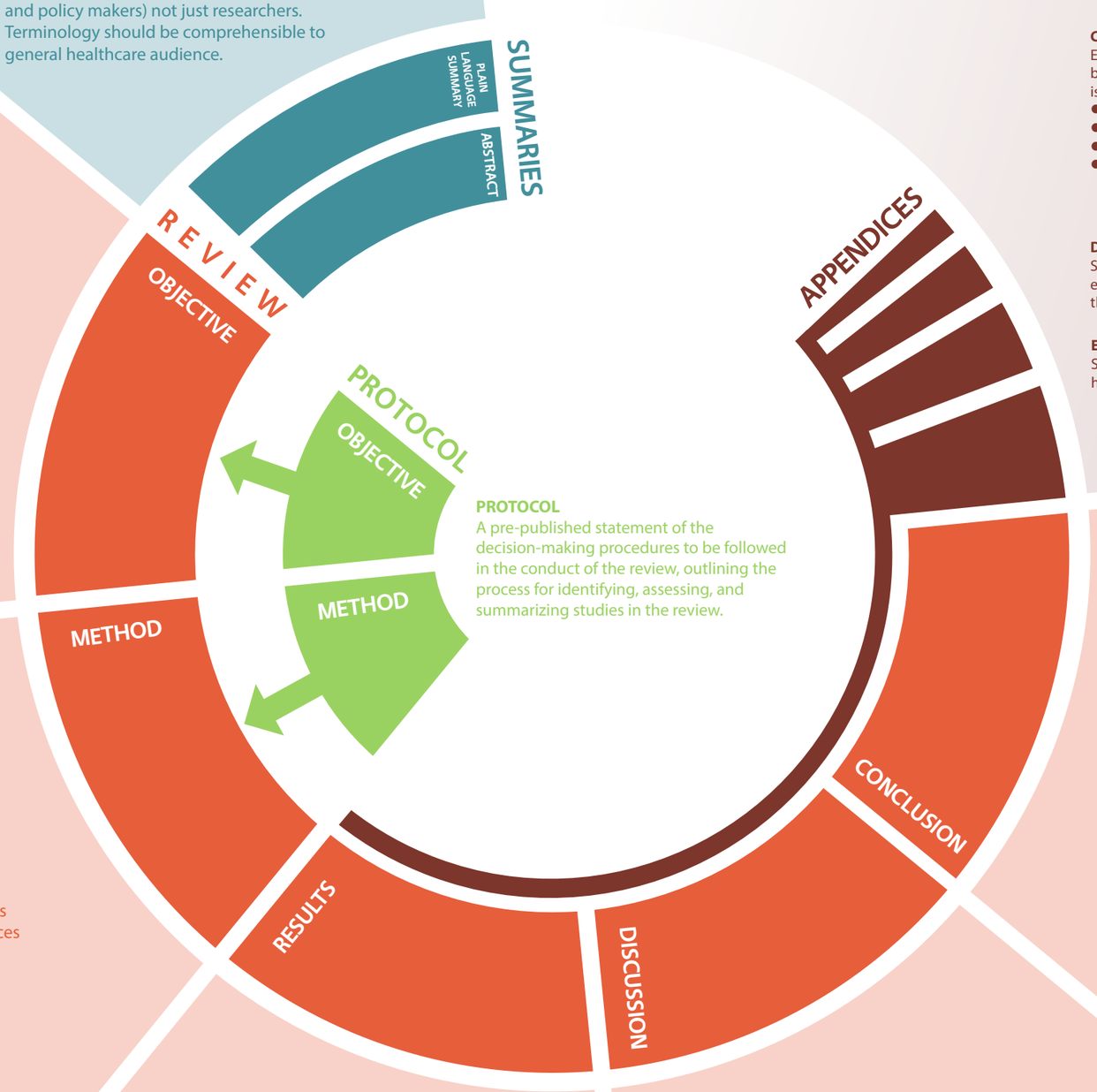
Systematic review and the future of evidence in chemicals policy

How systematic review techniques used in evidence-based medicine can advance the credibility and utility of chemical risk assessments, bringing us closer to a European Union in which chemicals policy is routinely based on the best available evidence

Paul Whaley
The Policy from Science Project

The anatomy of a Cochrane Review

Adapted from Higgins, Green (2008), chapter 4.



PLAIN LANGUAGE SUMMARY
Summarises the review in a straightforward style that can be understood by consumers of health care. Consists of a title, restating the original title using plain language terms, and a summary text of up to one page in length.

ABSTRACT
Summary for publication. Targeted at healthcare decision-makers (clinicians, informed consumers and policy makers) not just researchers. Terminology should be comprehensible to general healthcare audience.

OBJECTIVE
A precise statement of the primary objective of the review, ideally in a single sentence. The objective should be justified by a concise (one page) background statement, which sets the rationale for the review and justifies the specific formulation of the question.
- Description of the condition
- Description of the intervention
- How the intervention might work
- Why the review is important
- Reference to previous reviews

PROTOCOL
A pre-published statement of the decision-making procedures to be followed in the conduct of the review, outlining the process for identifying, assessing, and summarizing studies in the review.

METHOD
A description of what was done to obtain the results and conclusions of the current review. Any differences between protocol and method must be identified.
- The criteria studies have to meet in order to be considered for the review
- Search methods for the identification of studies
- Data collection and analysis, including meta-analysis if deemed appropriate

RESULTS
A description of the studies considered in the review.
- Summary of results of the search (number of studies retrieved, number considered eligible after screening)
- Statement of number of included studies, with succinct summary of "Characteristics of Included Studies" table
- List of excluded studies, which appear to meet criteria for inclusion but do not; reason for exclusion of each study (normally one is enough)
- Risk of bias analysis (table and chart is optional but highly recommended)

DISCUSSION
A structured discussion to aid the interpretation of the review.
- Summary of main results
- Overall completeness and applicability of evidence
- Overall quality of the body evidence
- Potential biases in the review process
- Agreements and disagreements with other studies or reviews

AUTHORS' CONCLUSIONS
Presentation of the significance of the information gleaned from the review.
- Implications for practice
- Implications for research

A: OTHER INFORMATION
Other information included as standard in a Cochrane Collaboration systematic review are:
- Authors and contact person
- Acknowledgements
- Contributions of authors
- Declarations of interest

B: TABLES
Tables are very important for the clear presentation of summarised data.
- Characteristics of included studies
- Risk of bias of included studies
- Characteristics of excluded studies
- Characteristics of studies awaiting classification
- Characteristics of ongoing studies
- Summary of findings

C: STUDIES AND REFERENCES
Each study which either does or would reasonably be expected to form part of the data in the review is accounted for, with tables listing:
- Included studies
- Excluded studies
- Studies awaiting classification
- Ongoing studies

D: OTHER STUDIES
Studies referenced in the review for e.g. explanatory purposes but do not constitute the data being reviewed.

E: DATA AND ANALYSES
Supplementary information showing whether and how meta-analyses are performed in the review.

F: FIGURES
- Illustrative figures with captions.
- Plots and graphs: funnel; forest; risk of bias graph and summary; other figures.

G: SOURCES OF SUPPORT
Acknowledgement of any sources of support for the review, including grants, material support, salary etc., both internal to the institution at which the review was conducted, and any external support.

H: FEEDBACK
Each piece of Feedback incorporated into a review is identified by a short title and a date. The summary of the feedback and authors' reply, and contributors to the reply are given.

I: SUPPLEMENTARY INFORMATION
Appendices provide a place for supplementary information such as detailed search strategies, lengthy details of non-standard statistical methods, data collection forms and details of outcomes such as measurement scales.

This diagram was produced as part of the Policy from Science Project
www.policyfromscience.com

Systematic review and the future of evidence in chemicals policy

How systematic review techniques used in evidence-based medicine can advance the credibility and utility of chemical risk assessments, bringing us closer to a European Union in which chemicals policy is routinely based on the best available evidence

Paul Whaley
Research Lead, Policy from Science Project
p.whaley@lancaster.ac.uk

Foreword

Our organisation, Réseau Environnement Santé (RES), was created to disseminate the growing knowledge-base being developed by the environmental health sciences and accordingly to support the update of health and environmental policies.

In France, RES has been closely involved in the national debate concerning the safety of Bisphenol A (BPA), which has ultimately led to ANSES' unfavourable re-evaluation of the hazards and risks posed by BPA, paving the way for a national ban in food contact materials.

In this course, BPA has not only become the poster child of emerging concerns on the hazards of endocrine disrupting chemicals, but is also seen to embody the dysfunctional nature of EU risk assessment and risk management procedures (well summarized in the BPA chapter of volume II of the European Environment Agency's Late Lessons from Early Warnings).

That this has happened during a period when EU's food safety authority EFSA has been challenged by MEPs and civil society to resolve problems with the conflicted interests of board members and panel experts, is doing little to increase public trust in the agency's first response to the BPA case.

In the near future, we can expect EU regulations to be equipped with updated test requirements and ad-hoc revisions to procedure, in order to capture and tackle the hazards of EDCs. We can also expect institutional bodies such as EFSA to drastically improve their internal rules on how to deal with conflicts of interest.

Further up, in a perfect world, public and independent research into chemical safety would be receiving sufficient funding, while regulatory instruments will become flexible enough to react promptly to early warnings and digest readily the advances of science.

In the meantime, we should welcome any toolkit that can help us move forward pragmatically to the most appropriate decision-making, despite all the imperfections and controversy. As science will always be about both knowledge and uncertainty, systematic review methodologies offer a meaningful interpretation to the growing EU regulatory focus on "weight-of-evidence" and "science-based" approaches.

Thus, if Evidence-Based Medicine can help medical and pharmaceutical research to walk less blindly in the darkness of our ignorance, why not learn from it and sow the seeds of an evidence-based approach to toxicology?

That is our objective in getting involved in the Policy from Science Project and supporting the work of Mr Paul Whaley, author of this report. We hope to have delivered a valuable contribution that will stir up a process of applying systematic review techniques to the risk assessment of chemicals, for the better regulation and for the provision of conditions for restoring public trust in the regulatory process.

Yannick Vicaire, November 2013.

Executive Summary

This report advocates the use of systematic review techniques first developed for use in medicine as a new approach to reviewing evidence in the conduct of chemical risk assessment, in order to strengthen the connection between the decisions made in chemicals policy and the evidence base which supports them.

Chemicals policy is increasingly characterized by controversy rather than consensus (part 1). For chemicals such as BPA, we see a range of opinions as to its safety, from EFSA's position that it poses no threat to health at current exposure levels, to Swedish regulators even banning its use in thermal paper. This diversity of opinion exists in spite of everyone having, at least in theory, access to the same evidence base.

