other health effects of a certain pollutant and the biological mechanisms involved
- Likelihood of a causal relationship between exposure and a certain health effect
- Shape of the ERF (e.g. a linear or log linear relationship) and the likelihood of a threshold level (which level?)
- Key epidemiological and toxicological literature
- Differences between the findings of studies. Short-term and long-term health effects of a certain exposure
- Exposure misclassification / exposure error
- (Residual) confounding / effect modification
- Time lag between certain exposure and health effect
- Mixture of certain pollutants

- Biases including publication bias

These uncertainties could be assessed in a more quantitative way by the use of an expert panel, which we potentially can then use as priors in multiple-bias modelling (see chapter 6).

3. Application of ERFs

Other questions relate to the application of a certain ERF in HIA, questions such as:

- Usability of a certain ERF in a specific situation e.g. differences between Europe
- Different ERFs for subgroups e.g. elderly, children

4.3 Comparison of methods

A major advantage in the Bayesian approach is the ease with which one can include study-specific covariates and that inference concerning the study-specific effects is done a natural manner through the posterior distributions. Interpretations for Bayesian estimators and probability intervals are very natural because of Bayesian analysis is based on true probability distributions. For example Bayesian 95% probability interval has interpretation that parameter of interest θ lies in that interval with probability 0.95. The frequentists confidence interval is a random interval meaning that if we would repeat our calculations 100 times with new data each time then, in average, 95 of our confidence intervals would contain θ . Thus, compared to classical methods Bayesian approach gives more complete representation of between-study heterogeneity and for sure more transparent and intuitive reporting of results.

However, Bayesian approaches are controversial because the definition of prior probability will often involve subjective assessments and opinion which runs against the principles of meta-analysis. Although in the Bayesian meta-analysis of Bell et al. (42) they make use of an uninformative/flat prior instead of a ‘subjective’ predefined prior. This flat prior indicates that all possible values of the summary estimate are considered approximately equally likely a priori, thus no subjective opinion is used in that case.

Other difficulties with Bayesian methods is the often far more complicated statistics and the relatively lack of knowledge (of most people in SP3) on these matters.

The proposed meta-regression methods described in paragraph 4.1.2 are two-level hierarchical models with studies/datasets constituting the highest level of hierarchy and individual study-specific data constituting the lowest level of hierarchy. This is equivalent to applying the Bayesian normal-normal hierarchical model and Monte Carlo Markov Chain (MCMC) methods for computation (80), but applies a different computational strategy based on the Berkey maximum likelihood estimation (56). Hence, the two approaches give similar results regarding the pooled estimates of exposure effects although they may differ in explaining heterogeneity.

As you can read in the above paragraph the formal method of an expert panel is a huge amount of work and is very time consuming. It is simply not feasible to recommend this expert panel method as the suggested methodology for the SP3 policy assessments to derive a certain ERF. Instead of the formal method of expert panel, one could also think of other more informal methods included the use of some experts judgements/opinions and guidance in some difficult ‘ERFs’ decisions to make transparent specific choices or a proof of concept from our ‘own’ experts within the INTARESE project itself.

4.4 Extrapolating from animal to human

While the methods of systematic review and meta-analyses described above can be applied equally well to animal and human studies, the use of animal studies require additional steps. Animal studies are produced for several different purposes and many studies are designed to characterize purely qualitative aspects of toxicology. The use of this type of data in risk assessment differs from the use of more quantitative animal data.

Any quantitative analysis of animal data has to be based on a careful analysis of the qualitative relevance of the animal data. Therefore we will review shortly the qualitative aspects of animal studies as well.

4.4.1 Qualitative aspects of animal studies

Animal studies often imply controlled exposure to a single very pure chemical. This is in contrast to exposure conditions in observational studies on humans, in which exposures often are complex and much less well characterized. In addition, potentially interacting factors such as dietary factors, temperature, sunlight, noise etc., are very well controlled in animal studies but less well controlled in observational studies.

Doses might have been selected to induce clear effects, and the purpose of the study might have been to characterize qualitative toxicological aspects (which type of effects does a certain chemical induce in animals? does a certain chemical have the capacity to cause a certain effect, such as cancer?). In risk assessment, many animal studies may primarily serve the purpose to answer qualitative questions about causality and to underpin correlations found in observational studies.

Under these conditions, doses used in animal studies are sometimes less important. Instead, a crucial question is whether an animal endpoint is relevant for a certain human health effect or a disease of interest. Alternatively, can it be expected that animal tests can be informative on a certain type of disease? Animal carcinogenesis has been well studied from this perspective (c.f. IARC). For example, a tumour response in animals is usually regarded as relevant for tumours in humans, even though several exceptions are known. A clear toxicodynamic example is carcinogenesis in male rats, implicating $\alpha 2\mu$ -globuline. This protein is not found in humans and the carcinogenic mechanism or mode of action can thus be regarded as irrelevant. A problem related to dose but of a qualitative nature is if high doses only have been shown to give tumours in animals. This may raise questions about relevance for humans if humans are exposed to comparatively low doses. In other words, is a linear dose-response or a threshold to be expected (see below)? In an effort to tackle this question, the issue of whether a genotoxic and a non-genotoxic mode of action are involved has to be resolved.

Even though only one single chemical has been classified as “probably not carcinogenic” by IARC, a negative result of an animal cancer test may also be taken as evidence for a lack of carcinogenicity, provided the test was performed according to the state of the art and if human data are very weak. Other effects or disease endpoints in animals are less well studied, and are sometimes questioned.

Clinically well known blood parameters are often used in animal studies to support causality and can often easily be interpreted in terms of human relevance. However, the situation becomes more complex if e.g. early pathological changes, such as early preneoplastic lesions, are used as endpoints. Animal models permit studies that are not possible to do in humans and many of these endpoints have not been characterized in humans. Subtle biochemical alterations might also be known to predispose to certain diseases in animals but their role in humans might be uncertain. In these cases, a “mode of action analysis” may be used for assessing human

relevance. The concept of mode of action analysis has been extensively discussed as a tool to analyse carcinogenic processes in animals (and in cell models) from a human relevance point of view. The sequence of key events leading to cancer in animals for a given chemical is analysed and compared to what is known about human effects for the same or related chemicals. This concept should also be possible, and fruitful, to use in the analysis of other diseases with a complex pathology.

These considerations indicate that extrapolations of qualitative animal data to humans cannot be assigned a fixed degree of uncertainty. In stead a case by case approach, based on a mode of action analysis should be performed. To take the extremes, an analysis of animal data for a certain chemical may lead to the conclusion that they are highly relevant for humans and perhaps even less uncertain than human data due to well controlled exposure conditions. Alternatively, an analysis may lead to the conclusion that they are non-informative. These considerations argue against generalized statements about errors in animal studies being larger as compared to errors in epidemiological studies.

Any quantitative analysis of animal data has to be based on a careful analysis of the qualitative relevance of the animal data. In case a chemical has produced a toxicological effect via a mechanism that is not working in humans, or deemed irrelevant due to other reasons, it is not meaningful to use them in a quantitative dose-response analysis.

4.4.2 Which animal models are relevant for humans?

As indicated above, cancer endpoints are well studied and bioassays for carcinogen testing are well standardized. In general, animal data obtained with such bioassays are regarded as informative, although the level of uncertainty varies and several exceptions are known. These uncertainties are expressed in the IARC evaluation system. However, carcinogen testing not based on state of the art protocols can be of little relevance. A clear example is the use of too short testing periods in cancer bioassays.

In many areas of toxicology test protocols cannot be regarded as standardized as is the case for cancer bioassays. Nevertheless, huge amounts of animal (and *in vitro*) data of significance for risk assessment have been published. Expert judgment and case by case analysis is necessary for assessing whether specific endpoints used in animal studies, or in *in vitro* studies, are

relevant for a specified human effect or disease. As mentioned above, such endpoints may include early or late biochemical, histological, gross pathology or behavioral effects. Comparable human parameters might not be available from clinical or epidemiological studies.

Certain common human diseases such as asthma or alterations in intellectual capacity are generally regarded as not easily modeled in animals. Furthermore, no validated and accepted animal model seems to exist for respiratory sensibilisation in humans. Test protocols have been elaborated for animal testing of teratogenic effects, but most human teratogens so far have been discovered in clinical practice or in epidemiological studies (81).

4.4.3 Quantitative aspects of non-cancer endpoints in animal studies

In cases when a genotoxic mode of action is not expected, animal studies are often designed to define critical effects of a given chemical or to define a threshold for a critical effect. Ideally, a broad spectrum of doses is used, and if the selection of doses has been successful, both Lowest Observed Adverse Effect Level (LOAEL) and No Observed Adverse Effect Level (NOAEL) can be defined. This can be done with a high degree of accuracy if the interval between the doses is small, and the degree of accuracy should be possible to calculate in numerical terms. Often, however, studies have not been optimally designed and e.g. a NOAEL is not possible to define.

Most dose metrics in animal studies are given as dose mass/day/kg body weight, and in these cases scaling factors are often used. Thus *per oral* doses for a small species as a mouse or rat are usually not directly extrapolated to humans, but divided by a scaling factor (7 for mouse and 4 for rat). This compensates for differences in caloric demand and metabolism, parameters that are more related to body surface than to body mass. As a rule, this gives that humans are to be considered more sensitive than laboratory animals in cases when a per oral dose is expressed as e.g. mg/kg body weight. Concentrations in inhaled air are not converted by scaling factors. However, toxicokinetic parameters may differ considerably between animals and humans and if kinetic data are available a physiologically based toxicokinetic modelling of target dose is preferred. If such parameters are not known default values for inter- and intra-species extrapolation may be used (see below).

Generally, there are three main types of dose scaling techniques. Dose scaling based on physical characteristics (e.g. body weight, body surface area, caloric requirements) is the most

frequently used type of approach. Among such approaches, scaling on the basis of body weight (isometric scaling) is most often applied in toxicology. The basic assumption is that numerous biological parameters show a linear correlation with body weight. Other factors, including absorption, plasma protein binding and biliary excretion are independent of body weight. As several of these functions correlate with body surface area, the latter was also used as a basis for interspecies extrapolation (allometric scaling). An alternative allometric scaling method to correct the intake/exposure is scaling based on caloric demand. Allometric scaling is more closely related to internal concentrations of a pollutant than to the external dose (26, 82).

Apart from dose scaling on the basis of physical characteristics two other approaches have been employed. The functional activity-based techniques basically scale the exposure/dose on the basis of lifespan and the multiple species regression. This latter approach requires conventionally adequate data from at least four species in order to estimate the equivalent dose for humans. However, because of the extreme data requirements of the method (studies performed in multiple species), it is not often applied (26).

Quantitative extrapolations to humans are often based on the use of safety factors or assessment factors (83). Thus, a NOAEL or a LOAEL from a particular animal study is divided with safety factors to accommodate areas of uncertainty in order to make a quantitative estimate of a safe dose for humans, i.e. at which no adverse effect would likely occur. Safety factors are not used to change the slopes of the ERF. Default assessment factors have been published and they may vary depending on in which sector of society the risk assessment will be used. For example, default values used in risk assessment of chemicals in the general environment are different from those recommended for risk assessment of industrial chemicals. In their Guidelines for risk assessment of carcinogens The US EPA provides guidelines for quantitative dose-response assessment (30).

Safety or Assessment factors

The EU document (83) describes several types of factors to be used when extrapolating data on threshold doses obtained in animals studies. These include factors for adequacy of the data base, route-to-route extrapolation, the use of LOAEL instead of NOAEL, duration of exposure, nature of the effects, interspecies (animal to human) extrapolation, and inter-individual variation in sensitivity among humans. For a detailed description of the reasoning behind the use of different safety factors, the scientific underpinning of default factors etc., see ref (83).

These descriptions comprehensively reflect limitations and weaknesses generally associated with animal studies.

Derivation of an overall assessment factor

An overall assessment factor is obtained by multiplying all single safety factors, and a LOAEL or NOAEL is then divided by this overall assessment factor. If all the assessment factors discussed here are multiplied as point estimates to obtain an overall assessment factor, one might end up with a very large factor, which would probably lead to a very high level of protection. Expert judgment and a case by case assessment are needed. It is recommended that distributions of the assessment factors should be used, if available, in the calculation of an overall assessment factor. In addition, distributions and point estimates can be used in parallel and be combined when necessary. Distributions are only available at present for the inter-species extrapolation factor and the factor for duration of exposure. Which percentile of the distribution should be chosen is a matter of policy (83).

It should be noted that the scientific foundation for the assessment factors is still unsatisfactory. Recent and ongoing studies on inter-species and inter-individual variability in pharmacokinetics provide information that may form the basis for new species- and metabolic pathway-related default assessment factors. More information, primarily concerning human inter-individual variation in sensitivity, is still needed (83).

4.4.4 Guidelines for carcinogen dose-response risk assessment

Shortly, for each effect observed, dose-response assessment should begin by determining an appropriate dose-metric. Several dose metrics have been used e.g. delivered dose, body burden and area under the curve. Selection of an appropriate dose metric considers what data are available and what is known about the agent's mode of action at the target site.

Physiologically based toxicokinetic modeling is the preferred approach for estimating dose metrics from exposure. Physiologically based models commonly describe blood flow between compartments and simulate the relationship between applied dose and internal dose.

The final analysis, however, should determine a human equivalent dose metric (30). Toxicokinetic modeling determines administered doses in animals and humans that yield equal tissue doses. Toxicodynamic modeling determines tissue doses in animals and humans that

yield equal lifetime risks. When toxicokinetic modelling is used without toxicodynamic modelling (standard) cross-species scaling procedures are available to cover toxicodynamic differences between animals and humans. For oral exposures, administrated doses should be scaled from animals to humans on the basis of equivalence of $\text{mg/kg}^{3/4}$ -d (milligrams of the agent normalized by the $3/4$ power of body weight per day (84).

For inhalation exposures animal exposures are replaced with human equivalent concentrations using EPA's methods for deriving inhalation reference concentrations (15), which give preference to the use of toxicokinetic modelling. When toxicokinetic models are unavailable, default dosimetry models are employed to extrapolate from animal exposures to human equivalent concentrations. The default dosimetry models typically involve the use of species-specific physiologic and anatomic factors relevant to the form of the agent and categorized with regard to whether the response occurs either locally or remotely.

The current default values of the parameters used in the default models are based on data from adults (15). More recently, the default parameters can be substitute by certain child-specific parameters(85).

Point of departure (POD)

All this modelling yields a dose estimate called the point of departure (POD). The POD is an estimated dose which is expressed in human equivalent terms. The POD is used as the starting point for subsequent extrapolations and analyses. For cancer dose-response data, the POD is an estimated dose 'near the lower end of the observed range without significant extrapolation to lower doses'. In the case of non-cancer dose-response assessment, the POD has generally been defined as the No observed adverse effect levels (NOAEL), the Lowest observed Adverse Effect Levels (LOAEL) or a modelled dose corresponding to an incremental e.g. the lower 95% limit of the dose or concentration corresponding to a 10% increase of response (LED_{10} or LEC_{10}). Use of the NOAEL or LOAEL has been criticized because of their dependence on study design and their lack of consideration of statistical error or the shape of the dose-response curve, and a Benchmark dose (BMD) method was proposed by Crump (86), and extended by Kimmel et al. (87).

Since risk at low exposure levels cannot be measured directly by animal experiments a number of mathematical models and procedures have been developed for use in extrapolating from high to low doses.

Linear extrapolation

For cancer risk assessment purposes the simplest assumption of mechanism is of a single transforming event that then leads to tumour formation with no further action needed, the ‘one-hit’ hypothesis, leading to linear extrapolation from the POD to lower doses. By linear extrapolation a line should be drawn from the POD to the origin, corrected for background. More complex models make assumptions about multiple events and different ways in which their effects interact to result in tumours, including the linear multistage model, the Weibull model, the probit and the logit model, see the recent review from Edler et al.(88). In practise mainly two of the mathematical models have been used in cancer risk assessments. These are the linearized multistage model and the low dose linear extrapolation model (see figure 6)

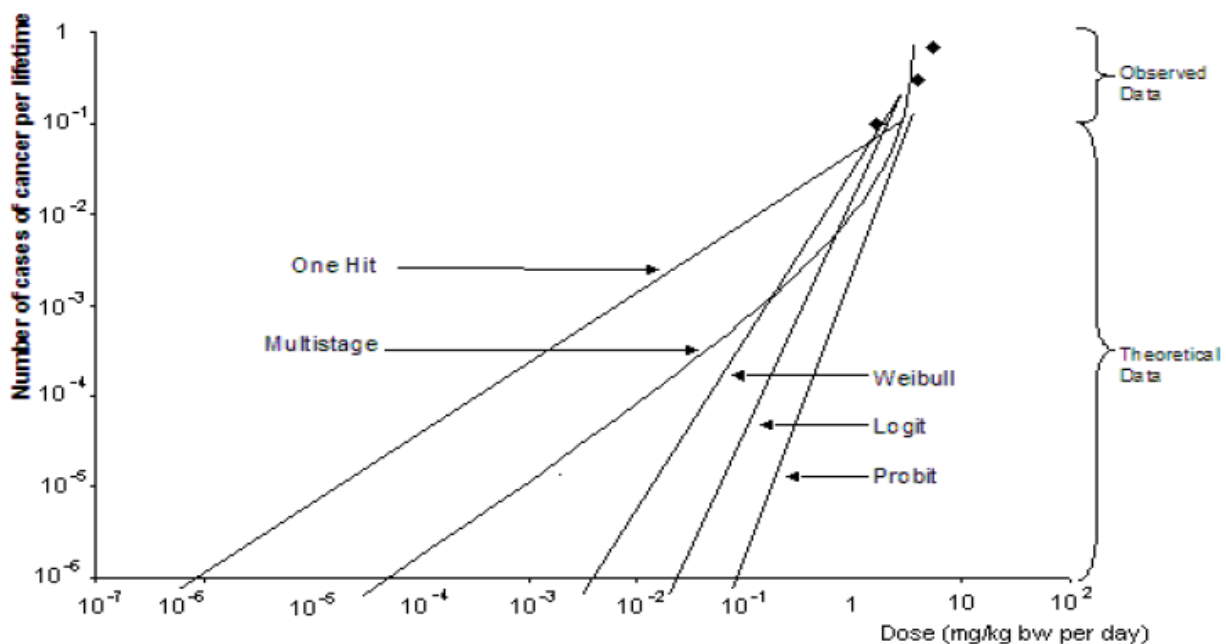


Figure 6 low dose extrapolation from animal carcinogenicity data using various models (89, 90).

The slope of the line, known as the slope factor, is an upper-bound estimate of risk per increment of dose that can be used to estimate risk probabilities for different low exposure levels. Its units are (proportion of individuals with tumours)/mg/kg bw/day (30).

Unit risk of cancer endpoints is often defined as the estimated upper-bound cancer risk at a specific exposure or dose from a continuous average lifetime exposure of 70 years. In this

approach, cumulative exposure is recommended as the appropriate dose metric. This implies that a higher dose received over a short period of time is equivalent to a corresponding lower dose spread over a longer duration (30). However, this is questionable for some exposures and cancer endpoints, and a potential large source of uncertainty.

Nonlinear extrapolation

The nonlinear extrapolation approach has been used when there are sufficient data to ascertain the mode of action and conclude that it is not linear at low doses and the agent does not demonstrate mutagenic or other activity consistent with linearity at low doses.

For nonlinear extrapolation the POD is used in the calculation of an oral reference dose (RfD) (91) or an inhalation reference concentration. (RfC) (15). The RfC is a benchmark estimate and is defined as ‘an estimate (with uncertainty spanning perhaps an order of magnitude) of a continuous inhalation exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious noncancer effects during a lifetime.’

In formula,

$$\text{RfC} = \text{NOAEL}_{[\text{HEC}]} / (\text{UF} * \text{MF}),$$

Where:

$\text{NOAEL}_{[\text{HEC}]}$ = The NOAEL or analogous effect level, dosimetrically adjusted to an health equivalent concentration

UF = Uncertainty factor(s) applied to account for the extrapolations required from the characteristics of the experimental regimen

MF = modifying factor to account for scientific uncertainties in the study chosen as the basis for the operational derivation.

The standard UFs applied are those for the following extrapolations: 1) data on effects of average healthy humans to sensitive humans; 2) laboratory animal data to humans; 3) studies of subchronic to chronic duration; 4) a $\text{LOAEL}_{[\text{HEC}]}$ to a $\text{NOAEL}_{[\text{HEC}]}$ and 5) from an incomplete data base. Actual determination of which UFs to apply and the magnitude of each is base on expert judgments (30).

There are other models and methods available for non linear extrapolation, including the methods of margin of exposure (MOE), which is defined as the ratio between a defined point on the dose-response curve for the adverse effect and the human intake. The defined point is

than usually the TD_{50} , TD_{25} or the BMD. See for example the methods used by the European Food Safety Authority (EFSA).

4.5 Informal methods

What if there is no existing ERF available from an authoritative and influential institute or organisation, like for example the WHO and due to time constraints you are unable to perform an appropriate systematic review and meta-analysis for every exposure-response from which there is sufficient evidence that it is causal and therefore want to include in your HIA.

In that case one could think of some adaptations of the formal suggested methods like:

- 1) Select some core quantitative literature instead of selecting all the relevant studies available of a certain exposure-response relation and try to combine those in a meta-analysis
- 2) Select an already published meta-analysis of good quality and use those estimates in your HIA
- 3) Select one individual (preferable international and multi-centre) study with a good quality and apply the individual derived ERF in your HIA
- 4) Look in former HIAs and related projects in your policy area and see which ERFs they have selected and whether you can use those as well for your HIA purposes

It is recommended that you consult expert(s) in your policy field (thus not a real expert panel!) for your ERFs decisions. They can give you advice about what the state-of-the-art is in the literature and they can suggest important meta-analyses/individual studies which can be used for your exposure-response assessment in your HIA.

Note that if you use more informal methods the checklist that needs to be considered in order to derive an ERF in box 1 (page 18) is still applicable! Also other paragraphs of this protocol are of course valid. This deals with questions like for which population and location is the ERFs developed, is the linear mathematical model reasonable for the ERF etc.

5 Review of methods for characterizing uncertainty in ERFs

Walker et al. (92) has presented a framework which is designed to help risk assessors conceive a broad spectrum of the uncertainties characterising their assessments. At the core of this framework is the notion that from a risk assessor point of view, uncertainty is best thought of as a two dimensional concept including the i) Location and ii) Level of uncertainty and iii) Nature of uncertainty.

Since the introduction, this framework has been applied several times, and also has been incorporated in the guidance for uncertainty assessment and communication of the Environmental Assessment Agency (RIVM/MNP), the Netherlands (63).

Within INTARESE WP 1.5 cross cutting issues this framework was adapted slightly, e.g. by leaving out the dimension of the nature of uncertainty. They will also slightly adapt the RIVM/NMP guidance for INTARESE purposes. The core of this framework is the notion that from a risk assessor point of view, uncertainty is best thought of as a two dimensional concept including the i) location and ii) level of uncertainty. The ‘location’ dimension refers to the aspect of the risks assessment model that is characterized by uncertainty and can be divided into certain categories of locations:

- ↳ Context
- ↳ Model structure
- ↳ Inputs
- ↳ Parameters
- ↳ Model outcome

The ‘level’ dimension refers to the severity of the uncertainty from the point of view of the decision maker and can be divided into different categories of levels of uncertainty:

- ↳ Statistically uncertainty
- ↳ Scenario uncertainty
- ↳ Recognized ignorance
- ↳ Total ignorance

In this chapter we will shortly review some methods for dealing with uncertainties in the exposure-response assessment phase of the risk assessment, assuming that one has performed a systematic review/meta-analysis (or make use of an existing combined estimate) of individual studies to derive a certain ERF. In paragraph 5.6 a broader method has been shortly described.

5.1 Introduction

Two broad types of error afflict individual epidemiologic studies: random error and systematic error (bias). In a study the random error is captured with the use of a statistical confidence interval (CI). A CI indicates the precision of the estimate as well as the likelihood that it is due to chance, but it reflects only the statistical uncertainty (random error) within the statistical model. In observational studies, systematic error (bias) is typically a much larger problem than random error. A study can be biased because of the way in which the subjects have been selected or as a result of factors that influence participation in the study (selection bias), the way the study variables are measured (information bias), or due to confounding factors. All these biases are not reflected in the CI.

Confounding factors are factors that change the strength of the association between the exposure and disease under study. Typical confounders may be age, sex, socioeconomic status, and smoking. Confounding can be controlled in statistical analysis using stratification and regression models. However, these methods can not adjust for unmeasured confounders, which can only be addressed with speculative discussions of how each might have biased the observed association. This is true also for uncertainty caused by selection bias and information bias. Thus in conventional analysis, the effects of these biases are only considered informally in the discussion, which often fails to capture sources of uncertainty and their interactions adequately. Therefore, conventional analysis produce misleadingly narrow interval estimates precisely when caution is most needed e.g. in meta-analyses and similar endeavours with potentially large policy impact (93).

5.2 Ordinary sensitivity analysis

Ordinary sensitivity analysis aims to estimate what the true association would be in light of the observed data and some hypothetical level of bias (94).

However, usually one bias at a time is examined, and thus interactions are overlooked.