The same problem has been faced in medicine (part 2), where there are many examples of how decisions made in healthcare have failed to match those which were best supported by the available evidence. The cause of the problem was determined to be a general failure to use scientific methods for identifying, appraising or synthesising information when conducting reviews of the literature. The solution? To develop systematic review techniques, the application of the basic scientific principle of using a reproducible methodology to the process of reviewing evidence.

In medicine, systematic review techniques cover seven basic elements: a clearly-stated objective; the use of a pre-published protocol defining the methods to be used in the review; a systematic search for evidence; clear criteria for electing evidence for inclusion in the review; an assessment of the methodological quality of the included studies; and a systematic synthesis of data and presentation of results.

A comparison between EFSA's recent Scientific Opinions on BPA (part 3) with a scientific approach to reviewing evidence produces a similar result as to that which was seen in medicine: review objectives are not sufficiently clearly stated; there are no pre-published protocols; methods for locating data are not consistently given; the criteria for selecting data for analysis are incompletely stated; how studies are evaluated for quality appears to be neither transparent nor consistent; the synthesis and presentation of results is unclear.

Without a transparent, reproducible method for evaluating toxicological data, it is not possible to be confident that the decisions made in chemicals policy are with those which are best supported by the evidence. As was the case in medicine, the solution is to develop systematic review techniques for reviewing toxicological data (part 4).

On this basis, we recommend the following measures to strengthen Scientific Opinions:

1. **In advance of developing Opinions, EU Agencies should publish and publicly consult on a review protocol, to cover: review objectives; methods to be used in searching for evidence; criteria for including and excluding evidence in the analysis; criteria for appraising the quality of the evidence; and the method for synthesising the evidence.**

2. **As a precondition of achieving a scientifically-robust review process, a toolkit for appraising the quality and directness of both individual studies and an overall body of evidence needs to be developed and validated.**

3. **Guidance should be issued for Working Groups and Scientific Committees on the structure and writing of Opinions, in order to enhance usability for stakeholders in chemical regulation.**

4. **Controls on the interests of the authors of Opinions need to be tightened, restricting direct financial conflicts of interest and developing further policies for the management and declaration of all interests which could be perceived by a user to have influence on the conclusions of an Opinion.**

5. **All decisions made in the review process need to be sufficiently documented so as to be transparent; all methodological considerations, such as criteria for inclusion and quality assessment of studies, must be consistently applied throughout.**

6. **An editorial and peer-review process for revising and accepting Opinions needs to be instituted, to ensure that published Opinions meet the quality criteria outlined above.**

There are a number of research initiatives in addition to the Policy from Science Project which can contribute to this process, with particular progress on review protocols being made by the Navigation Guide (University of California San Francisco, US), the Evidence Based Toxicology Collaboration and the US National Toxicology Panel. EFSA has also begun work in this area.

As a timetable for change, in the short term all imminent Scientific Opinions, including the next Scientific Opinion by EFSA on BPA (the hazard component of the overall risk assessment) should be structured to maximize ease of understanding. They should also include a comprehensive declaration of interests, present the full results of the evidence search and selection processes, and have a clear description of the methods used for appraising and synthesizing the studies included in the Opinions.

To support this programme, EU scientific staff and experts should receive training in systematic review techniques. In addition, researchers should be piloting more systematic reviews.

In the medium term, Scientific Opinions should be conducted according to pre-published protocols, developed in an open consultative process. Funding should be made available for education and research in systematic review methods.

In the long term an organization of similar function to medicine's Cochrane Collaboration needs to be established, to facilitate the production of and set the standards for systematic reviews of toxicological evidence.

Contents

Acknowledgments

This report originates from a concept first put to me in early 2010 by Rachel Stancliffe and Mark Starr of the Centre for Sustainable Healthcare, who saw the possibilities for how techniques used in evidence-based medicine might help clarify the evidence for a range of environmental concerns around the provision of healthcare. Since then, dozens of people have been involved in one way or another in the development of this work. There are too many to mention but in addition to Rachel and Mark, the following people deserve particular recognition for the support and encouragement they have given:

- Patrice Sutton
- Sebastian Hoffman
- David Taylor
- Marlene Agerstrand
- Jamie Page
- Yannick Vicaire
- Ninja Reineke
- Crispin Halsall
- Ruth Alcock
- Miriam Sturdee

I would like to thank Marlene Agerstrand, Sebastian Hoffman, Yannick Vicaire, David Taylor and Crispin Halsall for taking the time to review this report. Any shortcomings or errors which have since persisted are entirely of my own doing.

Paul Whaley. November 2013

Preface

Preface

0

Preface

Nobody benefits when the decisions made in regulating chemicals become routinely detached from the scientific evidence which might support or oppose them. It creates financial risk for companies because they cannot anticipate which chemicals will be restricted and which declared safe for use; it presents risks to health and the environment because regulatory controls do not reliably distinguish between those chemicals which are harmful and those which are sufficiently safe; it even presents a risk to regulators themselves, because inconsistent decision-making and the absence of a demonstrably impartial process undermines public trust in regulatory authority.

There is therefore a powerful and mutual interest in forging as tight a connection as possible between decisions made by regulators and the evidence-base which supports them.

This report advocates a new approach to forging that connection: the adaptation of systematic review techniques used in medicine to the synthesis of evidence in chemical risk assessment.

The aim of the report is not to present a complete solution to the challenges of using evidence in making chemicals policy. Instead, the intention is to encourage open discussion of how systematic review techniques can improve how all of us evaluate and communicate the complex body of evidence that goes into decision-making around chemical use, from MEPs and risk managers to expert committees, industry groups and environmental advocates.

Our intention is to encourage open discussion of how systematic review techniques can improve how all of us evaluate and communicate the complex body of evidence that goes into decision-making around chemical use

This report is critical of standard processes in reviewing toxicological data, using two EFSA Scientific Opinions on BPA as case studies. It is not, however, intended as an attack on anyone's competence or integrity, as it is well understood that EFSA's Opinions are put together at tremendous effort.

Equally, however, it should be clear that there are more and less scientific ways of synthesising research into Opinions and reviews. Many of these techniques are new (only really introduced even to medicine in the last 25 years) and the science of research synthesis is rapidly evolving. So while nobody should feel they are being blamed for limitations in the transparency and reproducibility of methods used for generating Opinions and reviews, many of the methods used to synthesise data and generate Opinions have been, and still are, less than scientific with substantial improvement both possible and necessary. As scientists, policy-makers, NGOs, manufacturers and users of chemicals, it is hoped that we all equally relish the challenges this involves.

0.1 Declaration of interests

Project funding: The report is part of the deliverables for the Policy from Science Project, funded by Fondation pour une Terre Humaine and the European Environment and Health Initiative (EEHI), and administered by Réseau Environnement Santé. The scope of the report was developed in discussion with the funders, though the funders had no input into the drafting of the report itself.

Financial: In the last five years, the author has been heavily involved in work for environmental NGOs advocating changes to various elements of chemicals policy. The author has carried out consultancy work for: ChemSec (researcher and project manager relating to flame retardant restrictions under the RoHS Directive); the Cancer Prevention and Education Society (retainer as Scientific Advisor, editor and producer of Health & Environment); UK Standardisation Network for Sustainability (Technical Committee Expert Representative working on flame retardants in flammability standards); Centre for Sustainable Healthcare (researcher and writer, working on systematic review techniques for risk assessment of chemicals); New Harbour (researcher, looking at the environmental profile of rayon manufacture); Health Care Without Harm (researcher, looking at toxicological data relating to the use of endocrine disrupting chemicals in medical devices). Ending 31 December 2009, the author was employed by Health Care Without Harm on projects relating to the use of chemicals in healthcare.

Academic: The author is working on a part-time PhD at Lancaster University on systematic review techniques for chemical risk assessment and is applying for grants to develop research capacity at Lancaster Environment Centre in this area.

Part One.

1 The need for systematic review

"We are, through the media, as ordinary citizens, confronted daily with controversy and debate across a whole spectrum of public policy issues. But typically, we have no access to any form of systematic 'evidence base' — and therefore no means of participating in the debate in a mature and informed manner." Adrian Smith, professor of statistics at Imperial College London, quoted in Chalmers et al. (2002).

1 The need for systematic review

1.1 Who should you believe when it comes to chemical safety?

1.1.1 The personal experience of an environmental consultant

1.1.1.1 Troubles with BPA

Speaking personally as the author of this report, when people find out I am an environmental health researcher and writer with a specific interest in the effects of environmental chemicals on human health, I start getting a lot of questions from people about compounds like BPA. They ask me: Should I drink from polycarbonate bottles? Will I benefit from eating less canned food? Should I refuse till receipts? Is exposure to BPA giving people cancer?

Embarrassingly, in spite of there being 6,000+ studies on the compound, I don't have any emphatic answers. I can describe tests which set a supposedly safe daily intake for BPA, and describe weaknesses in those tests which mean they may over- or underestimate this intake. I can even describe weaknesses in the tests which undermine the TDI tests, which suggests these tests may not in fact undermine the TDI after all.

So even though I have a reasonable working knowledge of BPA's toxicity, this doesn't seem to help very much in answering the questions which have been put to me, so I usually end up saying that it's best not to expose yourself to something if you aren't reasonably sure that it is safe. (But even this isn't straightforwardly true, if you consider how the relatively untested compound BPS is now being substituted for the heavily-researched BPA.)

Maybe, given that I am not a BPA specialist, I shouldn't be offering any sort of opinion myself at all – instead, I should point people towards the opinion of one of the many experts who know more about BPA than I do.

But which one?

Maybe, given that I am not a BPA specialist, I shouldn't be offering any sort of opinion myself at all – instead, I should point people towards the opinion of one of the many experts who know more about BPA than I do.

But which one?

Perhaps I could refer them to the European Food Safety Authority, whose "most critical commitment is to provide objective and independent science-based advice and clear communication grounded in the most up-to-date scientific information and knowledge" (European Food Safety Authority 2013).

In 2010, they reviewed recent scientific literature in terms of relevance for the risk assessment of BPA and its impact on tolerable daily intake (TDI) and "based on this comprehensive evaluation of recent toxicity data […] concluded that no new study could be identified, which would call for a revision of the current TDI" such that current exposure levels pose no risk to people's health. (EFSA Panel on Food Contact Materials 2010).

The problem is, for a supposedly definitive statement by a leading EU authority, it's not necessary to go very far to find an expert who disputes it. In this case, one need go no further than the back pages of the Opinion itself, where the Minority Opinion states that although current evidence indeed does not permit a new TDI to be calculated "there are significant uncertainties about the current validity of the NOAEL" and "due to the overall weight of evidence, the current TDI of 50 µg/kg body weight may not be confirmed as a full TDI and should be considered as temporary" (EFSA Panel on Food Contact Materials 2010).