5.3 Bayesian methods

Bayesian methods can be used to incorporate uncertainty regarding bias into the results of analysis. This method requires that the investigator specifies prior distributions (priors) for unknown parameters. As in conventional analysis one would then construct a model for the probability of the data given these parameters (i.e. the likelihood function). Finally using Bayes' theorem, the priors for unknown parameters would be combined with the probability for the parameter of interest. Bayesian methods can be somewhat involved and are not easy to implement with standard software. Analogous but simpler Monte Carlo sensitivity analysis has been proposed to account for likely bias (95).

5.4 Monte Carlo sensitivity analysis

Monte Carlo sensitivity analysis is an expanded version of ordinary sensitivity analysis. This analysis repeatedly re-estimates the effect measure of interest based on the observed data and the priors for bias sources. As in ordinary sensitivity analysis, one can 'correct' or adjust the effect rate ratio by dividing an observed (unadjusted) rate ratio by a bias factor. Probability distributions of all inputs and parameters, and the correlations between them, need to be specified. In Monte Carlo sensitivity analysis, a distribution for the parameter of interest is generated on the basis of repeated sampling from priors followed by adjustment, rather than sampling directly from a posterior distribution as in Bayesian analysis (95).

5.5 Multiple-bias modelling

In multiple-bias modelling, all major sources of uncertainty are systematically integrated in the data analysis by assigning hypothetical (prior) distributions to bias parameters. Thus, the analysis requires assumptions on which kind of biases exist, how biases act together and which priors should be used for them (96). Contemporary methods for multiple-bias modelling include Bayesian methods and Monte Carlo sensitivity analysis. When correcting for multiple biases, one has to make sure the correction order is reverse of the occurrence order of the biases (94).

According to Greenland (93) compared with conventional analysis and ordinary sensitivity analysis, multiple-bias modelling

- better captures uncertainty about effects
- requires the specification of a much larger model
- demands far more subject-matter knowledge
- requires much more presentation space and more effort by the reader
- if conducted and presented properly, depicts how, in the absence of experimental evidence, effects of interest are identified by prior distributions for bias sources rather than by data
- can produce firm conclusions only if indefensibly precise priors are used

5.5.1 Examples of Monte Carlo sensitivity analysis and Bayesian analysis

Steenland and Greenland (95) used Monte Carlo sensitivity analysis and Bayesian bias analysis to find out the degree of confounding of lung cancer rates by cigarette smoking in a US silica-lung cancer study. Ordinary sensitivity analysis was used to find a corrected point estimate for one hypothetical difference between the smoking habits of workers and the habits of the general population, as well as the effects of smoking on lung cancer. However, several scenarios should be considered in order to estimate the uncertainties about the relation of smoking to occupation and lung cancer. In the Monte Carlo sensitivity analysis 5000 randomly sampled confounding scenarios were used to repeatedly estimate the bias factor. Prior distributions for the smoking prevalence and smoking-lung cancer rate ratios were estimated from available data. The Bayesian analysis included also a prior distribution for the unknown smoking-adjusted rate ratio, but to parallel the Monte Carlo method, an essentially non-informative distribution was used. From the posterior distribution, 100 000 smoking-adjusted rate ratios were generated. The confidence limits of the conventional analysis gave the impression that the silica exposure is probably associated with an increase in the lung cancer rate of at least 30 % and quite possibly more than 90 % (relative to the US population), whereas the results of the Monte Carlo and Bayesian results gave the impression that this increase could easily be less than 20 % and is unlikely to be more than 90 %. According to Steenland and Greenland, in situations where the prior information is less precise, the bias distribution will lead to substantially wider Monte Carlo and Bayesian intervals relative to the conventional confidence intervals.

An example of Monte Carlo sensitivity analysis to quantify likely effects of exposure misclassification, given the sensitivity and specificity of classification, is presented in the article of Fox et al. (97). The method is applied to a study of the relation between occupational

resin exposure and lung-cancer deaths. The results are compared to the conventional results, which accounts for random error only.

Monte Carlo method for adding uncertainty about uncontrolled confounding and response bias to estimation of the health effects of magnetic fields can be found in the article of Greenland (98).

Another study of Greenland (93) illustrates the use of ordinary sensitivity analysis, Bayesian and Monte Carlo sensitivity analysis for pooled analysis of 14 case-control studies of residential magnetic field exposure and childhood leukaemia. Corrections were made for random error, response bias, confounding, and misclassification. When single corrections were compared, non-response was much larger source of uncertainty than random error, while confounding uncertainty was similar to uncertainty due to random error under the given priors. Classification error was the largest source of uncertainty because there were no relevant data or theory from which to develop a precise prior for misclassification. A Mantel-Haenszel analysis (corrected for random error only) produced an estimated odds ratio for the electric magnetic field-leukaemia association of 1.68, with 95% confidence limits of (1.27, 2.22). When the response bias, confounding and misclassification were combined with the effect of random error, the corrected Mantel-Haenszel odds ratio varied from 1.91 (0.95, 7.50) to 3.27 (0.92, 43.2) depending on the misclassification. However, Greenland (93) concluded that no agreement about the existence of an effect (let alone its size) could be forced by the data without more precise knowledge of the classification error. Simpler version of this study (omitting misclassification bias) is presented in a former article of Greenland (99).

5.6 The NUSAP approach

A whole different approach to characterise uncertainty is the NUSAP (Numeral Unit Spread Assessment Pedigree) approach. This approach aims to address quantifiable and unquantifiable uncertainties in a structured and transparent way rather than to correct an ERF by taken into account and quantify only a few uncertainties in the analysis.

Key dimensions of uncertainty are technical (inexactness), methodological (unreliability), epistemological (ignorance), and societal (social robustness). Ideally, all dimensions needs to be addressed. Quantitative methods, such as the Monte Carlo analysis described above, address only the technical dimension of uncertainty. They can be complemented with new qualitative

approaches, resulting in for example a novel approach to uncertainty assessment known as the NUSAP method (100).

NUSAP is a notational system proposed by Funtowicz and Ravetz (101), which aims to provide an analysis and diagnosis of uncertainty in the knowledge base of complex policy problems. It may be applicable to other environmental issues like for exposure-response assessment as well. For the uncertainty assessment, they make use of expert judgement. They make use of a selection of possible sources of errors/assumptions/input variables for example uncertainties due to different problem framings, choice of the shape of the ERF and the likelihood of a threshold, mixtures and time lags etc, and use this as a starting point for the elicitation of pedigree scores and probability distributions. After that Monte Carlo analysis has been done to trace out how the uncertainties in the inputs propagate in the calculation and to find out what the contribution of each input variable is to the variance in the overall estimate. They also combine the quantitative and the qualitative assessment by use of a diagnostic diagram, which provides insight on two independent properties related to uncertainty, namely spread and strength. Spread expresses inexactness whereas strength expresses the methodological and epistemological limitations of the underlying knowledge base (100).

For more information, see Risbey et al (79) and Van der Sluijs et al. (102) for an application of the NUSAP method in the emission monitoring process of volatile organic compounds from paints in the Netherlands. No published study has been identified which has characterize and quantify the uncertainties of a certain derived ERF by the use of NUSAP.

6 Combining animal and human studies in exposure-response assessment

Studies contributing knowledge to risk assessment are often divided into epidemiological and toxicological studies. Epidemiological studies can be divided into 1) observational studies and 2) experimental studies. Toxicological studies can be divided into 3) in vivo studies including 3a) controlled (experimental) animal studies and 3b) controlled (experimental) human studies and 4) in vitro studies. Epidemiological experimental studies (clinical trials) in the environmental epidemiology field are scarce. ‘In vitro’ studies are necessary in particular for characterising the mode of action/mechanism, but these findings need to be validated in whole animal studies, and therefore these in vitro toxicity studies are beyond our WP1.3 scope.

Usually in environmental health risk assessment there is only animal or in vitro studies available on the substance under study. However, for the most important exposures causing the main public health concerns and the most difficult risk management decisions, there is typically also epidemiological data on humans available. To make more valid risk management decisions, we need better methods to combine animal and human data.

Animal and human studies contribute both to hazard identification and to exposure/dose-response assessment. In hazard identification, combining these two sources of data is slightly less difficult, as the outcome of hazard identification is typically qualitative, and several attempts have been made to achieve this (see paragraph 6.4). In contrast, animal and human studies have rarely been combined in quantitative exposure/dose-response assessment.

The ultimate aim of both animal and human studies is to contribute to the estimation of the true slope and shape of the exposure/dose-response function of the substance under study in humans. The main difference is the sources of uncertainty, e.g. animal to human extrapolation in animal studies or multiple biases in human studies. The only valid possibility to quantitatively combine animal and human studies in exposure/dose-response assessment would be to estimate the size and direction of these uncertainties and what the effect is on the estimated combined ERF by making use of a common metric; the effect estimate and the standard error. Ultimately, this is currently not possible, but there are several recent

developments that have taken us several steps closer to the solution. However, first we need to analyze more carefully the uncertainties in different study types.

6.1 Uncertainties in the different study types

It is generally agreed that the best empirical way to test hypothesis about determinants of health in humans is the randomised clinical trial. The main strengths of randomised clinical trials are that it studies the correct species, which is the obvious limitation of animal and in vitro studies, and randomisation. Randomisation diminishes the possibilities for bias, i.e. that other factors than the exposure under study creates the observed difference in outcome. Bias is the main problem of observational studies.

Epidemiological studies are mostly observational for obvious ethical reasons. Because of this, observational studies are liable to different kind of biases like:

- information bias like poor assessment of exposure, confounders and health outcome
- selection bias
- (residual) confounding / effect modification

It has been suggested that the main problem with using epidemiological studies in quantitative risk assessment is their poor exposure assessment. Many epidemiological studies use broad exposure categories like grouping subjects into three groups based on exposure and often there is no direct information on personal exposure. However, also such exposure data can be used in risk assessment and often the errors are small compared to possible errors from interspecies extrapolation (103).

Other problems with environmental epidemiology are the relatively low power of many studies to find associations and the problem of separating the effect of the exposure of interest among multiple exposures (because people are often exposed to a mixture rather than one pollutant). The main advantages of epidemiological studies in risk assessment is that their study the relevant species. The exposure scenarios and the low levels of exposure studied are also closer to the situation that the risk assessment is trying to address.

To use animal studies for risk assessment purposes, the following extrapolation uncertainties have to be considered:

- extrapolations from animal to human (interspecies variations)
- extrapolation from human to human (difference in sensitivity, intraspecies variations)
- extrapolations from high to low doses
- extrapolations from short-term to long-term exposure
- extrapolations from one pollutant to an often complex mixture of pollutants

Note that apart from the extrapolation uncertainties, other (more methodological) uncertainties of the animal study (often unrecognized!) can also be present. One could think of issues like (104):

- how clearly the agent was defined and, in the case of mixtures, how adequately the sample characterization was reported
- whether adequate animals and animals strains were used
- whether the dose was monitored adequately, particularly in inhalation experiments
- whether the dose, duration of exposure and route of exposure were appropriate
- whether there were adequately numbers of animals per group
- whether the choice of the comparison group was appropriate
- whether animals were allocated randomly to groups and the investigators blinded for exposure and health effect
- whether the duration and timing of observation of the health outcomes was adequate and
- whether the data were reported and analysed adequately

Thus, in animal studies, high exposure concentrations and relatively short exposure time are used. However, one of the main problems often considered with the use animal studies in risk assessment is the uncertainties in extrapolation from the test animal to human, i.e. interspecies extrapolation. Because of this, risk assessment based on animal studies has come under strong attack and it has even been suggested that animal studies can be used only to rank possible carcinogenic hazards (105). On the other hand, animal studies can be done before any humans

have been exposed. They can also more easily accommodate detailed physiological and pathological measurements.

There can be other extrapolation uncertainties present which can be as important as the extrapolation of animal to human like the extrapolation from a controlled laboratory exposure setting to the real world exposures.

See Table 9 and 10 in chapter 7 for a checklist to characterise and rate the possible sources of uncertainty within individual studies.

6.2 Recent developments in quantitatively combining human studies

With the advent of evidence based medicine in the last decades, methods for combined analysis of different human studies have greatly developed. The main focus has been on randomized clinical trials, which have fewer biases and therefore are easier to combine than observational studies. In meta-analysis (see paragraph 4.1), studies are no longer mechanically combined, but current meta-analysis involve sophisticated investigation of sources of heterogeneity using meta-regression (see paragraph 4.1.2) and adjustment for biases, like misclassification bias or measurement error and publication bias.

The recently developed multiple-bias modelling (see paragraph 5.3) provides a method to incorporate the estimated size and direction of the main biases in observational human studies, i.e. selection bias, misclassification bias, and confounding. Multiple-bias modelling does not solve the problem how to estimate the true size and direction of these biases, as the biases are different in each individual study. Best method currently available to estimate these biases is expert judgment combined with meticulous analysis of available data on the biases. In the best situation, the investigators have addressed these biases themselves by doing validation studies within the main study. If such validation studies are not available, validation studies done in similar situation or other available data need to be used.