The Minority Opinion does not even seem to be much of a minority. In 2011 the French food safety authority ANSES recommended that tight restrictions be placed on the use of BPA. Although ANSES and EFSA eventually agreed that ANSES had in fact only performed a hazard assessment, French authorities nonetheless legislated a total ban on BPA in food contact materials, commencing in 2015, while there is a consensus view among a large number of environment groups including Greenpeace, the US Environmental Defence Fund and the EU Health and Environment Alliance that BPA poses enough risk of harm that it should be banned from use more-or-less across the board. Danish authorities have recently been followed by the US Food and Drug Administration in banning BPA in infant food contact materials while Swedish authorities have gone even further, announcing restrictions even on the use of BPA in thermal paper.

On the other hand, there are prominent researchers who have bemoaned the fuss about what is, in their opinion, an innocuous compound (Sharpe 2010), while the UK Food Standards Agency and various industry associations maintain that BPA is safe as currently used.

All these experts have their opinions as to the risks posed by BPA and what should be done about them – but I am not able to judge which expert is right. Working that out requires an ability to distinguish between the better opinions from the worse – but to do this, I would have to go back to the literature myself – which of course doesn't help, because I'm looking at the opinions of experts precisely because I don't trust my own ability to digest down the enormous quantity of data on BPA into a definitive statement of the toxicity of BPA.

1.1.1.2 In contrast: vitamin C

My experience with BPA contrasts starkly with my experience with medicine. I suffer quite stubborn colds, for example, so I have an interest in whether taking vitamin C will in some way prevent or treat them. Knowing a little about medicine, I know to look for Cochrane Collaboration reviews – and I find one on exactly on this topic, published in March this year (Hemilä, Chalker 2013).

This Cochrane review tells me that the effectiveness of vitamin C for preventing and treating the common cold has been of interest for decades, partly due to Nobel-prize-winning chemist Linus Pauling's interest in the matter. In that time, researchers have conducted a total of 29 high-quality placebo-controlled trials involving 11,306 participants looking at the effect of vitamin C supplementation on preventing colds. These trials show no overall effect on incidence of the common cold in the general population.

Regular supplementation does, however, seem to have a modest but consistent effect in reducing the duration of symptoms. Based on 31 study comparisons with 9745 common cold episodes, vitamin C supplementation reduces the duration of a cold by about half a day for an average of 6-7 days overall duration of the cold. Very interestingly, in 5 trials with 598 participants exposed to short periods of extreme physical stress (such as marathon runners and skiers), vitamin C halved the risk of common cold.

There is also a little bit of inconclusive evidence that high doses of vitamin C administered after onset of a cold can reduce the duration of symptoms, though more research needs to be done here if we are to know if it is truly beneficial. Likewise, harm from vitamin C supplementation is under-researched; sustained high dose supplementation may be harmful, at this stage we just don't know.

What is strange for me here is that, even though I know far more about BPA than I do about vitamin C, after reading the Cochrane Review I feel I have more confidence in how vitamin C supplements might help me with my stubborn colds than I do about whether BPA should be used in the lining of food cans.

1.1.2 What can we learn from the conduct of reviews in medicine for use in chemical risk assessment in the EU?

The difference between the confusion of opinions on BPA and the clarity provided by the Cochrane Review of vitamin C makes me wonder: is there something in the approach taken by the Cochrane Collaboration to resolving controversy, in how it evaluates and communicates complex evidence about controversial issues in healthcare, which might be applied to the evaluation and communication of chemical risk, so that a heavy user of synthesised toxicological research such as myself stands a better chance of being able to trust and use the results?

To find out, we will take a quick tour through why review techniques in medicine have developed as they have, before going into some detail on

exactly what these review techniques are and why, for the purposes of synthesizing complex bodies of evidence, they are considered superior to traditional techniques of narrative review. This should give us a working understanding of the role of systematic review methods in resolving controversies in medicine.

Emphasis will be on systematic review techniques as developed by the Cochrane Collaboration, an international network of 31000 people involved in preparing, updating and promoting the accessibility of systematic reviews of the efficacy of medical interventions (Cochrane Collaboration 2013).

We will then look at two Expert Opinions published by the European Food Safety Authority, to give us concrete examples of the strengths and weaknesses of current review practices in risk assessment as it compares to best practice in medicine, which we can use to inform a series of recommendations for reforming the evidence review process in chemical risk assessment.

Medical decision-making is not and never has been as evidence-based as one might hope or assume:

These two Opinions are important because they occur at either ends of a period of controversy for EFSA, when the agency has been subject to criticism in its handling of conflicts of interest of expert members of its scientific committees and working groups (European Ombudsman 2011), but has also initiated work on introducing systematic review techniques into its processes for conducting risk assessment (European Food Safety Authority 2010).

The case studies should reveal how much progress EFSA has made in introducing systematic review techniques into risk assessment.

1.2 A crashing realization in medicine

Medical decision-making is not and never has been as evidence-based as one might hope or assume: the history of medical care is littered with examples of missed opportunities, wasted resources and counter-productive policy which would have flown in the face of the available evidence, if only we had been better at assembling and acting on it.

1.2.1 Failures in the use of evidence in administering healthcare

The following examples, widely-cited in the literature (e.g. Goldacre 2012; Evans et al. 2011; Mulrow 1994) show how failings in the use of evidence in healthcare resulted in thousands of unnecessary deaths. There are many more – some tragic, some just a waste of resources.

1.2.1.1 Steroids and mortality in premature infants

The first fair test of the use of steroid drugs in women expected to give birth prematurely was conducted in 1972. This showed that the babies of mothers who had received the steroid were less likely to die. Over the next ten years, more trials were done but they were small and their individual results confusing. When in 1989 all the data was collected together and assessed, very strong evidence of the efficacy of the steroid treatment was revealed – but because this could have been known years earlier, tens of thousands of premature babies had died unnecessarily (Reynolds & Tansey, 2005).

1.2.1.2 Sleeping position and cot death

Dr Spock's 1956 edition of his famous book on childcare changed advice from sleeping supine to sleeping prone. The first study of sleeping manner in 1965 suggested harm from sleeping prone; in 1971 the second study also suggested harm; by 1985 three further studies suggested harm. Finally, in the mid-90s "Back to Sleep" campaigns were run in the US and UK to reverse Spock's advice (Evans et al. 2011).

1.2.1.3 Anti-arrhythmia drugs and heart attack

After a heart attack, people who develop heart rhythm abnormalities are at greater risk of death than those who do not. Since there are drugs which suppress arrhythmia, it seemed logical that they should reduce the risk of dying after a heart attack and they were therefore prescribed in large quantities. Tragically, because they actually increase risk of death (the clinical trials had only looked at reduction in arrhythmia, not mortality), at the peak of their use in the late 1980s, one estimate suggests they may have been killing more American men every year than had been killed in the entire Vietnam war (Veronesi et al. 2002); furthermore, the first systematic review of trial data in 1983 already showed no reduction in death rates – but the drugs were still prescribed until the early 90s.

Recommended reading: Evans, Imogen; Thornton, Hazel; Chalmers, Iain; Glasziou, Paul (2011): Testing Treatments. Better research for better healthcare. 2nd ed. London: Pinter & Martin Ltd.

1.2.2 The need for a new approach to reviewing evidence

These errors did not happen for want of regular reviews; indeed, in medicine the opinion of experts was sought and offered continuously in the form of practice guidelines and reviews published in the peer-reviewed literature.

So something else was going wrong, in the way that evidence was assembled (in that, it seemed not to be producing the right answers) and the way the evidence was disseminated (in that, even though there was sufficient evidence for best practice, those practices were not making it into mainstream care with sufficient alacrity).

Something was needed to make the outcomes of the review process and resultant changes in healthcare practices more reliably connected to what the available evidence actually said rather than what people simply thought it said. That something was systematic review.

Something was needed to make the outcomes of the review process and resultant changes in healthcare practices more reliably connected to what the available evidence actually said rather than what people simply thought it said. That something was systematic review.

Part Two.

Systematic review

"Science is supposed to be cumulative, but scientists only rarely cumulate evidence scientifically. This means that users of research evidence have to cope with a plethora of reports of individual studies with no systematic attempt made to present new results in the context of similar studies." (Chalmers et al. 2002)

2 Systematic review

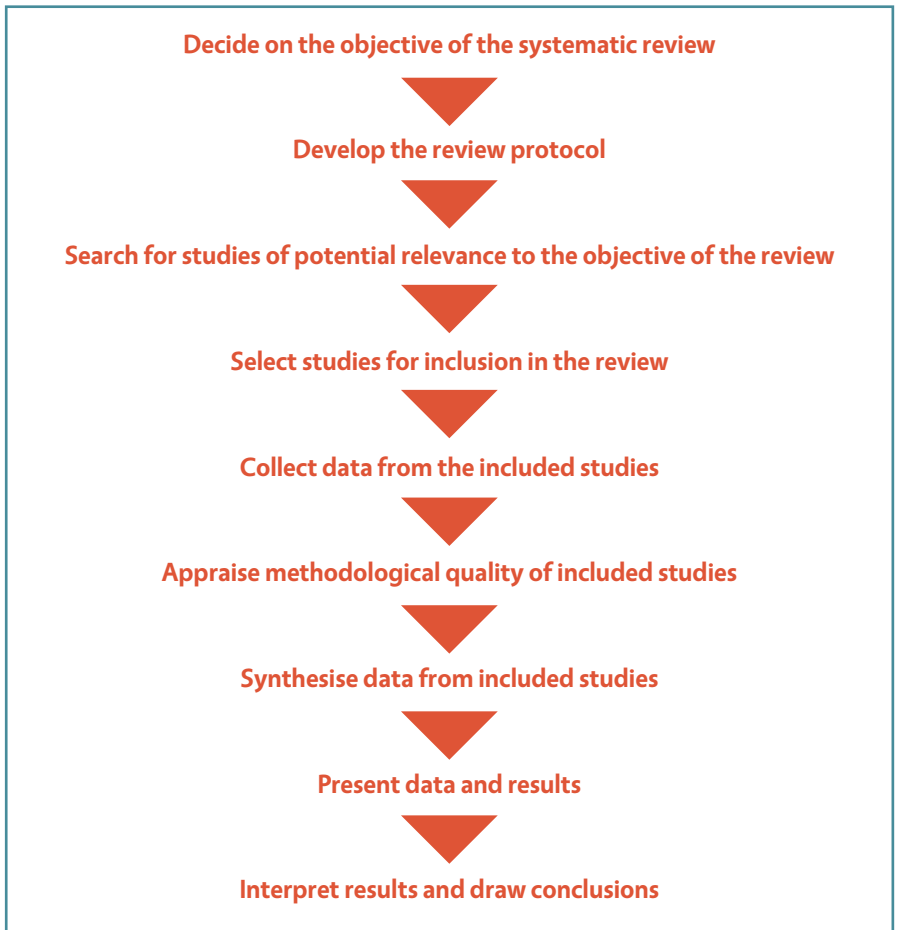
2.1 What is "systematic review"?