6.3 Recent developments in quantitatively combining animal studies

To be able to combine results from several animal studies, it is obligatory that the single animal studies are analyzed in a way, which gives the best estimate (and its uncertainty) about the

dose-response slope based on that data, not ‘worse case’ or other estimate. The use of safety factors or upper confidence intervals or ‘worst case’ scenarios should be done only after the studies have been pooled.

Systematic reviews and meta-analyses of toxicological studies are uncommon (29), although the use of these methods to evaluate animal evidence has increased over time (31).

However, meta-analytic techniques are directly applicable to experimental animal studies and would provide a formal method to combine results from several studies and to explore possible differences in the toxicity of the chemical e.g. between different species. Meta-analysis could benefit toxicological risk assessment, which today often bases its judgment on one or two studies, which are selected to be the most relevant, discarding the evidence from other studies. However, in many situations meta-analysis is not possible or needed due to the small number of animal studies on a given substance/stressor.

A few detailed methods have also been developed (106) to quantitatively combine evidence from several toxicological studies. However, these methods do not attempt to combine the ERF. In contrast, they combine the information from several studies into a score on potential for endocrine disruption (107) or score on ecological risk (108) or potential for interactive effects (109).

The toxicological risk assessment has, however, recently been forced to more explicitly address uncertainties. The most recent, and possibly most influential, is the Proposed Risk Assessment Bulletin in early 2006 from U.S. Government Office of Management and Budget (110). This will set the stage in the future for development of the methods to assess uncertainties of epidemiological and toxicological studies using a comparable metrics and then, ultimately, combining both sources of data in a coherent and transparent way in environmental health risk assessment.

6.4 Combining human and animal studies

In hazard identification, combining animal and human data is slightly less difficult, as the outcome of hazard identification is typically a qualitative assessment about the potential causal association. In epidemiology, the main criteria used to assess causality are the Hill’s criteria.

They incorporate the criteria on ‘biological plausibility’, which typically in environmental health risk assessment comes from animal studies. Proctor et al (111), based on U.S.EPA guidelines, have developed similar criteria to the Hill’s criteria, but also incorporating criteria for animal studies. IARC has developed a scheme, where toxicological and epidemiological hazard identifications proceed mostly separately, but are combined in the final phase into the IARC classification of carcinogenicity.

A different approach to combine animal and human data qualitatively for hazard identification is the use of mode of action (MOA) information by the identification of key events to assess the relevance of animal tumours for human risk assessment. Drawing on US EPA and IPCS proposals for animal MOA evaluation for carcinogens, Meek (112) proposed a human relevance framework (HRF) which include a systematic evaluation of comparability between the postulated animal MOA and related information from human data sources. See for an application of the HRF both for non-DNA-reactive as well as DNA-reactive carcinogens the articles of Meek et al.(112) and Preston et al. (113).

There are also some attempts to combine animal and human studies quantitatively. Bayesian methods allow a flexible tool to combine different types of data and also prior information. These methods have been applied to combine dose-response slopes from animal and human studies on the association between trihalomethane and low birth weight (114). The analysis assessed mainly random error, but assessed the sensitivity of the results to assumptions on dose-response models, body weight, and water consumption. However, it did not assess the effect of potential biases in epidemiological studies or in animal to human extrapolation.

It has also been suggested that we should build more sophisticated mechanistic models (115) and thereby link animal and human data into risk assessment.

However, whatever method will be ultimately used, the procedure how this could be done is already visible. First, it is probably best to identify and rate the most important sources of uncertainty in each individual study separately. After that try to come up with some quantified estimation of the size and direction of the certain biases (e.g. making use of validation studies, expert judgement etc.) and estimate the more ‘refined’ effect estimate. Note that in the case of animal studies no safety factors should be used at this stage. After that try to combine those refined estimates by making use of a meta-analysis, for both human and animal studies

separately and eventually, try to combine the combined human ERF and the combined animal ERF using the selected method, e.g. meta-analysis, Bayesian methods, or mechanistic modelling. The latter step is potentially even more demanding than the first step due to the different types of data to be combined. Using Bayesian approaches the whole procedure can potentially be combined into a single hierarchical model.

Another possibility is to first combine the human and animal studies separately, and then estimate the size and direction of certain important uncertainties which play a role in all the studies used for your meta-analysis and quantify what the potential effect is on the combined estimates (rather than the effect of biases in your individual study). After that the refined combined ERFs for both human and animal studies should be combined. The latter approach is not tempting because the biases and size and direction are different in each individual study.

Methods to quantitatively combine animal and human studies are in their infancy. The main challenges are to develop 'general' methods to assess the important uncertainties of each study type as the biases are different in each individual study and study type. If the uncertainties can be assessed, statistical methods to combine the results have already been suggested. The ultimate challenge is to propose criteria for decisions in situations with very large uncertainty, as is often the case in environmental health risk assessment.

7 Suggested methodology

For the ‘first pass’ assessments we recommend that if there is already a published and up-to-date ERF available, preferably from an authoritative and influential institute or organisation, like for example the World Health Organization, one should use that in the SP-3 assessments. If not available, we recommend using the frequentist systematic review (including if appropriate a meta-analysis) to derive a certain ERF for the policy assessments.

If it is not possible to use these formal methods suggested above we suggested to make use of the more informal methods described in paragraph 4.5.

In addition to the recommended systematic review, one should consider some important sources of uncertainty by filling in the tables below for each epidemiological and animal study separately. We suggest that you fill in the tables below (table 9, 10) by trying to: 1) describe the uncertainties qualitatively in a few sentences and 2) rate the uncertainties on a +, ++, +++ and a -, --, --- scale. One plus implies that you estimate that the specific uncertainty (e.g. measurement error) results in fewer than 20 % increase of effect estimates; two pluses is between 20 and 50 % estimated increase of effect estimates and three pluses is more than 50 % estimated increase of effect estimates. The same division is made for the minuses for an estimated decrease in the effect estimate.

These tables may help to characterise and rate the most important sources of uncertainty which affecting your ‘true’ ERF. Maybe in the second pass assessment it is possible to quantify the most important sources of some core individual studies and come up with a more ‘refined’ estimate of the ERF. In appendix 4 the potential sources of uncertainty (biases) which can affect individual studies are described shortly.

Note that these tables are focused on the potential uncertainties in individual studies. There are of course uncertainties as well in the combination of these individual studies including the issue of heterogeneity and publication bias (see paragraph 4.1), which are important and needs to be considered as well.

In table 8 the WP 1.3 contact persons and their email addresses are given which can give help and guidance in your SP-3 exposure-response assessment work.

Table 8 WP 1.3 contact persons who can give help and guidance in the exposure-response assessment work

Workpackages in SP3	Name	email address
WP 3.1 Transport	Hanna Boogaard	j.m.c.boogaard@iras.uu.nl
WP 3.2 Housing	Joachim Heinrich	joachim.heinrich@gsf.de
WP 3.3 Agriculture land use	Gerard Hoek	g.hoek@iras.uu.nl
WP 3.4 Water	Mark Nieuwenhuijsen	mnieuwenhuijsen@imim.es
WP 3.5 Chemicals in household products	Johan Hogberg	johan.hogberg@imm.ki.se
WP 3.6 Wastes	Francesco Forastiere	forastiere@asplazio.it
	Juha Pekkanen	juha.pekkanen@ktl.fi
WP3.7 Climate	Klea Katsouyanni	kkatsouy@med.uoa.gr

Table 9 Checklist to characterise and rate possible sources of uncertainty within epidemiological (observational) studies and human experiments

POTENTIAL SOURCES OF UNCERTAINTY¹	Epidemiological study 1	Epidemiological study 2	Epidemiological study 3	Epidemiological study X
I. Selection bias				
1. Ascertainment bias				
2. Participation bias				
3. Choice of comparison group				
3. Loss to follow-up bias				
4. Other selection biases				
II. Information bias				
1. Measurement bias of exposure ²				
2. Measurement bias of health effect				
3. Recall bias				
4. Interviewer bias				
5. Other information biases				
III. Confounding				
1. Known confounders.				
2. Unknown confounders.				
3. Misclassification of confounder variables				

¹Potential sources of biases in epidemiological studies are described in appendix 4.

²Measurement bias of exposure not only included sampling errors but also questions like 1) Is the (measured/modelled) exposure relevant and specific to the etiologic hypothesis and disease outcome (biological relevant exposure, also in respect with lag time). 2) Is the chosen indicator of exposure a good representation of the often complex mixture people are exposed to in the real life? 3) Is the chosen exposure metric (e.g. the 1-hr, 8-hr or 24-hr average/ peak concentration / cumulative exposure) a relevant representation? 4) Is the (measured/modelled) exposure a good proxy for the average individual (personal) exposure/dose?

Table 10 Checklist to characterise and rate possible sources of uncertainty within whole animal studies

	Animal study 1	Animal study 2	Animal study 3	Animal study X
POTENTIAL SOURCES OF UNCERTAINTY				
I. (Methodological) uncertainties				
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				
II. Extrapolation uncertainties				
1.				
2.				
3.				
4.				
5.				
III. Other uncertainties				

8. Further work

In the next years we will work on the refinement of ERFs by the use of both animal and human studies. We will assess possibilities to integrate epidemiology and toxicology and try to make an effort in the integration of these two. We envisage development of a principal (methodological) paper and application to a few selected exposures.

The integration of human and animal data is strongly linked to the topic of taking into account uncertainty more systematically. Proposed methods included a formal expert panel, more informal methods like expert judgements/opinions and guidance, multiple-bias (Bayesian) modelling and the NUSAP approach.

To test and illustrate the suggested methods, a set of case studies of selected exposure-health effects relationships will be performed. Table 11 shows the selection of exposure-health effects to test and illustrate the suggested methodology. We will focus probably on the following exposures including dioxins, black smoke, ultrafine particles, disinfection byproducts e.g. chlorination and traffic noise. Some other exposures are suggested as well in the case of unexpected (methodological) problems of the exposures above.

Table 11 Selection of exposure-health effects to illustrate the suggested methodology

Exposure	Examples of health effect being considered	Responsible persons
Dioxins	- Several forms of cancer	Juha Pekkanen
Black smoke/ black or elemental carbon/diesel root	- Respiratory diseases - Mortality - Inflammation? - Cancer	Xanthi Pedeli / Klea Katsouyanni
Ultrafine particles (PM 0.1)	- Mortality - Lung inflammation?	Hanna Boogaard / Gerard Hoek
Disinfection by-products e.g. chlorination	- Bladder cancer - Reproductive effects	Mark Nieuwenhuijsen / James Grellier
Traffic noise	- Cardiovascular diseases effects (hypertension, blood pressure changes, angina pectoris, myocardial infarction) - Annoyance	Paul Fischer
Other suggestions (in the case of unexpected problems of exposures above):	- polycyclic aromatic hydrocarbons (PAHs) - Electromagnetic fields - Phtalates - Benzene	

References

1. European Centre for Health Policy. Health impact assessment: main concepts and suggested approach. Gothenburg consensus paper. Brussels; 1999.
2. Scott-Samuel A, Birley M, Arden K. The Merseyside guidelines for health impact assessment. Liverpool: International Health Impact Assessment Consortium; 1996.
3. World Health Organisation. Evaluation and use of epidemiological evidence for environmental health risk assessment. Copenhagen: WHO Regional Office for Europe (also: *Environ Health Persp* 2000;108:997-1002); 2000.
4. AIRNET. Air-pollution health impact assessment, Health Impact Assessment. AIRNET Work Group 4. <http://airnet.iras.uu.nl/>; 2005.
5. Lock K. Health impact assessment. *Bmj* 2000;320(7246):1395-8.
6. Douglas MJ, Conway L, Gorman D, Gavin S, Hanlon P. Developing principles for health impact assessment. *J Public Health Med* 2001;23(2):148-54.
7. Taylor L, Quigley R. Health impact assessment:: a review of reviews. London: NHS Health Development Agency; 2002.
8. Zartarian V, Bahadori T, McKone T. Adoption of an official ISEA glossary. *J Expo Anal Environ Epidemiol* 2005;15(1):1-5.
9. Nieuwenhuijsen M, Paustenbach D, Duarte-Davidson R. New developments in exposure assessment: The impact on the practice of health risk assessment and epidemiological studies. *Environ Int* 2006.
10. Nieuwenhuijsen M. Exposure assessment in occupational and environmental epidemiology. Oxford, UK: Oxford University Press; 2003.
11. Van der Hazel P, Zuurbier M. PINCHE project: Final report WP1 Exposure assessment. Arnhem: Public Health Services Gelderland Midden; 2005.
12. National Research Council. Environmental epidemiology 1. Public health and hazardous waste. Washington DC, USA: National Academy Press; 1991.
13. de Hollander AEM. A framework for assessing the significance of health impacts of environmental exposures. In: *Assessing and evaluating the health impact of environmental exposures*. Utrecht: Utrecht University; 2004.
14. American Thoracic Society. What constitutes an adverse health effect of air pollution? Official statement of the American Thoracic Society. *Am J Respir Crit Care Med* 2000;161:665-673.