The fundamental concept of the systematic review is disarmingly simple: it is the use of a documented and reproducible method for synthesising evidence relating to a hypothesis.

A systematic review is therefore a literature review conducted according to scientific method. An objective and/or hypothesis is stated; a search for all the evidence is conducted; the evidence is appraised and synthesised according to a transparent method; a statement is given of whether the hypothesis is likely to be true or false.

Once a review is completed by a research team, a second team should be able to follow the first's documentation and produce the same answer (or if not, they should at least be able to explain why not).

Recommended reading: Garg, Amit X.; Hackam, Dan; Tonelli, Marcello (2008): Systematic review and meta-analysis: when one study is just not enough. In Clin J Am Soc Nephrol 3 (1), pp. 253–260.



The systematic review process. Adapted from European Food Safety Authority (2010).

2.1.1 Why bother with systematic reviews?

The process of conducting a systematic review may look like a lengthy and involved process (though not necessarily obscenely so – many are conducted by two primary authors within the space of a year; on average they take about 18 months). So why bother? The following reasons represent a selection of those described in the literature. (See for example: Bero et al. 1998; Evans et al. 2011; Garg et al. 2008; Greenhalgh 2010; Kane 1995; Mulrow 1987; Mulrow 1994; Mulrow et al. 1997; Rennie, Chalmers 2009; Woolf 2000; Hartung 2009.)

To manage data volume. There are too many publications in any given field for a single working person to take into account when making a decision. This information has to be reduced down into manageable quantities to allow users of the information to make timely, evidence-based decisions. Without integrating this data, as a society we will not reliably keep abreast of developments, justify and refine hypotheses, avoid pitfalls in previous work, identify adverse effects and covariates, or formulate effective guidelines and treatment strategies. (That said, if there are only two data points in an area of research, a systematic review is probably not necessary – though it might take a systematic search to determine if this is the case.)

The fundamental concept of the systematic review is disarmingly simple: it is the use of a documented and reproducible method for synthesising evidence relating to a hypothesis.

A systematic review is therefore a literature review conducted according to scientific method.

To be cost-effective. Although sometimes arduous, it is usually quicker and cheaper to conduct a systematic review than do unnecessary research (because this is a waste of money) or conduct a review which is not sufficiently systematic (where you incur the costs of being wrong). To put it another way: it is cheaper to spend more on being right and relevant than less on being wrong or redundant.

To generalise. Multiple reviewed studies provide an interpretative context not available from single studies. Because the tendency of effects to occur in the same direction and be of the same magnitude given the variations in study methods can be determined, the consistency of relationships can be assessed.

To increase power. By pooling the results of individual, less certain observations, a more definitive effect size can be calculated, increasing precision in estimates of risk or effect size.

To improve accuracy. Systematic reviews are distinguished from narrative reviews by the application of explicit principles aimed at reducing random and systematic error. While it is difficult to prove this yields greater accuracy, the findings of traditional reviews tend to lag behind or differ from systematic reviews. In addition, explicit methods allow assessment of what was done in the course of a review and thus increases the ability to reproduce results or understand why results and conclusions might sometimes differ.

To enhance credibility. Systematic methods allow the reader to take the findings of a review on more than just trust of the authors, with a demonstration that a systematic review is not just "a story told by (knowledgeable) authors who present their personal views on their topic of interest in a more or less well disguised manner" supported by literature which is "largely what has been accumulated over time and shaped the opinion of the author(s)" (Hartung 2009).

To identify knowledge gaps. It would be a mistake to consider the purpose of a systematic review as to give a definitive answer to a research problem – after all, they can only tell you what is currently known. What they are very good for, however, is telling you what is not known – the areas in which research data are lacking, weak or inconclusive, thereby paving the way for targeted research.

Recommended reading: Mulrow, C. D. (1994): Rationale for systematic reviews. In BMJ 309 (6954), pp. 597–599.

2.1.2 The basic elements of a systematic review

Systematic reviews are scientific experiments – the difference being, instead of collecting observations from test-tubes and mass spectrometers, researchers are collecting observations from a dispersed body of evidence. There are two basic parts to a systematic review (adapted from Higgins, Green 2008):

2.1.2.1 A methodology, or protocol

The protocol for a systematic review should be published prior to conduct of the review itself, and incorporates:

- a clearly-stated set of objectives, describing what it is the research aims at finding out, why this is important, and what sort of evidence will be considered as being relevant to this aim;
- an explicit, reproducible methodology, covering all of the decisions the review team will make, from where they will look for evidence, to which evidence they will pay attention, how they will distinguish better evidence from worse, and how they will combine all this evidence into an answer to their research aim.

2.1.2.2 The review itself

Following the methodology laid out in the Protocol, a systematic review incorporates:

- a systematic search that attempts to identify all the evidence which is relevant to the aim of the review;
- an assessment of the validity of the findings of the included studies, covering whether or not there are random or systematic errors in the evidence-base, and the degree to which some of the evidence is more or less relevant to the aims of the review;
- a systematic synthesis and presentation of the characteristics and findings of the included studies, leading to a statement of what can be concluded from the evidence and how confident one can be in that conclusion.

2.1.3 A (very) brief history of systematic review

Acknowledgment of the need for reviews in medicine goes as far back as the 18th century, when the Scottish naval surgeon, early medical pioneer and scurvy researcher James Lind conducted not only the first recorded clinical trial but in order to "root out prejudices" prefaced it with a "full and impartial view of what had hitherto been published on the scurvy" – arguably the first recorded attempt at a systematic review (Lind 1753).

In 1987, Cynthia Mulrow revealed that none of the review articles published at the time in four major medical journals had used scientific methods for identifying, appraising or synthesising information.

In 1907, Joseph Goldberger became the first person to carry out something approaching the modern concept of the systematic review, in the course of evaluating the effectiveness of inoculation against enteric fever. Goldberger performed a full literature search with comprehensive references and exclusion criteria, and conducted a statistical analysis of pooled data from the included studies (Chalmers et al. 2002).

It was another 80 years, however, before Cynthia Mulrow, a pioneer of evidence-based medicine and the use of systematic review in healthcare, began the process of articulating for the health professions the scientific issues which need to be addressed in synthesising information, in an article published in 1987 in the Annals of Internal Medicine (Mulrow 1987).

This paper concluded that none of the review articles published at the time in four major medical journals had used scientific methods for identifying, appraising or synthesising information.

The need for improving the review process was considerably sharpened by a paper published in 1992 which found that clinicians' recommendations for treatment of heart attacks was correlated not with the treatments best supported by the available evidence, but instead on whichever analyses of evidence to which the clinicians happened to have access. This hardly amounted to evidence-based practice, leading to the paper's call for a specialised discipline of summarising evidence (Antman et al. 1992).

Systematic reviews have since become the most-cited publications in the medical literature (Patsopoulos et al. 2005) and conducting a systematic review is often a prerequisite of performing new clinical research (Young, Horton 2005). In excess of 700 textbooks and 25,000 journal articles offer perspectives on the basics of evidence-based medicine (Greenhalgh 2010).

The history of systematic review shows two things: firstly, that for all the simplicity and obviousness it has when viewed in hindsight, systematic review was difficult to invent and articulate de novo; secondly, that once articulated the techniques of systematic review had rapid and wide-ranging impact on the medical disciplines.

So our questions are: what are the basic elements of systematic review, and should we be thinking about adapting them for use in evaluating chemical safety?

Recommended reading: Chalmers, Iain; Hedges, Larry V.; Cooper, Harris (2002): A brief history of research synthesis. In Eval Health Prof 25 (1), pp. 12–37.

2.2 The anatomy of a systematic review

2.2.1 A clear objective

In excess of 700 textbooks and 25,000 journal articles offer perspectives on the basics of evidence-based medicine (Greenhalgh 2010)

Often, a systematic review addresses a question which could in theory be answered by a single practical experiment (European Food Safety Authority 2010) – the reason for conducting the review rather than doing the experiment being there might already be enough evidence to answer the question without doing the experiment at all. However, the objective of a review can also be to identify research trends and needs, evaluate the strength of evidence supporting a particular policy or intervention, or acquire any other sort of knowledge which a systematic appraisal of existing research would give you.

To prevent waste of limited resources, the objective should be set in such a way as to produce a review which directly answers a question of importance to those making decisions in medical care, be they policy-makers, practitioners or patients. The need for the review should be justified by the research and social context in which the review is being conducted.

2.2.1.1 Examples of objectives

Screening for breast cancer with mammography
"To assess the effect of screening for breast cancer with mammography on mortality and morbidity." (Gøtzsche, Nielsen 2011)

The authors explain that they conducted this review because although screening is widely practised due to the belief that it reduces mortality and morbidity from breast cancer, as yet there is no clear evidence as to how effective it is (if at all).

Vitamin C for preventing and treating the common cold
"To find out whether vitamin C reduces the incidence, the duration or severity of the common cold when used either as a continuous regular supplementation every day or as a therapy at the onset of cold symptoms." (Hemilä, Chalker 2013)

The authors explain that they conducted this review because vitamin C is cheap and easily accessible, while the common cold is a highly prevalent illness; if vitamin C really were to help with colds, it would be a major boon to public health.

2.2.2 A pre-published protocol

The Cochrane Collaboration requires all systematic reviews to be preceded by the pre-publication of a review protocol. The protocol is peer-reviewed to ensure robustness and describes in detail what the review team will do, from how they will search for evidence of possible relevance to the review objective, how they will decide which evidence will be included in the review, how they will assess the quality of the evidence, and how they will synthesise that evidence into a conclusion for the review.

The pre-published protocol is vital to minimising bias in a systematic review: registration of the protocols discourages non-publication of reviews with negative findings (as is the case for individual trials, reviews with "uninteresting" results are less likely to get published), while a pre-specified methodology discourages selective presentation of findings in a review (The PLoS Medicine Editors 2011; Liberati et al. 2009).

The pre-published protocol is vital to minimising bias in a systematic review: registration of the protocols discourages non-publication of reviews with negative findings, while a pre-specified methodology discourages selective presentation of findings in a review

Recommended reading: Liberati et al. (2009): The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. In PLoS Med 6 (7), pp. e1000100.

Recommended reading: US National Toxicology Program (2013): Draft Protocol For Systematic Review To Evaluate The Evidence For An Association Between Bisphenol A (Bpa) Exposure And Obesity.

2.2.2.1.1 What goes into a protocol

Derived from European Food Safety Authority (2010), Higgins, Green (2008).