15. US Environmental Protection Agency. Methods for derivation of inhalation reference concentrations and application of inhalation dosimetry. Research Triangle Park, NC: Office of Health and Environmental Assessment, Environmental Criteria and Assessment Office; 1994.
16. Federal Register. Guidelines and methodology used in the preparation of health effects assessment chapters of the consent decree water criteria documents; 1980.
17. Murray CJ, Lopez AD. The global burden of disease; a comprehensive assessment of mortality and disability from disease, injury, and risk factors in 1990 and projected to 2020. Global burden of disease and injury series: Harvard University Press; 1996.
18. Murray CJ, Lopez AD. Mortality by cause for eight regions of the world: Global Burden of Disease Study. *Lancet* 1997;349(9061):1269-76.
19. Swaen GM. A framework for using epidemiological data for risk assessment. *Hum Exp Toxicol* 2006;25(3):147-55.
20. Hill AB. The environment and disease: association or causation? *Journal of the royal Statistical Society of Medicine* 1965;58:295-300.
21. Le Tertre A, Schwartz J, Touloumi G. Empirical Bayes and adjusted estimates approach to estimating the relation of mortality to exposure of PM(10). *Risk Anal* 2005;25(3):711-8.
22. Medina S, Plasencia A, Ballester F, Mucke HG, Schwartz J. Apehis: public health impact of PM10 in 19 European cities. *J Epidemiol Community Health* 2004;58(10):831-6.
23. Martuzzi M, Krzyzanowski M, Bertollini R. Health impact assessment of air pollution: providing further evidence for public health action. *Eur Respir J Suppl* 2003;40:86s-91s.
24. De Vries CGJCA, Gerlofs-Nijland ME, Cassee FR. Concepts assessment of effects from combined exposure. Draft. WP 1.5. Cross cutting issues. 2007.
25. Cassee FR, Groten JP, van Bladeren PJ, Feron VJ. Toxicological evaluation and risk assessment of chemical mixtures. *Crit Rev Toxicol* 1998;28(1):73-101.
26. Dybing E, Doe J, Groten J, Kleiner J, O'Brien J, Renwick AG, et al. Hazard characterisation of chemicals in food and diet. dose response, mechanisms and extrapolation issues. *Food Chem Toxicol* 2002;40(2-3):237-82.
27. Bliss CI. The toxicity of poisons applied jointly. *Ann Appl Biol* 1939;26:585-615.
28. Samet JM. What can we expect from epidemiologic studies of chemical mixtures? *Toxicology* 1995;105(2-3):307-14.
29. Sandercock P, Roberts I. Systematic reviews of animal experiments. *Lancet* 2002;360(9333):586.

30. US Environmental Protection Agency. Final guidelines for carcinogen risk assessment. Risk assessment forum. Washington, DC: U.S. Environmental Protection Agency; 2005.
31. Peters JL, Sutton AJ, Jones DR, Rushton L, Abrams KR. A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. *J Environ Sci Health B* 2006;41(7):1245-58.
32. Egger M, Davey Smith G, Altman D. Systematic reviews in health care: meta-analysis in context. Williston: BMJ Books; 2001.
33. Dickersin K. Systematic reviews in epidemiology: why are we so far behind? *Int J Epidemiol* 2002;31(1):6-12.
34. Shapiro S. Meta-analysis/Shmeta-analysis. *Am J Epidemiol* 1994;140(9):771-8.
35. Anderson R, Atkinson R, Peacock J, Marston L, Konstantinou K. Meta-analysis of time-series and panel studies on Particulate Matter (PM) and Ozone (O3). Report of a WHO task group. Copenhagen: WHO Regional Office for Europe; 2004.
36. Egger M, Smith GD. Bias in location and selection of studies. *BMJ* 1998;316(7124):61-6.
37. Gregoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? *J Clin Epidemiol* 1995;48(1):159-63.
38. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. *Int J Technol Assess Health Care* 1996;12(2):195-208.
39. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;283(15):2008-12.
40. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995;16(1):62-73.
41. Blair A, Burg J, Foran J, Gibb H, Greenland S, Morris R, et al. Guidelines for application of meta-analysis in environmental epidemiology. ISLI Risk Science Institute. *Regul Toxicol Pharmacol* 1995;22(2):189-97.
42. Bell ML, Dominici F, Samet JM. A meta-analysis of time-series studies of ozone and mortality with comparison to the national morbidity, mortality, and air pollution study. *Epidemiology* 2005;16(4):436-45.
43. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7(3):177-88.

44. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med* 1998;17(8):841-56.
45. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *BMJ* 1997;315(7121):1533-7.
46. Carlin JB. Meta-analysis for 2 x 2 tables: a Bayesian approach. *Stat Med* 1992;11(2):141-58.
47. Sterling T. Publication decisions and their possible effects on inferences drawn from tests of significance- or vice versa. *J Am Stat Assoc* 1959;54:30-34.
48. Sterling T, Rosenbaum W, JJ W. Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *Am Stat* 1995;49:108-112.
49. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315(7109):629-34.
50. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994;50(4):1088-101.
51. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000;56(2):455-63.
52. Higgins J, Beyene J. An introduction to meta-regression.
<http://www.cochrane.org/colloquia/abstracts/ottawa/W-007.htm>: The Cochrane Collaboration,; 2006.
53. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998;351(9096):123-7.
54. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21(11):1559-73.
55. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med* 1995;14(4):395-411.
56. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Stat Med* 1998;17(22):2537-50.
57. Touloumi G, Atkinson R, Le Tetre A, al. e. Analysis of health outcome time series data in epidemiological studies. *Environmetrics* 2004;15:101-117.
58. Congdon JP. *Applied Bayesian Modelling: Wiley series in probability and statistics*; 2003.
59. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian data analysis.*: Chapman & Hall; 2004.

60. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med* 1995;14(24):2685-99.
61. Dominici F, Zeger SL, Samet JM. A measurement error model for time-series studies of air pollution and mortality. *Biostatistics* 2000;1(2):157-75.
62. Samet JM, Dominici F, Curriero FC, Coursac I, Zeger SL. Fine particulate air pollution and mortality in 20 U.S. cities, 1987-1994. *N Engl J Med* 2000;343(24):1742-9.
63. Van der Sluijs JP, Janssen P, Petersen A, Kloprogge P, Risbey J, Tuinstra W, et al. RIVM/MNP Guidance for uncertainty assessment and communication: tool catalogue for uncertainty assessment. Utrecht: Copernicus Institute for Sustainable Development and Innovation; 2004. Report No.: NWS-E-2004-37.
64. Cooke R. Experts in uncertainty- Opinion and subjective probability in science. New York: Oxford University Press; 1991.
65. Cooke R, Goossens L. Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research* 2004;7(6):643-656.
66. Goossens L, Cooke R. Applications of some risk assessment techniques: Formal expert judgement and accident sequence precursors. *Safety Science* 1997;26(1/2):35-48.
67. Arnell NW, Tompkins EL, Adger WN. Eliciting information from experts on the likelihood of rapid climate change. *Risk Anal* 2005;25(6):1419-31.
68. Tuomisto JT, Wilson A, Evans JS, Tainio M, Cooke R. Uncertainty in mortality response to airborne fine particulate matter: elicitation of european air pollution experts. Manuscript. 2006.
69. Goossens LHJ, Cooke RM, Woudenberg F, Van der Torn P. Expert judgement and lethal toxicity of inhaled chemicals. *Journal of Risk Research* 1998;1(2):117-133.
70. Cooke RM, Goossens LHJ. Procedures guide for structured expert judgement. Brussels-Luxembourg; 2000. Report No.: EUR 18820.
71. Morgan MG, Henrion M. Uncertainty. A guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge: Cambridge University Press; 1990.
72. Spetzler C, Von Holstein S. Probability Encoding in Decision Analysis. *Management sciences* 1975;22(3):340-358.
73. Industrial Economics. An expert judgement assessment of the concentration-response relationship between PM2.5 exposure and mortality. Cambridge, MA: Industrial economics; 2004.

74. Industrial Economics. Expanded expert judgement assessment of the concentration-response relationship between pm2.5 exposure and mortality. Cambridge, MA: Industrial Economics; 2006.
75. Hogarth R. Judgement and Choice. New York: Wiley; 1987.
76. Van der Fels-Klerx IH, Goossens LH, Saatkamp HW, Horst SH. Elicitation of quantitative data from a heterogeneous expert panel: formal process and application in animal health. *Risk Anal* 2002;22(1):67-81.
77. Cooke RM, Goossens LH. TU Delft Expert judgement data base
<http://www.rff.org/expertjudgementdocuments/documents/relevantdocs/CookeGoossens-EJDatabase.pdf>; 2006.
78. Mansfield C. Peer review of expert panel. Review on the following document: An expert judgement assessment of the concentration-response relationship between PM2.5 exposure and mortality.
http://www.rff.org/expertjudgementdocuments/documents/relevantdocs/Peer_Review_US_EPA.pdf; RTI International; 2004.
79. Risbey J, Van der Sluijs J, Ravetz J. A protocol for assessment of uncertainty and strength of emissions data. Utrecht: Department of Science, Technology and Society; 2001. Report No.: NW&S-E-2001-10.
80. Everson PJ, Morris CN. Inference for multivariate normal hierarchical models. *J R Statist Soc A* 2000;62:399-412.
81. Brent RL. Utilization of animal studies to determine the effects and human risks of environmental toxicants (drugs, chemicals, and physical agents). *Pediatrics* 2004;113(4 Suppl):984-95.
82. Calabrese EJ, Kenyon EM. *Air Toxics and Risk Assessment*. Chelsea, Michigan: Lewis Publishers; 1991.
83. Swedish National Chemicals Inspectorate. Proposals for the use of assessment (uncertainty) factors, Application to risk assessment for plant protection products, industrial chemicals and biocidal products within the European Union, KemI report No 1/03 (PDF, 485 kB)
http://www.kemi.se/upload/Trycksaker/Pdf/Rapporter/Rapport1_03.pdf. Solna, Sweden; 2003.
84. US Environmental Protection Agency. A cross-species scaling factor for carcinogen risk assessment based on equivalence of mg/kg³/4/day. *Federal Register* 57 (109):24152-24173; 1992.

85. US Environmental Protection Agency. A review of the reference dose and reference concentration process. Risk Assessment Forum. Washington, DC; 2002.
86. Crump KS. A new method for determining allowable daily intakes. *Fundam Appl Toxicol* 1984;4(5):854-71.
87. Kimmel CA, Gaylor DW. Issues in qualitative and quantitative risk analysis for developmental toxicology. *Risk Anal* 1988;8(1):15-20.
88. Edler L, Poirier K, Dourson M, Kleiner J, Mileson B, Nordmann H, et al. Mathematical modelling and quantitative methods. *Food Chem Toxicol* 2002;40(2-3):283-326.
89. Committee on carcinogenicity of chemicals in food cpatcC. Guidance on a strategy for the risk assessment of chemical carcinogens.
www.advisorybodies.doh.gov.uk/coc/guideline04.pdf. 2004.
90. EFSA. Opinion of the scientific committee on a request from EFSA related to a harmonised approach for risk assessment of substances which are both genotoxic and carcinogenic
http://www.efsa.europa.eu/etc/medialib/efsa/science/sc_committee/sc_opinions/1201.Par.0002.File.dat/sc_op_ej282_gentox_en3.pdf. *EFSA Journal* 2005;282:1-31.
91. Barnes DG, Dourson M. Reference dose (RfD): description and use in health risk assessments. *Regul Toxicol Pharmacol* 1988;8(4):471-86.
92. Walker W, Harremoes P, Rotmans P, Van der Sluijs JP, Van Asselt MVA, Janssen PHM, et al. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Journal of Integrated Assessment* 2003;4(1):5-17.
93. Greenland S. Multiple-bias modelling for analysis of observational data. *J R Statist Soc A* 2005;168(Part 2):267-306.
94. Greenland S. Basic methods for sensitivity analysis and external adjustment. In: Rothman K, Greenland S, editors. *Modern Epidemiology*. Philadelphia, PA: Lippincott-Raven Publishers; 1998. p. 343-57.
95. Steenland K, Greenland S. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *Am J Epidemiol* 2004;160(4):384-92.
96. Hofler M. The Bradford Hill considerations on causality: a counterfactual perspective. *Emerg Themes Epidemiol* 2005;2:11.
97. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *Int J Epidemiol* 2005;34(6):1370-6.

98. Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. *Int J Epidemiol* 2004;33(6):1389-97.
99. Greenland S. The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukaemia. *Journal of the American statistical association* 2003;98:47-54.
100. van der Sluijs JP, Craye M, Funtowicz S, Kloprogge P, Ravetz J, Risbey J. Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: the NUSAP system. *Risk Anal* 2005;25(2):481-92.
101. Funtowicz SO, Ravetz JR. *Uncertainty and Quality in Science for Policy*. Dordrecht: Kluwer; 1990.
102. Van der Sluijs JP, Risbey JS, Ravetz J. Uncertainty assessment of VOC emissions from paint in The Netherlands using the NUSAP system. *Environ Monit Assess* 2005;105(1-3):229-59.
103. Smith AH. Epidemiologic input to environmental risk assessment. *Arch Environ Health* 1988;43(2):124-9.
104. International Agency for Research on Cancer. *IARC monographs on the evaluation of carcinogenic risks to humans. Preamble*. Lyon, France: IARC; 2006.
105. Ames BN, Magaw R, Gold LS. Ranking possible carcinogenic hazards. *Science* 1987;236(4799):271-80.
106. Weed DL. Weight of evidence: a review of concept and methods. *Risk Anal* 2005;25(6):1545-57.
107. Calabrese EJ, Baldwin LA, KostECKI PT, Potter TL. A toxicologically based weight-of-evidence methodology for the relative ranking of chemicals of endocrine disruption potential. *Regul Toxicol Pharmacol* 1997;26(1 Pt 1):36-40.
108. Menzie C, Henning MH, Cura J, Finkelstein K, Gentile J, Maughan J, et al. Special report of the Massachusetts weight-of-evidence workgroup: A weight-of-evidence approach for evaluating ecological risks. *Human Ecological Risk Assessment* 1996;2(2):277-304.
109. Mumtaz MM, Durkin PR. A weight-of-evidence approach for assessing interactions in chemical mixtures. *Toxicol Ind Health* 1992;8(6):377-406.
110. US Government Office of Management and Budget. *Proposed Risk Assessment Bulletin*; 2006.
111. Proctor DM, Otani JM, Finley BL, Paustenbach DJ, Bland JA, Speizer N, et al. Is hexavalent chromium carcinogenic via ingestion? A weight-of-evidence review. *J Toxicol Environ Health A* 2002;65(10):701-46.

112. Meek ME, Bucher JR, Cohen SM, Dellarco V, Hill RN, Lehman-McKeeman LD, et al. A framework for human relevance analysis of information on carcinogenic modes of action. *Crit Rev Toxicol* 2003;33(6):591-653.
113. Preston RJ, Williams GM. DNA-reactive carcinogens: mode of action and human cancer hazard. *Crit Rev Toxicol* 2005;35(8-9):673-83.
114. Peters J, Rushton L, Sutton A, Jones D, Abrams K, Muggleston M. Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence. *Applied Statistics* 2005;54(part 1):159-172.
115. Portier CJ. Linking toxicology and epidemiology: the role of mechanistic modelling. *Stat Med* 2001;20(9-10):1387-93.
116. World Health Organization. Quantification of the health effects of exposure to air pollution. Report of a WHO working group. Bilthoven, the Netherlands: WHO regional office for Europe; 2001.
117. Mindell J, Hansell A, Morrison D, Douglas M, Joffe M. What do we need for robust, quantitative health impact assessment? *J Public Health Med* 2001;23(3):173-8.
118. Douglas M, Scott-Samuel A. Addressing health inequalities in health impact assessment. *J Epidemiol Community Health* 2001;55(7):450-1.
119. Katsouyanni K, Touloumi G, Spix C, Schwartz J, Balducci F, Medina S, et al. Short-term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project. *Air Pollution and Health: a European Approach*. *BMJ* 1997;314(7095):1658-63.
120. Katsouyanni K, Touloumi G, Samoli E, Gryparis A, Le Tertre A, Monopoli Y, et al. Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 project. *Epidemiology* 2001;12(5):521-31.
121. Abbey DE, Nishino N, McDonnell WF, Burchette RJ, Knutsen SF, Lawrence Beeson W, et al. Long-term inhalable particles and other air pollutants related to mortality in nonsmokers. *Am J Respir Crit Care Med* 1999;159(2):373-82.
122. Dockery DW, Pope CA, 3rd, Xu X, Spengler JD, Ware JH, Fay ME, et al. An association between air pollution and mortality in six U.S. cities. *N Engl J Med* 1993;329(24):1753-9.
123. Pope CA, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, et al. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *Am J Respir Crit Care Med* 1995;151(3 Pt 1):669-74.

124. Health Effects Institute. Special Report: Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. In: HEI; 2000.
125. Goldbohm RA, Tielemans EL, Heederik D, Rubingh CM, Dekkers S, Willems MI, et al. Risk estimation for carcinogens based on epidemiological data: A structured approach, illustrated by an example on chromium. *Regul Toxicol Pharmacol* 2006.
126. Samet J, Burke T. Epidemiology and risk assessment. In: *Applied epidemiology: theory and practice*. New York: Oxford University Press; 1998. p. 137-175.
127. Anderson HR, Atkinson RW, Peacock JL, Sweeting MJ, Marston L. Ambient particulate matter and health effects: publication bias in studies of short-term associations. *Epidemiology* 2005;16(2):155-63.
128. Kunzli N, Kaiser R, Medina S, Studnicka M, Chanel O, Filliger P, et al. Public-health impact of outdoor and traffic-related air pollution: a European assessment. *Lancet* 2000;356(9232):795-801.
129. Medina S. Presentation of APHEIS. Monitoring and assessment of health effects from air pollution meeting 26 -27.5.2005. www.apheis.net/IspraMeeting2005/Apheis.ppt; 2005.
130. de Hollander AEM. Valuing the health impact of air pollution: Deaths, DALYs or Dollars? In: *Assessing and evaluating the health impact of environmental exposures. "Deaths, DALYs or Dollars"*. Utrecht: Utrecht University; 2004.
131. Mindell J, Boaz A, Joffe M, Curtis S, Birley M. Enhancing the evidence base for health impact assessment. *J Epidemiol Community Health* 2004;58(7):546-51.
132. Joffe M, Mindell J. A framework for the evidence base to support Health Impact Assessment. *J Epidemiol Community Health* 2002;56(2):132-8.
133. Institute for Environment and Health. Risk Assessment Approaches used by UK Government for Evaluating Human Health Effects of Chemicals. Leicester: Institute for Environment and Health; 1999.
134. Health Canada Revised Health Canada Risk Management Framework. Draft for Discussion. April 1999. Canada; 1999.
135. US Environmental Protection Agency. Guidance for risk characterisation. Washington.: US Environmental Protection Agency Science Policy Council; 1995.
136. Organisation for Economic Cooperation and Development. OECD Guidelines for the Testing of Chemicals. Section 4: Health Effects. Vol 2, 10th Addendum. Paris: Organisation for Economic Cooperation and Development; 1998.

137. Environmental Health Risk Assessment (enHealth Council). Guidelines for assessing human health risks from environmental hazards. (<http://enhealth.nphp.gov.au/council/pubs/pubs.htm>). Canberra, Australia.; 2004.
138. Roemer W, Hoek G, Brunekreef B. Pollution effects on asthmatic children in Europe, the PEACE study. *Clin Exp Allergy* 2000;30(8):1067-75.
139. Sackett DL. Bias in analytic research. *J Chronic Dis* 1979;32:51-63.
140. Rothman KJ. Biases in study design. In: *Epidemiology: an introduction*. New York: Oxford University Press; 2002. p. 94-112.
141. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med* 1998;55(10):651-6.
142. Armstrong BG. Exposure measurement error: consequences and design issues. In: Nieuwenhuijsen M, editor. *Exposure assessment in Occupational and Environmental Epidemiology*: Oxford University Press; 2003.
143. Klimisch HJ, Andreae M, Tillmann U. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul Toxicol Pharmacol* 1997;25(1):1-5.
144. National Centre for the Replacement Refinement and Reduction of Animals in Research. www.nc3rs.org.uk. accessed november 2006.
145. Center for Alternatives to Animal Testing. caat.jhsph.edu/. accessed november 2006.
146. Organisation for Economic Co-operation and Development. Good laboratory Practice. http://www.oecd.org/departement/0,2688,en_2649_34381_1_1_1_1_1,00.html. Accessed november 2006.
147. Ghio AJ, Huang YC. Exposure to concentrated ambient particles (CAPs): a review. *Inhal Toxicol* 2004;16(1):53-9.

Appendices

Appendix 1 Review of health impact assessment issues

A WHO working party defined a HIA as a combination of procedures, methods and tools by which a policy, program or project may be judged as to its potential effects on the health of a population, and the distribution of those effects within the population (1). Specifically, the purpose of a HIA is (2):

- to assess the potential health impacts, both positive and negative, of projects, programmes and policies
- to improve the quality of public decision making through recommendations to enhance predicted positive health impacts and minimise negative ones

The major steps that play a role in a HIA are the following (3, 4):

1. specify the purpose and framework of the HIA
2. decide which exposure-effect pathways will be quantified
3. identify and characterise the population at risk
4. select or develop a suitable set of exposure-response functions (ERFs) that link (individual) pollutants with specific health endpoints, i.e. % increase in morbidity per $\mu\text{g}/\text{m}^3$ of a pollutant
5. derive the population exposure distribution
6. estimate the background rates (i.e. prevalence and/or incidence) of the relevant health endpoints in the population at risk
7. calculate the burden of disease or death in the population at risk
8. value the burden of disease or death in the population at risk
9. assess and quantify the uncertainty of the HIA

There has been much debate about HIA methodology; different approaches have been proposed (2, 5-7). It has been concluded that there is no single ‘blueprint’ for HIA that will be appropriate for all circumstances (6). For a more detailed description of HIA methods, see the

Merseyside Guidelines for health impact assessment (2) or check www.who.int/hia/en/ for more (country specific) guidelines.

Table 1 Overview of health impact assessment (HIA) issues

Steps in HIA	HIA issues
1. Specify the purpose and framework of the HIA	Transparency
2. Decide which exposure-effect pathways will be quantified	Systematic approach Judgements Integration several disciplines
3. Identify and characterise the population at risk	Mixture of pollutants Combined effects Vulnerable individuals Environmental justice
4. Select or develop a suitable set of exposure-response functions (ERFs)	Limited data Transferability Mathematical choices Limited exposure assessment Integration several disciplines Precautionary principle Methodology for developing ERFs
5. Derive the population exposure distribution	Linkage with ERFs
6. Estimate the background rates (i.e. prevalence and/or incidence) of the relevant health endpoints in the population at risk	Limited data Non-comparability
7. Calculate the burden of disease or death in the population at risk	Use of RR of RD Competing risks
8. Valuate the burden of disease or death in the population at risk	Discount rates Severity/quality weightings Double counting
9. Assess and quantify the uncertainty of the HIA	Measurement error modelling

Identified problems in HIA are the lack of transparency and the lack of a systematic approach. Typically there is limited data, so that a lot of choices are made that are not often made explicit. Another issue in HIA is the poor integration between various disciplines like exposure and health effect assessment or epidemiology/toxicology. Another important topic is how to deal adequately with uncertainty. In table 1 an overview is given of all the identified HIA issues. The issues marked in bold are possible innovation ambitions and are important not only for WP 1.3 but for the INTARESE SP-3 assessments in general.

The section below summarizes some main HIA issues related to the steps listed above.

1. Specify the purpose and framework of the HIA

The purpose of the HIA and framework of the HIA should be made clear because decisions concerning the choice of epidemiological and other data will depend on the objectives of the assessment. Ideally, policy-makers, scientists and also stakeholders should be involved in defining the scope of the assessment (2, 3).

Transparency is very important in all stages of HIA. To ensure transparency, HIA results should include assessments of the reliability of the impact estimates that are reported (both qualitative and quantitative). More importantly, a clear, systematic and detailed reporting of the methods used and all the judgements and assumptions made, will help to ensure that policymakers and stakeholders can understand what was done and have the information necessary to evaluate the HIA critically (4).

2. Decide which exposure-effect pathways will be quantified

In every HIA, certain judgements need to be made about what exposure-effect pathways are relevant for your HIA first. Should an HIA only model those effects for which there is sufficient evidence for causality or also include effects for which the evidence is more limited? These kinds of decisions often involve making judgements on issues where there is no scientific consensus yet. These judgements need to be based on various disciplines such as toxicology, exposure assessment and epidemiology (4).

The integration of the different disciplines (e.g. toxicology, epidemiology) in HIA frequently results in problems.

Often exposure is a complex mixture of different pollutants and it is very difficult to disaggregate a mixture into its component parts to identify ‘relevant’ exposure-effect pathways. Use of a single or a multi-pollutant model depends on the question being asked by the HIA (4).

People are often exposed to several environmental exposures, for example air pollution and noise. These combined exposures might have synergistic, additive or antagonistic effects. However, until now, no methodology for the estimation of combined effects is available for HIA.

3. Identify and characterise the population at risk

Generally, the population at risk may be defined as all those living in areas where the relevant pollution levels have changed (4). In most HIAs, the impacts of an exposure are assessed for the total population concerned. However, highly vulnerable individuals (or highly exposed individuals) within the population may be affected to a much greater extent (116). In air pollution and health, people with a lower social economic status are often exposed to higher levels of air pollution. HIAs should consider these health inequalities and HIAs need to include an analysis of the health impacts on the most vulnerable population groups within the population at risk. This is also very important as issues of environmental justice become increasingly important in public decision making (6, 117, 118).

4. Select or develop a suitable set of exposure response functions (ERFs)

According to WHO guidelines (3) the ERFs may be reported as a slope of a regression line or as a relative risk for a given change in exposure. ERFs may be derived either from multicity studies, such as APHEA (119), APHEA-2 (120) and NMMAPS (62), a meta-analysis (35) or from the individual studies in the field of epidemiology and/or toxicology.

However, when the HIA is being undertaken in a local area, it is difficult to choose the most appropriate ERF. Ideally what one would like is an ERF that reflects any real differences between the local situation and studies elsewhere, but not chance differences. This suggests a weighted mean of the local-city value and for example the general meta-analysis value. Such an approach has been recently developed by APHEIS-3 (21) using the statistical concept of a shrunken estimate (4).

This issue of transferability of the estimates to populations other than the study population from which the estimate has been derived should be considered. While substantial differences in terms of susceptibility to for example air pollution seem unlikely, there are factors that may affect transferability such as: differences in daily pattern of activity, climatic conditions, housing etc, that would result in different exposures from the same ambient concentration; differences in the pollution mixture; different importance of confounding factors that might not have been properly controlled for in the epidemiological studies; different techniques in air pollution concentration measurement, and others. (4, 22, 23).

In air pollution and mortality, currently, three US cohort studies have been extensively used for HIAs (121-123). The transferability of these cohort study estimates to populations in Europe or

other regions is a concern. Differences in the mixture of pollutants, measurement methods and the extensive use of air conditioners in the US can play a role in the possible non-transferability of results. In the US, education has been found as one major effect modifier of particulate air pollution on mortality (124). However, the role of education is not well understood, and it is by no means clear that it should be expected to modify air pollution risks in the same way in Europe (116). Moreover, the US studies do not address key aspects of the exposure response relation, such as the appropriate lag time (124).

Also to derive ERFs the choice of the mathematical model should be considered. A variety of shapes of the exposure-response relation can be used, for example a simple linear model, log-linear model, square root model, spline model, threshold/non-threshold model etcetera. These basic model assumptions can have a great influence on the results of the HIA (125).

To quantify and estimate the role of different mathematical models on study-specific estimates and overall HIA results, different models should be applied in the individual (epidemiological) studies. However, epidemiological studies are rarely sufficiently abundant to provide powerful discrimination among different mathematical models, often due to low sample size (126).

Air pollution epidemiological studies underlying the ERFs, typically involve estimation of a statistical relationship between the frequency of a specific health outcome observed in a given study population and air pollution concentrations measured at fixed-site monitors in the study area (116). This estimation assumes that all individual subjects in a specific study area have the same exposure and ignores spatial variability as well as possible differences in outdoor/indoor exposure and time-activity patterns between the different subjects. Ideally, one would like to know exactly how much of a pollutant each individual comes into contact with (4). However, it is generally not possible to measure this personal exposure for large numbers of people and derive more refined ERFs. There have been some studies conducted, which try to assess exposure in a more individual way, but until now their value for deriving more refined ERFs for HIA are limited.

One method to develop and quantify a set of ERFs for HIA is to conduct a meta-analysis. The WHO has recently conducted a meta-analysis of peer reviewed studies to obtain summary estimates for certain short-term health effects linked to the exposure to particulate matter (PM) and ozone (O₃) (35). Potential problems in meta-analysis are often the lack of accurate data, an inadequate reporting in the articles, existence of publication bias, heterogeneity of results etc.

Especially, publication bias is often a problem in meta-analysis (32). Publication bias occurs when studies showing evidence for associations in a particular direction are selectively published (127). More details on systematic reviews and meta-analyses see paragraph 4.1

Ideally, to estimate ERFs by use of meta-analyses, both epidemiological and toxicological studies should be included. However, this may lead to many complex problems when integrating two whole different disciplines and paradigm's of science. In air pollution epidemiology studies exposure is often a mixture of pollutants whereas in toxicology often one chemical is examined. Other problems involve the selection of toxicological studies. Which toxicological studies should be included (in vitro studies of animal cells, in vitro studies of human cells, in vivo studies of animals etc.)? Do we have to taken into account computational methods such as quantitative structure-activity relationships and biokinetic models etc.?

Another issue which have to be considered in selecting studies for meta-analysis is which quantitative risk estimations should be used. In toxicological animal studies, it has been usual practice to select only studies with the most conservative results. Although the precautionary principle has been applied to epidemiological studies as well for the purpose of HIA, we can be sure that in the long run this approach systematically overestimates the impacts of exposure on disease or death (125). Actually, in HIA, the most appropriate aim is to provide the 'best' practicable estimates of the health impacts of a proposed policy (4).

5. Derive the population exposure distribution

In HIA, we need to use the same exposure level as the one used to derive ERF. Exposure in air pollution epidemiology studies are often measured at fixed-site monitors stations. So in most cases, we need to use that exposure for HIA. However, it is necessary to describe pollution levels across the population-at-risk and data from monitoring stations are often sparse that interpolation or extrapolation may be necessary, and modelling may be needed to link exposure with information about population density and other demographic characteristics within the population at risk (4).

6. Estimate the background rates of the relevant health disease endpoints in the population at risk

Baseline frequencies outcomes should preferably be obtained from data regarding the population for which HIA is being made. If such data is not available, health frequency data

from other populations may sometimes be used. Especially within a large at-risk (European) population, information is needed from different sources and about many different locations. These data should be selected carefully. Differences in recording and classification of health endpoints among and within countries lead to non-comparability of background rates (4, 128). Problems can also occur in the linkage between background rates and the health endpoints underlying the chosen ERFs. APHEIS found that, even if most of the cities have hospital data from registries that use quality-control programme, the lack of comparability for morbidity indicators makes it difficult to compare these cities in a HIA (22).

7. Calculate the burden of disease or death in the population at risk

Incidence, prevalence and mortality are the most quantitative parameters used to describe the burden of disease. Some HIAs have derived impacts estimated variously as attributable cases, changes in life expectancy or estimated years of life lost (YLL).

Attributable number of cases is the estimated number of cases attributed to the exposure in the population. The ERF of a certain disease, together with the proportion of the population and the background rates of a certain disease, enables calculation of the attributable cases (4). In figure 1 the general HIA model of APHEIS is shown.

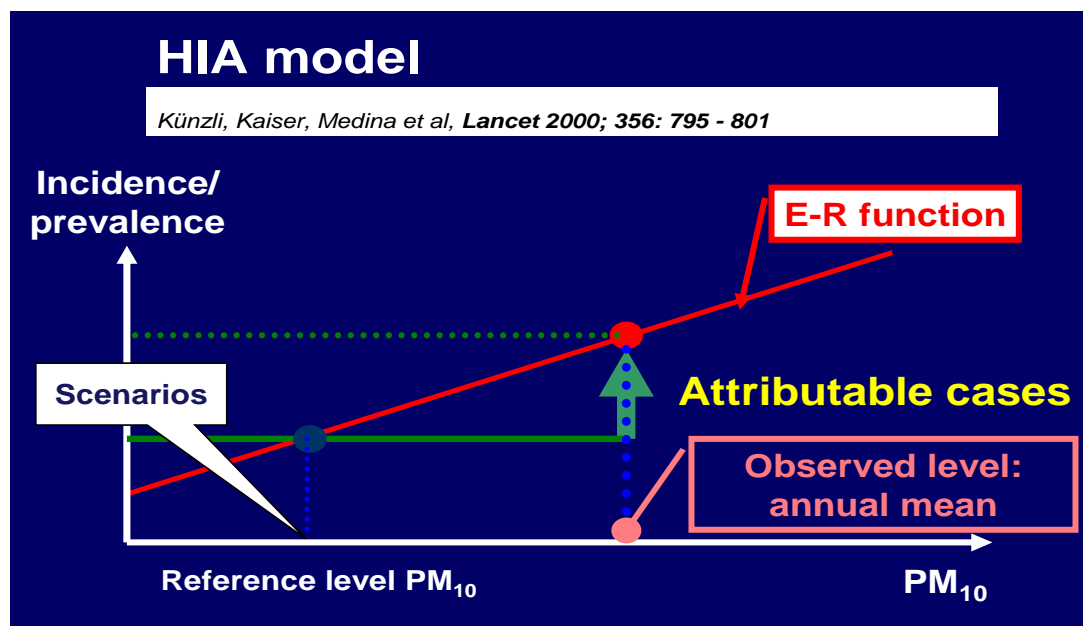


Figure 1 The HIA model how to calculate annual attributable cases (129)

Calculations of health impact measures are generally based on and derived for a given study population from the relative risks or the risk difference from the individual studies selected. RRs can better transfer to other populations than RD, because RD depends more on underlying background risks/diseases rates in the study population (4).

In HIAs competing risks have to be taken into account for cause-specific mortality (116). Ignoring competing risks may lead to an overestimation of the burden of a certain disease or death. In a study of Goldbohm et al. competing risks were taken into account by calculation of excess lifetime risk through a life table (125).

8. Value the burden of disease or death in the population at risk

Others calculate the monetary value of the estimated impacts, or some other measures (QUALYs, DALYs) which taken into account quality of life aspects (4).

In economic valuation, discounting is routinely applied and the discount rate chosen often proves to be crucial to the final result. Individuals show a clear preference for valuing a benefit that can be obtained today over a benefit that will be obtained in the future. Up to the present day, there seems to be no broad consensus which discount rate is the most appropriate in absolute terms (4).

Cost/benefit analysis aims to quantify benefits in monetary terms in order to make ‘direct’ comparisons of goods that are per se incomparable. They assume that all is exchangeable against each other. The danger is that nothing is considered indispensable but that it is just a matter of price (4).

DALYs include the years lived with disability as well as the years of life lost. These estimates are then given an age weighting and discounted accordingly. Inherent on DALYs and QUALYs are the severity/quality weightings. Some may see it as a disadvantage that this weightings approach involves social and individual preferences, and so it is not very compatible with general scientific procedures. However, health preference measurements have been found to be rather stable and reproducible, even across countries, if they carried out cautiously. Nevertheless, the low disability weightings are very sensitive to variation. In addition, they generally concern less-severe health outcomes that affect large number of people. This can lead to greater uncertainty in the assessment of mild diseases (4).

The use of DALYs and QALYs implies that people with a disease count less than healthy people. Disabled individuals are subject to double jeopardy. People with poor health suffer a disadvantage twice: first they are disabled, and secondly the saving of one year of their lives counts less than that of a healthy person (130).

When aggregate the burden of disease of for example the complex mixture of air pollution, HIAs is normally focused on individual pollutants using a single-pollutant model. Adding pollutant-specific effects may be justified when levels of the specific pollutants are clearly not correlated. However, most air pollutants are strongly correlated (PM and NO₂) and adding these pollutants leads to double counting, which should be prevented (4, 116).

9. Assess and quantify the uncertainty of the HIA

A HIA has inherent uncertainties and requires a set of assumptions. Any approximations and assumptions made in a HIA, and its consequent limitations, must be made explicit (1, 3, 117, 131, 132). They usually reported confidence intervals (CI), which reflects only the statistically uncertainty within the statistical model. Potential uncertainties, due to for example misspecification of the model, the population under study, and the measurement error in the data used to estimate the model, are not reflected in the CI. Often it is usual to distinguish between statistical uncertainty and model uncertainty. Both statistical and model uncertainties arise during the various stages of a HIA. Model uncertainty also arises in the linkages between those stages (4).

The use of CIs for the HIA results is one important aspect of the characterisation of uncertainty in a HIA as a whole. One approach to assess model uncertainties is not only to identify and discuss them, but also to carry out sensitivity analyses. Sensitivity analyses are a useful tool to obtain insight into the impact of assumptions made and the variability of the underlying data (4).

A more comprehensive analysis shows that key dimensions of uncertainty can be classified as technical (inexactness), methodological (unreliability), epistemological (ignorance), and societal (social robustness) uncertainty. Quantitative methods, such as Monte Carlo analysis, address the technical dimension only. They can be complemented with new qualitative approaches, resulting in for example a novel approach to uncertainty assessment known as the NUSAP (Numeral Unit Spread Assessment Pedigree) method (100).