Protocol Elements		Description
Background		Reasons for doing the review; theoretical underpinning of the review topic
Objective		Clear statement of the objective of the review, normally stated as a question to be answered; statement of the inclusion criteria for studies to be reviewed.
Methods	Search strategy	Explanation of how potentially relevant studies will be located, including a statement of search terms and the information sources which will be queried; process for managing references.
	Study selection	How studies yielded by the search strategy will be screened for meeting inclusion criteria, including who will be doing the screening and how disagreements over eligibility will be resolved.
	Data collection	Details of the data which will be retrieved from each study, presentation of data collection forms, number of reviewers collecting data, resolution of disagreements, how data which is missing from study reports will be dealt with (such as by contacting study authors).
	Assessment of methodological quality (risk of bias)	The method used for assessing methodological quality of each included study.
	Data synthesis	Description of the strategy for data synthesis, including the conditions under which meta-analysis might be conducted, how data will be synthesized in the event that meta-analysis is not possible, any sensitivity analysis to determine the effect of study design, methodological quality etc. on overall findings, investigation of publication bias and other analysis of the data which might be useful for fulfilling the objective of the review.

2.2.3 Systematic search

Searching for literature of relevance to the review question is a two-stage process, beginning with a general search for all the evidence which may be relevant to answering the question, before those studies deemed of specific relevance to the review question are selected for appraisal (are included) in the review.

2.2.3.1 Search

The evidence relevant to a review can be highly dispersed through physical and digital libraries and an unpublished grey literature of rejected papers, technical reports, internal and commissioned research, and so forth. The more of this which is examined, the more likely a systematic review will represent all of the evidence on a certain topic. Systematic reviews therefore require a comprehensive and reproducible search of a range of sources of information in order to identify as many relevant studies as possible.

For transparency, the search strategy used should be fully documented, covering the search terms used, databases queried, whether hand-searching and referencing-mining from individual study bibliographies were employed, and so forth. The results of the search method should be presented so the user of the review can judge how comprehensive the search was.

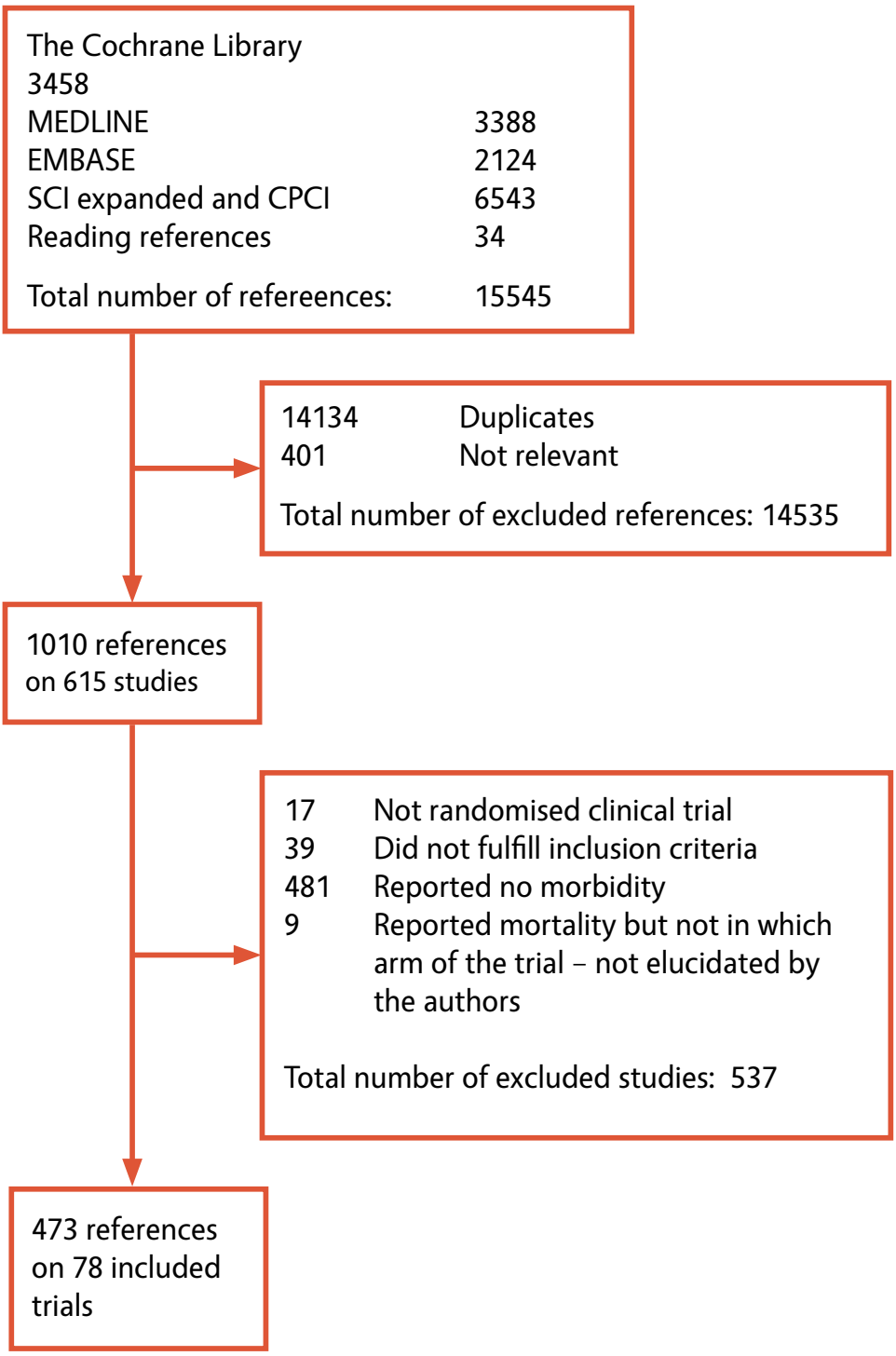
To ensure the selection process does not bias the results of the review, it is necessary to set in the protocol the criteria which each piece of evidence needs to meet in order to be included in the review

2.2.3.2 Selection

Not all of the studies found by the search strategy will actually contain information which helps answer the question in the review. These might be off-topic, or they might contain a methodological flaw so fundamental that they cannot be considered as evidence of anything (medical trials without controls, for example, are routinely excluded from Cochrane reviews). These will be excluded from the review.

To ensure the selection process does not bias the results of the review, it is necessary to set in the protocol the criteria which each piece of evidence needs to meet in order to be included in the review, and assess each piece of evidence found in the search strategy for whether or not it meets those criteria. These studies are then put forward to analysis in the review.

For transparency, the number of studies excluded should be stated, along with the principle reason for their exclusion. It is not necessary to discuss excluded studies at length – it is sufficient to present a brief note stating which criteria were used for excluding a study (one is usually enough).



A PRISMA flow diagram from a Cochrane Review showing the results of the search strategy (first box) then the process by which references not relevant to the review were excluded from the analysis. From Bjelakovic et al. (2012).

2.2.4 Assessment of the validity of included studies

The studies determined to be of relevance to answering a review question will vary both in terms of how well they have been conducted and how directly relevant they are to the issue at hand. Since weaker research of less direct relevance should count for less than stronger research of more direct relevance, it is necessary to systematically assess the quality and relevance of the included research.

In Cochrane Reviews, this process of assessing quality and relevance is deliberately described as the assessment of internal validity and external validity of a study. Use of the term "quality" is discouraged in relation to individual studies because it is considered ambiguous, so is instead reserved as a term for the overall quality of a whole body of evidence. The following discussion follows this convention (hence the title of this section).

Author's note: The following explanation of some of the basic principles of evaluating study quality is lengthy. This is partly because some of it is counter-intuitive, partly because much of it is all too easy to get wrong, and partly because a lot of it is just very interesting and fundamental to the science of information synthesis. Any errors in appraising the validity of a study can introduce errors into a review of data. In the same way that not using a sufficiently robust method in a laboratory introduces error into the results of an experiment, so does methodological error in the review method.

The studies determined to be of relevance to answering a review question will vary both in terms of how well they have been conducted and how directly relevant they are to the issue at hand.

2.2.4.1 Internal validity

The point of assessing the internal validity of a study is to address the credibility of a study: how confident are we that the method of the study has not introduced any systematic errors (bias) into the results?

2.2.4.1.1 How methodological quality of research is assessed in Cochrane Reviews

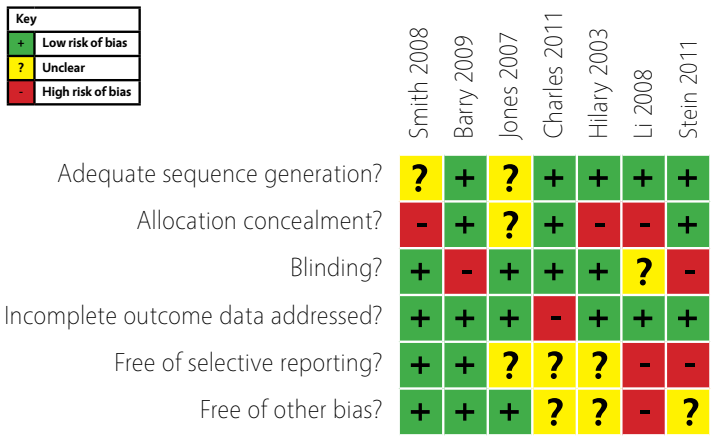
The Cochrane Collaboration has developed a specific toolkit for assessing the methodological quality of individual studies, concerned with the extent to which a study's method is at risk of introducing systematic error into the study results – the risk of it being biased. Once researchers have an understanding of the risk of bias of the individual studies in a review, they can adjust their confidence in their overall conclusions accordingly.

The Cochrane risk of bias assessment tool covers five key domains which need to be controlled if a medical trial is to be at low risk of being biased. These are: selection bias; performance bias; detection bias; attrition bias; and reporting bias; then there is one other category ("other") which is for biases which affect specific research methods.

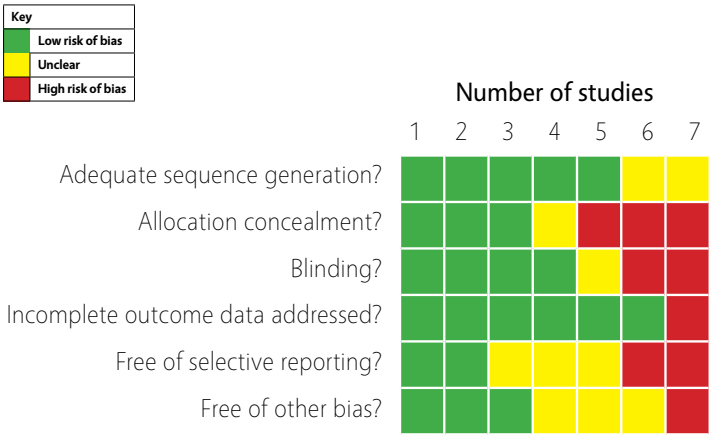
Depending on the methods used in a study, the reviewers rate the study as being at either low, unknown or high risk of bias for each of these domains, according to the decision-making procedures articulated in the review protocol.