Appendix 2 Hazard identification

Hazard assessment comprises Hazard identification and dose – response assessment. Hazard identification is defined as the identification, from animal and human studies, *in vitro* studies and structure – activity relationships, of adverse health effects associated with exposure to an agent (133). In the case of chemicals, not only the agent itself but also the breakdown products may need to be assessed as well. Hazard identification involves determining:

- what types of (adverse) health effects might be caused by the problem; and
- how quickly the problems might be experienced (134).

According to US EPA (135), hazard identification examines the capacity of an agent to cause adverse health effects in humans and animals. It is a qualitative description based on the type and quality of the data, complementary information (e.g. structure-activity analysis, genetic toxicity, pharmacokinetic) and the weight of evidence from these various sources. Key issues include:

- the nature, reliability and consistency of human and animal studies;
- the availability of information about the mechanistic basis for activity; and
- the relevance of the animal studies to humans.

Hazard identification uses both animal and human data. Animal data are usually assessed by toxicological methods, while human data are assessed by either epidemiological methods, when groups of people are involved, or by toxicological methods when using case studies and acute chamber studies. Other data, such as structure-activity data or *in vitro* data assessed by toxicologists, may also be included. However, Hazard Identification mostly relies on the results of *in vivo* toxicity studies conducted according to standard protocols (e.g.(136). The data may come from a range of sources, such as ad hoc data, anecdotal data, case-report data and data collected from epidemiological registries (such as cancer or pregnancy outcome data). In each instance, the quality of the study design and methodology and the resulting data will need to be rigorously assessed.

Differences in hazard identification based on epidemiological and toxicological data may be seen in the matter of “site concordance”. The epidemiological data may suggest lung cancer is of concern whereas the toxicological data may suggest liver cancer. Similar conflicts can arise when there are suggestions of a problem from epidemiological data unsupported by toxicological evidence (126).

Generally speaking, epidemiology has a number of potential advantages over animal toxicology in the area of hazard identification:

- it directly assesses human health risk;
- absorption, metabolism, detoxification and excretion may vary between humans and the animal species studied does not need to be taken into account in epidemiological studies;
- sample sizes for human studies may be much larger than those available for animal studies;
- genetic diversity may be broad in humans compared to selected animal strains used in toxicological studies;
- epidemiological studies may include different groups (e.g. the young, old and susceptible) that may not be included in the usually relatively homogeneous groups used in toxicological studies; and
- effects in some aspects of mental function or behavior, and more subjective effects such as nausea or headache, can be better assessed in human studies.

On the other hand, when Hazard Identification is based on toxicological appraisals a series of important questions should be thoroughly assessed. Such questions are summarized in the following checklist which has been adapted with slightly modification from US EPA (135):

1. what is the key toxicological study (or studies) that provide the basis for health concerns?
 - how good is the key study?
 - are the data from laboratory or field studies? Are the data from single species or multiple species?
 - in case of a carcinogenic hazard, evidence is coming from observation of single or multiple tumour sites? Benign or malignant tumours occurred? Were

there any tumour types not linked to carcinogenicity? Was the maximum tolerated dose used?

- if the hazard is other than carcinogenic, what endpoints were observed and what is the basis for the critical effect?
 - availability of other studies that support this finding?
 - availability of valid studies which conflict with this finding?
 - as many relevant studies as possible should be collected and rigorously assessed as to their strengths and weaknesses to determine the key studies. This is particularly important where quantitative risk estimates will be undertaken or where there are apparently contradictory studies. In the latter case, the studies that are considered to be adequate in their design and interpretation will need to be appraised to determine the overall weight-of-evidence.
2. besides the health effect observed in the key study, are there other health endpoints of concern?
- what are the significant data gaps?
3. are there any relevant epidemiological or clinical data available? For epidemiological studies:
- what type of studies were used, i.e. ecologic, case-control, cohort?
 - were exposures adequately described?
 - were confounding factors adequately accounted for?
 - were other causal factors excluded?
4. how much is known about the biological mechanism by which the agent produces adverse effects?
- are there any relevant studies, including metabolism studies, on mechanisms of action?
 - does this information aid to the interpretation of the toxicity data?
 - what are the implications for potential health effects?
5. are there any negative or equivocal findings in animals or humans? Were these data considered in the Hazard identification? (137)

Appendix 3 WHO meta-analysis

WHO project ‘Systematic review of health aspects of air pollution in Europe.’

Meta-analysis of time-series and panel studies of particulate matter (PM) and ozone (O₃) (35).

The WHO task group developed search criteria to identify relevant studies; guidelines for study selection were discussed and determined. Questions which have been addressed include the following questions:

- which outcomes should be included?
- which pollutants should be included?
- which lags should be chosen?
- which averaging time should be chosen?
- from which geographical areas should studies be selected?
- if more than one study is available for a city/region, which one should be used?
- to what extent can differing definition of a disease category be combined?
- to what extent should different seasons be analysed?
- how to deal with publication bias?
- how to deal with heterogeneity?

Fixed and random-effects summary estimates were calculated for a 10 µg/m³ increase in the pollutant. Results were dominated by large multicity studies like the PEACE study (138) and the APHEA 2 study (120). For a more detailed description of this meta-analysis, we refer to the original paper including the annexes (35).

Appendix 4 Possible sources of uncertainty within individual studies

1. Observational studies

Two broad types of error afflict individual observational studies: random error and systematic error. Random errors affect the precision of a study. Because it is not feasible to study an entire population, a sample of the population is chosen. Random sampling error can result and reflects variability or chance variation that may occur from sample to sample. Studies with a small sample size are more prone to this type of error.

In observational studies, systematic error (bias) is typically a much larger problem than random error and thus, we concentrate on this type of error.

Systematic errors are due to problems of the internal validity of a study (in comparison with the external validity: generalisation of a study). Many classification schemes for some possible systematic errors have been published in the literature but they have all their drawbacks as there are innumerable **types of biases** and it is often **difficult to make a distinction** between certain forms. Already in 1979, Sackett et al. (139) has reviewed more than 50 different types of bias.

While different types of bias overlap, it is useful to classify bias into three broad categories: and use these categories as a starting point for possible sources of bias in the environmental epidemiology:

- selection bias
- information bias
- confounding

Within these broad categories we can distinguish several common types of biases, as specific as possible, which can affect environmental observational studies.

Selection bias is systematic error in the process of identifying the study population. The preferential selection of subjects is related to their case/control status and/or their exposure status (140). Different types and subtypes of selection bias included for example:

- Ascertainment bias
Ascertainment bias can bias the results when the person assessing the health outcome, whether an investigator or the participant itself, knows already the exposure status.

- Participation bias

Participation bias can bias the results when participants from the study are different than the refusers, both in terms of exposure and response (and confounders).

- Choice of comparison group

Famous example of this is the healthy worker effect; for example the frequent use of general population rates as the referent group for occupational studies often introduces a form of selection bias called healthy worker effect.

- Loss to follow-up bias

Bias due to differences in completeness of **follow-up** between the participants and that this is related to personal characteristics, outcome and /or exposure. Bias due to loss of follow-up can be a major source of bias in prospective cohort studies.

Information bias results from differences in the methods in which information is collected about or from participants and may result in misclassification. The bias can originate from the researchers or interviewers, the methods of data collection, or from the subjects itself. Misclassification can be differential if it is related to exposure or outcome, or non-differential if it is unrelated. Different types of **information bias** included:

- Measurement bias of exposure → exposure misclassification

1. Is the (measured/modelled) exposure relevant and specific to the etiologic hypothesis and disease outcome (biological relevant exposure, also in respect with lag time). For example cognitive function in the elderly might be related to cumulative lead exposure, which is better assessed by bone lead (with a half-life of 10-15 years) than blood lead (half-life of 45 days measurements) or lead in the air.
2. Is the chosen indicator of exposure a good representation of the often complex mixture people are exposed to in the real life?
3. Is the chosen exposure metric (e.g. the 1-hr, 8-hr or 24-hr average/ peak concentration / cumulative exposure) a relevant representation?
4. Is the (measured/modelled) exposure a good proxy for the average individual (personal) exposure/dose?

- Measurement bias of health effect → misclassification of health effect

There is a range of methods to measure health effects. The subject can be asked whether he/she suffers from a particular disease. Alternatively, diagnoses made by specialists can be used or the causes of death recorded on the death certificate can be used. All these different methods have their own uncertainties (19).

Sometimes it is difficult to disentangle certain types of health effect (e.g. in the case of cardiovascular and respiratory endpoints), leading to some measurement bias of health effects.

- Recall bias

Recall bias occur when exposure information is differentially misclassified for cases and non-cases. It can happen in all epidemiological studies, but in the epidemiological case-control design recall bias is most likely; with the assessment of exposure relying mainly on participants' self-reports from the past. Cases may reflect about potential causes of their disease and therefore may call former exposures more accurately or even estimate these exposures. Controls would tend to forget former exposures, particularly when the exposure was a long time ago. (can be the case of a case-control study and a retrospective cohort when the exposure and disease have already occurred)

- Interviewer bias

Investigators can ask cases and control differently about their exposures. (can be the case of a case-control study and a retrospective cohort when the exposure and disease have already occurred)

Confounding is another type of systematic error. It results when additional factors or variables are associated with both the exposure and independently with the disease status, and when these variables are not intermediate factors between the exposure and health effect of interest. (Main) confounding factors (including possible effect modifiers) have to be identified and accounted for in the analysis of a study. Often the general main confounding factors are gender, age, smoking, and other pollutant factors. Uncertainties can arise if there are:

- known confounding factors, but no data available to correct for it
- unknown confounding factors

- misclassification for confounding factors (inadequate data or e.g. only measured at baseline, inappropriate level) leading to residual confounding

Errors in confounders compromise the ability to control for their effect, leaving residual confounding. For example the ratio adjusted with the approximate confounder will on average lie between the crude, unadjusted ratio and the ratio adjusted with the true (unknown) confounder. There are a few exceptions. Entirely systematic error (everyone underreporting their smoking; basically never the case) will not usually compromise control for confounding (141, 142).

2. Toxicological studies

We make the distinction between controlled animal experiments and controlled human toxicological studies. Unfortunately, there are no widely used guidelines to evaluate the quality of animal experiments and different kinds of potential biases in contrast with the bulk of literature about systematic errors/biases in observational studies (31). However, there are some attempts for example a whole issue of the ILAR journal (Institute for Laboratory Animal Research; Vol 43, No 3, 2002) is dedicated to give guidance in the design and analysis of animal toxicology studies and an approach for evaluating the quality of toxicological studies has been proposed (143). In the UK, the National Centre for the Replacement, Refinement and Reduction of Animals in Research is working towards improvements in experimental design (144) and guidelines for animal testing are also available from a number of US and international organizations, such as the Center for Alternatives to Animal Testing (CAAT) based at the John Hopkins University (145) and the Organisation for Economic Co-operation and Development (OECD) (146)

Important methodological problems of animal experiments (apart from the extrapolation uncertainties, see below) are:

- how clearly the agent was defined and, in the case of mixtures, how adequately the sample characterization was reported
- whether adequate animals and animals strains were used
- whether the dose was monitored adequately, particularly in inhalation experiments
- whether the dose, duration of exposure and route of exposure were appropriate

- whether there were adequately numbers of animals per group
- whether the choice of the comparison group was appropriate
- whether animals were allocated randomly to groups and the investigators blinded for exposure and health effect
- whether the duration and timing of observation of the health outcomes was adequate and
- whether the data were reported and analysed adequately

Extrapolation uncertainties can involve:

- extrapolations from animal to human (interspecies variations) (e.g. differences in breathing rates, organ sizes, basal metabolism, life spans). Also the question which animal species is the most relevant in terms of the ‘susceptible’ human given a particular health outcome involves inherent uncertainty
- extrapolation from human to human (difference in sensitivity, intraspecies variations)
- extrapolations from high to low doses
Often the exposures/concentrations/doses are several orders of magnitude higher in animal studies than exposures in the real world (e.g. ambient air, household consumer product)
- extrapolations from short-term to long-term exposure
Most animal studies have investigated exposures-health effects for a relatively short time period (days, weeks).
- extrapolations from one pollutant in the laboratory setting to an often complex mixture of pollutants in the real world

In the case of **controlled human experiments** some extrapolation uncertainties are not valid anymore, as they are investigated the right specie. However, most of the issues described above for animal studies also play a role in the controlled human experiments. One difference is the possible bias introduced by the use of volunteers in the experiment.

Of course this is a general list and not all of these uncertainties may be met in the individual studies. One illustrated example is the studies were they use the ambient fine particle concentrator to manage ambient air in such a way that the particle concentration is multiplied many times. One of the major strengths is that they are ‘real’ world particles (147).