This information is then summarized in a table and risk of bias chart. The table quickly and clearly shows the reasoning behind a risk of bias assessment for each study included in a review, while the charts give an easy visual representation of the credibility of each study in the review and the overall degree to which the whole body of evidence is at risk of bias.

At present, it does not appear to be the case that EU Expert Committees use a rigorously-defined methodology for appraising study quality.



Risk of bias charts offer a visual representation of the results of the risk of bias assessment, showing the user how well the each individual study performs against the criteria for internal validity of included research, and allows quick comparison between different studies in the review.



An additional way of presenting the data which can also assist users in interpreting the overall quality of a body of evidence is to show the proportions of data which are at various degrees of risk of bias. Note that each study occupies an equal amount of space on the scale rather than a proportion of space which is relative to the amount of data they contribute to the review, meaning a chart such as this should be interpreted with care.

Alison, Brian, Child et al. (2001)		
Domain	**Support for Judgement**	**Judgement**
Sequence generation	Quote: "A random number generator was used to assign patients to treatment and control groups."	Low
Allocation concealment	No mention of method for allocation concealment. Study authors uncontactable.	Unclear
Blinding of participants and personnel	Quote: "The placebo was prepared to be identical in appearance and taste to the tablets being taken by the test group." However, authors confirmed that neither researchers nor lab technicians were blinded.	High
Blinding of outcome assessors	Quote: "Outcome assessors were given samples in random order." Comment: Although blinding in this way could be broken, the consistency of results between assessors was checked and the result is not especially subjective.	Low
Attrition	Missing outcome data was balanced across groups. Reasons for missing data were similar across groups.	Low
Selective reporting	A primary outcome was reported using a data subset not pre-specified in the protocol.	High
Other	No measures were in place to prevent drug pooling.	High

This table shows how a study performs against the criteria being used in a review for evaluating the internal validity of included research. Presentation of data in tabular format makes it easy for users to determine whether or not each study in the review had been subject to the same test for internal validity and makes the basis of each judgment readily discernible.

2.2.4.1.2 Five key components of the Cochrane tool for assessing risk of bias

Many instruments for assessing research quality have been developed. In addition to the Cochrane Collaboration's emphasis on internal validity rather than quality of a study, a number of features of their approach to assessing the credibility of research deserve comment, in order to better understand why the Collaboration considers their approach to constitute best practice in systematic review.

2.2.4.1.2.1 The need to distinguish between bias and precision in the analysis

The distinction between systematic and random error, or bias and precision, in a study is important. A study is biased if, on repeat performance, it consistently overestimates or underestimates the true intervention effect. Lack of precision, on the other hand, is simply random error. If repeated enough times an imprecise study will zero in on the true size of the intervention effect but a biased study never will – so with an imprecise study, you can be sure that the true answer is within the range of answers the study gives, but with a biased study you cannot be so confident.

Because a very precise study can be biased while a very imprecise study can be unbiased, it does not follow that increased precision equates with increased study quality. Bias and precision therefore need to be distinguished in the analysis: bias needs to be dealt with first, because it gives a sense of how far away a study is from indicating the true size of effect; the precision of estimated size of effect can be worried about later, when the data from included studies is pooled together.

***Bias vs. Precision**. The green box shows the range of results given by an imprecise but unbiased study.Because the study is unbiased, you know that the true result of the study lies somewhere within the boundaries of the range of results.*

The red box represents the results given by a precise but biased study. In this case, you do not know where the true result of the study lies, whether it is inside or outside the range of results given by the study.

However, if you know which biases affect the red study, along with their likely magnitude and direction, you can make an educated guess as to how biased the red study is and therefore your confidence that even the biased study might be giving you an approximately correct answer.

This is why it is incorrect to hold that a precise study is better than an imprecise study without also determining the risk of bias in each study.

2.2.4.1.2.2 Rejection of the use of scores and scales for describing study validity

In assessing risk of bias, the temptation can be to give a scored estimate of risk of bias, to quantify the level of confidence one can have in a particular study and facilitate the comparison of methodological quality of one study with another. As appealing as this possibility is, the Cochrane Collaboration explicitly advises against the use of scales and checklists to give a scored estimate of risk of bias.

Because a very precise study can be biased while a very imprecise study can be unbiased, it does not follow that increased precision equates with increased study quality

This may seem surprising, given the widespread use of scores and checklists for study quality, but the reason is straightforward: there is no evidence that any scoring system yet devised accurately measures risk of bias. Empirical research into the use of scales has shown that, while a high degree of inter-rater consistency can be achieved in the application of scales, the judgments of quality fail to correlate with observed effect sizes: studies with lower quality scores should give estimates of effect further from the truth, but this turns out not to be the case (Jüni et al. 1999; Berlin, Rennie 1999; Glasziou et al. 2004).

This means all the effort going into scoring studies for quality is not helping identify the studies which get closest to reporting the true magnitude of effect.

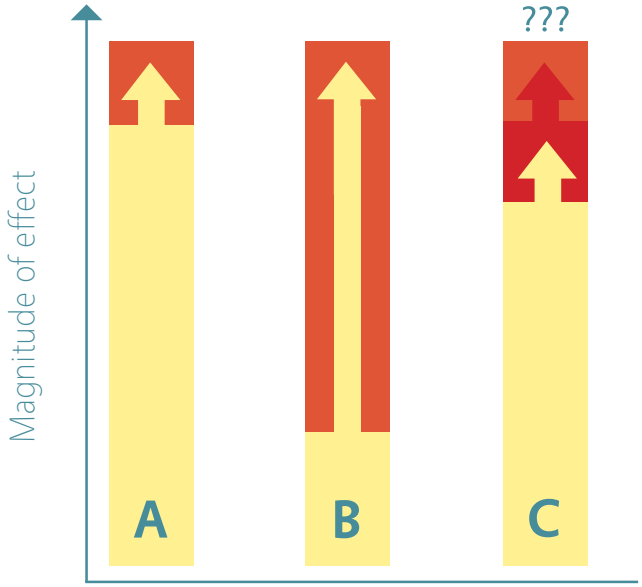
This is an issue for chemicals policy because although scores and scales such as the Klimisch criteria (Klimisch 1997), with 62 citations in the literature (Ågerstrand et al. 2011b), are in wide use in regulatory risk assessment and prominent in the European Chemicals Agency guidance on reporting weight-of-evidence evaluations (European Chemicals Agency 2010), nobody has ever proven that they work. Overall, 30 instruments exist for assessing the internal validity of animal studies but only two have been tested for validity or reliability (Krauth et al. 2013).

2.2.4.1.2.3 Focus on material risk of bias

The extent to which a study is at risk of bias varies from case to case, depending on the particular kind of bias and the outcome of interest. In one study bias may be certain yet compared to the size of intervention effect may introduce only a small error; in another study the risk of bias might be moderate yet lead to a significant error; in another a high risk of bias might lead to a moderate error but only to underestimating a beneficial effect.

The point is that each of these scenarios will have a different effect on the confidence one has in the data and the decisions one makes on the basis of it. For example, one might prescribe a drug for which the evidence is biased but only towards underestimating benefit, while not prescribing a drug which the evidence might be biased but benefit greatly overestimated. (In risk assessment, data which systematically underestimates harm might lead to difficulties in calculating a TDI but could still inform restrictive measures in risk management.)

This is why the Cochrane Collaboration emphasises the importance of material rather than hypothetical risk of bias when assessing the internal validity of a study, with size and direction of bias to be given due consideration in judging the credibility of a piece of research.



Why risk of bias must be material. *Risk of bias is the same concept as any kind of risk calculation: risk = likelihood of effect x magnitude of effect. Here, assuming that the likelihood of bias for each study is the same, we can see how the estimated magnitude of bias affects the material risk of bias in the study. For study A, the magnitude is small enough that it's a minimal material risk. For B, the magnitude of effect of bias is enormous and material risk therefore significant. For C, we are unsure of the magnitude of effect so the risk of bias in uncertain. None of these three cases should be treated equally.*

2.2.4.1.2.4 Understanding that poor reporting of a study is not the same as poor conduct of a study

Cochrane also remind us, when assessing a study for risk of bias, of the distinction between adequacy of the conduct of a study and adequacy of reporting. The quality of reporting obviously affects the ability of the reviewers to assess the risk of bias; however, quality of reporting is not directly related to risk of bias (Higgins et al. 2011).

Because studies which fail to report various methodological elements will sometimes have carried out these elements and sometimes not, it follows that on average a study which is not fully reported will be more reliable than a fully-reported but poorly-conducted study, while being less reliable than a study which is both fully-reported and well-conducted.

The quality of reporting obviously affects the ability of the reviewers to assess the risk of bias; however, quality of reporting is not directly related to risk of bias

Studies which are at unclear risk of bias should therefore be treated as being between low and high risk of bias. There is empirical evidence supporting this, where studies with inadequate allocation concealment were found to exaggerate the effect of an intervention by 41%, while studies at unclear risk of bias from allocation concealment exaggerated the effect of intervention by only 30% (Schulz et al. 1995).

Since being unclear about risk of bias is of course unsatisfactory, when study reports leave doubts about risk of bias Cochrane reviewers are instructed to follow up with researchers in order to clarify study methods.

2.2.4.1.2.5 Understanding that being at risk of bias is not necessarily a good reason for excluding a study from a review.

It may be tempting to only take into account data from the studies at the lowest overall risk of bias, on the argument that only taking data from the most credible studies must naturally produce the most credible results. Again, however, the Cochrane Collaboration advises caution in the face of intuitive appeal.

One concern with using risk of bias as an exclusion criterion is the trade-off between bias and precision: an analysis including all studies could be very precise because it uses the most data, but be seriously biased because of the flawed conduct of the studies which are included. Conversely, only using a very few of the least biased studies reduces the size of your data set and can lower precision – so you can end up with an unbiased but very imprecise measure of effect size.

A second concern is that excluding data from the analysis because of worries about its credibility may make too many assumptions on behalf of the users of that data as to what it is useful for and what it is not. Because people use reviews to inform decisions made in all kinds of different contexts, only some of which can be anticipated by the authors of the review, excluding data due to risk of bias potentially reduces the utility of the review.

Thirdly, differences in risk of bias between studies can also help explain why the studies included in a review might have obtained different results, which is why it is useful to account for risk of bias in studies rather than use risk of bias as an exclusion criterion.

This is not to say that all data should be treated as equally valid regardless of its credibility – that would be to misunderstand what is meant by a study being included in a systematic review. In this context, for data to be included only means that it be analysed, not necessarily that it be believed. So when we say that risk of bias should not be used as an exclusion criterion, what we are saying is that data at higher risk of bias should be included in the review, be accompanied by a description of its credibility and weighted accordingly, so that weak data is not overly influential in drawing conclusions but is nonetheless accounted for in the review.

To give a concrete example, imagine that a systematic review looked at the efficacy of various means of preventing transmission of malaria in sub-Saharan Africa. Now suppose that all the studies looking at the efficacy of mosquito nets were at high risk of bias. If the review used high risk of bias as an exclusion criterion it would end up saying nothing about research into malaria nets except that it had been excluded, when actually there might be a lot of information which can be extracted from those studies which would be useful to people involved in preventing morbidity and mortality from malaria – even if for example it only produces research recommendations which would improve the evidence base.

Instead of excluding studies at risk of bias, the Cochrane Collaboration advises conducting sensitivity analyses to measure the difference that including studies at higher and lower risk of bias makes to the size of the intervention effect

Instead of excluding studies at risk of bias, the Cochrane Collaboration therefore advises conducting sensitivity analyses to measure the difference that including studies at higher and lower risk of bias makes to the size of the intervention effect (The Cochrane Collaboration 2002).

Recommended reading: Higgins, J. P. T.; Altman, D. G.; Gotzsche, P. C.; Juni, P.; Moher, D.; Oxman, A. D. et al. (2011): The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. In BMJ 343 (oct18 2), pp. d5928.

2.2.4.2 External validity

External validity concerns the matter of whether or not a study is of the right kind to answer the question at hand. For example, a study which looks at the effect of vitamin C supplementation on the risk of cardiovascular disease has very little

external validity if the issue of concern is the effect of vitamin C supplementation on the duration of a cold.

External validity is not a binary matter. For example, studies which look at the effect of vitamin C supplementation on colds in the general population may, because of differences in exercise habits, allow few inferences to be drawn about effects of vitamin C on marathon runners yet allow some inferences to be made people who take regular light exercise.

Interpreting external validity is challenging and there is no straightforward formula which can be followed; as such, any decisions about applicability in informing conclusions should be transparently stated and justifications supplied.

Judgements of external validity in chemical risk assessment are even more complicated, where the relevance of a mouse model or in vitro assay for drawing inferences about effects on human health can be hotly contested.

2.2.5 Systematic presentation and synthesis of results

Systematic reviews need discussion sections to help users interpret the results of the review. Typically, discussion should consist of:

- a summary of results of data extracted from studies in the review, plus any meta-analyses;
- a statement of the completeness and external validity of the evidence;
- a statement of the overall credibility of the body of evidence;
- acknowledgement of potential biases in the review process;
- and acknowledgement of agreements and disagreements with other studies or reviews.

2.2.5.1 Summary of results

Tables summarising results help a review fulfil its objective of being easy to use. A standard format is used to achieve consistency and ease of use across reviews. Standard Cochrane "Summary of findings" tables include the following six elements (Higgins, Green 2008):

- A list of all important outcomes, both desirable and undesirable.
- A measure of the typical burden of these outcomes
- Absolute and relative magnitude of effect, as appropriate
- Numbers of participants and studies addressing these outcomes
- A rating of the overall quality of evidence for each outcome
- Space for comments

2.2.5.2 The completeness and validity of the evidence

Addressing the extent to which a review is relevant to the purpose to which it is being put is a shared job, with responsibilities for both the user of the review (because it is the user who has to interpret the review for their particular decision-making context) and the authors of the review (because it is the authors' job to help the user as much as practicably possible). One thing the authors have to do is be absolutely clear on the scope of the review, on the population, intervention and outcomes which they are addressing, and any gaps in knowledge due to absence of research.

2.2.5.3 Describing the overall quality of the evidence

An assessment of the overall quality of the body of evidence is essential to informing the user of the confidence they can have in the existing research (often, when evidence is weak, it is this statement of what is not known which is the most important finding of a review; it would be a mistake to think that systematic reviews necessarily provide answers when they often simply bring clarity to what in fact we do not yet know).

...often, when evidence is weak, it is this statement of what is not known which is the most important finding of a review; it would be a mistake to think that systematic reviews necessarily provide answers when they often simply bring clarity to what in fact we do not yet know

Cochrane Reviews use the Grading of Recommendations Assessment, Development and Evaluation Working Group (GRADE) approach to describing the overall quality of a body of evidence supporting an estimate of effect of an intervention (GRADE Working Group 2013; Atkins et al. 2004).

GRADE defines quality of evidence as "the extent to which one can be confident that an estimate of effect or association is close to the quantity of specific interest" (Higgins, Green 2008, pp. Kindle 9582) and is described as either "high", "moderate", "low" or "very low". High quality evidence is convincing, very low quality is unconvincing.

The grade given to a body of evidence is a matter of judgment but is made within a transparent structure, involving explicit consideration of within-study risk of bias (internal validity), directness of evidence (external validity), variance in results between studies (heterogeneity), precision of the estimates of effect, and risk of publication bias. The quality of the evidence is graded according to the extent which weaknesses in these areas undermine confidence in the estimate of size of effect.

Note that this is a scheme specifically developed for medical research so cannot be recommended for direct application to syntheses of toxicological research.

Recommended reading: Atkins, David; Best, Dana; Briss, Peter A.; Eccles, Martin; Falck-Ytter, Yngve; Flottorp, Signe et al. (2004): Grading quality of evidence and strength of recommendations. In BMJ 328 (7454), p. 1490.

2.2.5.4 Acknowledgement of potential biases in the review process

The authors of Cochrane reviews are encouraged to discuss any risk that the findings of their review might be biased.

For example, in a Cochrane review of exercise as a treatment for depression, the authors acknowledged that publication bias may have exaggerated any effect size they saw, that they changed their method slightly to reduce risk of bias resulting from some post-hoc decisions made in earlier versions of the review, but they were using data from study arms with the largest effect rather than the largest dose – a limitation they stated they will address in the next update to the review (Rimer et al. 2012).

For the efficacy of influenza vaccine, the risk of bias was considered by the review authors to be stark enough for inclusion in the Abstract (Jefferson et al. 2012):

> *"WARNING. This review includes 15 out of 36 trials funded by industry (four had no funding declaration). An earlier systematic review of 274 influenza vaccine studies published up to 2007 found industry funded studies were published in more prestigious journals and cited more than other studies independently from methodological quality and size. Studies funded from public sources were significantly less likely to report conclusions favorable to the vaccines. The review showed that reliable evidence on influenza vaccines is thin but there is evidence of widespread manipulation of conclusions and spurious notoriety of the studies. The content and conclusions of this review should be interpreted in light of this finding."*

2.2.5.5 Disagreements with other studies or reviews

This section of a Cochrane review allows the authors to put their findings in the context of other studies and reviews, to explain to readers the reasons for similarities and differences in their findings.

2.3 Some examples of systematic reviews

These reviews, of varying quality, are worth exploring to understand the sorts of questions which systematic reviews in medicine can address, and also how reviewers handle bodies of data of varying accessibility, quality and experimental type. Abstracts and plain-English summaries are freely available but the bodies of Cochrane Reviews are not open-access.

Showell, Marian G.; Brown, Julie; Clarke, Jane; Hart, Roger J. (2013): **Antioxidants for female subfertility.** In *Cochrane Database Syst Rev 8*, pp. CD007807.

Taylor, Fiona; Huffman, Mark D.; Macedo, Ana Filipa; Moore, Theresa Hm; Burke, Margaret; Davey Smith, George et al. (2013): **Statins for the primary prevention of cardiovascular disease**. In *Cochrane Database Syst Rev 1*, pp. CD004816.

Waters, Elizabeth; Silva-Sanigorski, Andrea de; Hall, Belinda J.; Brown, Tamara; Campbell, Karen J.; Gao, Yang et al. (2011): **Interventions for preventing obesity in children.** In *Cochrane Database Syst Rev* (12), pp. CD001871.

Rimer, Jane; Dwan, Kerry; Lawlor, Debbie A.; Greig, Carolyn A.; McMurdo, Marion; Morley, Wendy; Mead, Gillian E. (2012): **Exercise for depression.** *In Cochrane Database Syst Rev 7*, pp. CD004366.

Krogsbøll, Lasse T.; Jørgensen, Karsten Juhl; Grønhøj Larsen, Christian; Gøtzsche, Peter C. (2012): **General health checks in adults for reducing morbidity and mortality from disease**. In *Cochrane Database Syst Rev 10*, pp. CD009009.

2.4 Maintaining standards in systematic review

Part of the value in systematic reviews is having a recognizable gold standard by which they are conducted. This helps users trust them and encourages their uptake by the general medical community. This gold standard, however, needs maintaining, with continual development of research methods and ensuring that each review published is of sufficient standard. Maintaining this standard and level of trust is the function of the Cochrane Collaboration.

2.4.1 Keeping reviews clear of real and perceived bias

There is an acute awareness within the Cochrane Collaboration of the need to be demonstrably free of bias, whether it is real or perceived. This is necessary because Cochrane Reviews will only be used if they are trusted; any general slippage even in perception that the reviews are compromised by conflicts of interest and that trust could be lost.

The Cochrane Collaboration policy on conflicts of interest is comprehensive, with the guidance stating that it should cover anything which might be perceived by readers as capable of influencing an author's judgments (Higgins, Green 2008, pp. Kindle 2249). Measures to manage interests are as follows:

a) The extreme (usually financial) cases of conflicts of interest are barred from involvement in a review. Receipt of funding, hospitality or any other kind of subsidy from a source which may be perceived to have an interest in the outcome of the review is absolutely forbidden.

b) Other relevant interests such as personal conflicts, political, academic and other interests must be declared. Although it should be avoided if possible, authors must state if they have been involved in a study which is included in the review. (An interest is relevant if a user of the review might perceive it as having an influence over the review's results.)

c) The stipulation that reviews are carried out according to pre-published protocols by teams whose members have differing interest profiles. This is on the understanding that when interests impinge on decision-making, disagreements will then result. These can then be documented and the adequacy of their resolution judged by the reader.

d) The requirement to disclose contributions to a review.

2.4.1.1 Describing authors' contributions to a Cochrane Review

Cochrane Reviews include a dedicated section for information about the authors, a contact person, acknowledgements, declarations of interest and a detailed breakdown of the activities of whoever contributed to the review. This is to help make sure due credit is given for work done, and also ensure accountability and transparency.

Part of the value in systematic reviews is having a recognizable gold standard by which they are conducted.

Cochrane guidance on describing contributions to a systematic review (Higgins, Green 2008)

- Conceiving the review.
- Designing the review.
- Coordinating the review.
- Data collection for the review.
  - Designing search strategies.
  - Undertaking searches.
  - Screening search results.
  - Organizing retrieval of papers.
  - Screening retrieved papers against eligibility criteria.
  - Appraising quality of papers.
  - Extracting data from papers.
  - Writing to authors of papers for additional information.
  - Providing additional data about papers.
  - Obtaining and screening data on unpublished studies.
- Data management for the review.
  - Entering data into RevMan.
- Analysis of data.
- Interpretation of data.
  - Providing a methodological perspective.
  - Providing a clinical perspective.
  - Providing a policy perspective.
  - Providing a consumer perspective.
- Writing the review (or protocol).
- Providing general advice on the review.
- Securing funding for the review.
- Performing previous work that was the foundation of the current review.

2.4.2 Assuring the on-going production, relevance, utility and quality of reviews

One of the notable things about the Cochrane Collaboration is how much it achieves with so few resources. Key to success is how the majority of authors contribute their time to the production of reviews free of charge: essentially, the Cochrane Collaboration solves the problem of resourcing the production of large numbers of high-quality systematic reviews by outsourcing it to volunteers.

This approach functions because volunteering academics view authoring reviews as part of their existing efforts to keep up-to-date in their areas of interest. Since Cochrane Reviews are particularly prestigious and well-cited authors get a deal of credit for writing them.

The organizational infrastructure behind the voluntary production of systematic reviews is of critical importance to the whole systematic review process. Authors would not be able to contribute reviews without the existence of an organizational structure which set the standards for reviews, promoted their use, conducted research into review methods, ensured reviews are accessible and well-read, and ensured the correct topics are being covered. The Cochrane Collaboration is therefore made up from the following units (The Cochrane Collaboration 2013):

...volunteering academics view authoring reviews as part of their existing efforts to keep up-to-date in their areas of interest.

Cochrane Review Groups (CRGs): Responsible for the preparation and maintenance of Cochrane Reviews, a CRG consists of a Coordinating Editor and editorial team which plans, coordinates and monitors the CRG's work. A Managing Editor is appointed to organize the day-to-day activities of the CRG. Each CRG is supported by people working in Methods Groups, Fields and Centres.

Besides determining the standards and procedures for reviews in their particular area, the CRG reduces the burden placed on individual authors by providing support for conducting systematic searches for relevant studies, supplying studies to authors and ensuring that authors receive the methodological support they need.

Methods Groups: Because the science of research synthesis is still young and evolving fast, Methods Groups have been established to advise the Cochrane Collaboration on how the validity and precision of systematic reviews can be improved.

Fields: Fields ensure that priorities and perspectives in their particular areas of interest are reflected in the work of CRGs, to ensure the output of the CRGs is maximally relevant to the various healthcare disciplines. The Cochrane Consumer Network provides information for consumers and is a liaison point for consumer groups.

Centres: Centres facilitate the work of CRGs, Methods and Fields with activities such as training, and are responsible for promoting the objectives of the Cochrane Collaboration at the national level.

Steering Group: The Steering Group is the board of trustees of the Cochrane Collaboration, elected by registered members of the CRGs, Methods Groups, Fields, Consumer Network and Centres. The Steering Group governs the Collaboration by making decisions in line with the goals set out in the Collaboration's Strategic Plan.

Cochrane Operations Unit: A small staff based in Oxford UK, providing support to the Steering Group.

The Cochrane Library: The main output of the Cochrane Collaboration, the library contains the full Cochrane Database of Systematic Reviews. The library is updated quarterly and distributed via the Internet and CD-ROM. The Library has its own Editor in Chief and Editorial Unit. In addition to systematic reviews, the Library contains a number of other databases including the Methodology Register (a set of references to literature on the science of reviewing research).

2.5 Conclusions

This brings our tour of systematic review, as realized by the Cochrane Collaboration, to a close. Hopefully this has given some insight into the rationale, the techniques and the organizational structures which underpin the production of systematic reviews:

- Traditional review methods were found to be inadequate for transmitting into practice the medical decisions best supported by the available evidence because they were insufficiently scientific in their approach to finding and analysing relevant data.
- It is believed that applying the scientific principle of reproducibility of method to the process of synthesizing data should at least ensure everyone has theoretical access to the full set of evidence which should inform practice, even if it won't change practice on its own.
- Systematic review techniques are a complex constellation of methods which cover the determination of review objectives, definition of and adherence to protocols, finding of data of potential relevance to the review, selection of studies for analysis, appraisal of the validity of included studies, synthesis of data and presentation of results.
- The production of systematic reviews is enabled by an organization which develops and protects a standard for the conduct of reviews and ensures that each review has maximum value to its users.

Part Three.

Comparing current review practices at EFSA with standards expected of systematic reviews in medicine

"Unfortunately, medical reviews are often subjective, scientifically unsound, and inefficient. Strategies for identifying and selecting information are rarely defined. Collected information is reviewed haphazardly with little attention to systematic assessment of quality. Under such circumstances, cogent summarization is an arduous, if not insurmountable, task." (Mulrow 1987)

3 Comparing current review practices at EFSA with standards expected of systematic reviews in medicine

3.1 Introduction

What follows is a comparison between two formal reviews of evidence by the European Food Safety Authority and the standards required of Cochrane reviews, in order to crystallise a set of general recommendations for reform to the processes by which EU Agencies review evidence for the purpose of chemical risk assessment.

The review of interest are the 2010 Scientific Opinion on BPA (EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids 2010) and the 2013 Draft Opinion on exposure to BPA (EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids 2013).

To develop the comparison, the author of this report analysed the methodological quality of the Opinions according to eight key components of systematic review. Seven are key to systematic reviews in medicine, as articulated in the Cochrane Handbook for Systematic Review (Higgins, Green 2008) and related literature, with the eighth additional element (assessment of external validity of evidence) of specific relevance to chemical risk assessment.

These elements are:

1. The clarity of question for review
2. The use of a pre-published protocol
3. A comprehensive declaration of interests
4. A systematic search method for capturing all evidence of potential relevance to the review aims
5. The selection process for putting forward all research of actual relevance to analysis
6. The assessment of the external validity of included studies
7. The assessment of the internal validity of the included studies
8. The clarity of the answer to the question for review

The purpose of the analysis is to determine whether or not the documents developed by EFSA are sufficiently transparent and robust to be capable of adjudicating in a matter of scientific dispute – in this case, the toxicity of BPA. The analysis is not concerned with the validity of the findings of either of the two Scientific Opinions, instead developing a structured presentation of a user's subjective concerns about their methodological robustness of the review process.

This exercise will hopefully pull apart for the user the various things a review should be doing in order to secure confidence in the validity of its findings – that is, attaining transparency and reproducibility in each of the eight areas under scrutiny. This exercise also responds to the political dimensions of the debate about BPA, presenting a view

of what an agency would need to do if it were to develop documents of sufficiently demonstrable impartiality and scientific robustness to be capable of cutting through controversy in risk assessment and securing the trust of their users.

What would undoubtedly have strengthened this section of the report would have been other case studies in addition to the two Opinions by EFSA. There were plans to include ANSES' risk assessment of BPA for comparison with EFSA's Opinions, while with hindsight including reviews from other sources, such as NGO reviews of the toxicity of BPA and some reviews from the academic literature would have been beneficial, allowing broader lessons to be learned both about using the toolkit and the general landscape when it comes to review methods used by different stakeholders in chemical regulation. Resources did not extend to this.

Note on citations. When discussing research cited in EFSA's Opinions, the citation is prefixed with "[EFSA]". These citations should be sought from the bibliographies and literature search results of the relevant Opinions, not the bibliography of this report.

3.2 EFSA 2010 Opinion on BPA

EFSA Panel on food contact materials, enzymes, flavourings and processing aids (CEF). Scientific Opinion on Bisphenol A: evaluation of a study investigating its neurodevelopmental toxicity, review of recent scientific literature on its toxicity and advice on the Danish risk assessment of Bisphenol A. EFSA Journal 2010;8(9):1829. [110pp.] doi:10.2903/j.efsa.2010.1829.

3.2 .1 Clarity of question for review

Reviews should ask a clear, unambiguous question, the formulation and usefulness of which is justified by a presentation of the context in which the review is being conducted.

The background to the Opinion is clearly presented - in light of concerns about low-dose effects, EFSA was asked to update its 2006 Opinion on BPA, incorporating new research which had since been conducted. The primary focus of the Opinion was to determine if any new data had been published which would lead to a change in the TDI for BPA.

The objective of the Opinion is less clear. The terms of reference from the Commission were: "Update, if necessary, the currently applicable tolerable daily intake for Bisphenol A" (p10); however, there is no clear statement in the Opinion as to how EFSA interpreted the remit, except "EFSA should address any new scientific evidence that may affect the conclusions of the previously adopted opinions on BPA" and "the CEF Panel undertook the task of reviewing new toxicological data that may have an impact on the previous risk assessments of BPA".

This is inadequate because "impact" is undefined, allowing for potentially conflicting interpretations of the objective of the review. This undermines reproducibility – and indeed, two interpretations appear to be presented in the Opinion.

The main body of the Opinion appears to work from a narrow interpretation of "may have an impact". Here the Panel seems to have considered its objective to be to recalculate the tolerable daily intake (TDI) for BPA, so that if a study potentially undermined the TDI but did not produce data which directly permitted recalculation of the TDI, then the study would be considered to have no impact on the risk assessment of BPA.

The Minority Opinion, on the other hand, appears to operate a broad interpretation of "may have an impact". Here, the objective of the Opinion is understood as not necessarily being to recalculate a TDI but to evaluate whether or not any new evidence has implications for the accuracy of the TDI, regardless of whether that evidence allows a new TDI to be calculated.

This ambiguity seems to drive the differences in conclusions between the conclusions of the main body of the Opinion and the Minority Opinion, where the former states there is no grounds for changing the TDI, while the latter says while a new TDI indeed cannot be calculated, the TDI as it stands could well be incorrect.

A clear statement of whether the interpretation of "may have an impact" is narrow or broad is therefore needed. If a narrow interpretation of remit is preferred, a convincing explanation of why this is preferable to a broad interpretation will need to be provided. This is because it seems intuitive that studies which do not themselves permit calculation of a TDI may undermine a TDI anyway (by, for example, calling into doubt the methods of the studies which generate the TDI), while a broad interpretation will give the most informative response to the terms of reference of the Opinion.

3.2.1.1 Conclusion

The objective is not consistent with a scientific standard for reviewing evidence because it is ambiguous, while insufficient justification is given in the text for the Panel's choice of operating a narrow interpretation of the terms of reference for the Opinion.

3.2.2 Use of a pre-published protocol

Cochrane reviews require Review Teams to develop and follow robust review protocols which are published prior to conducting a review, in order to avoid bias from subjective and ad-hoc decision-making in the conduct of the review.

There was no pre-published protocol for the review. Instead, an expert Working Group was convened to conduct unspecified preparatory work for the Panel, while the Panel was responsible for writing of the Opinion.

3.2.2 .1 Conclusion

The lack of pre-published protocol is not consistent with a scientific standard for reviewing evidence.

3.2.3 Comprehensive declaration of interests

Cochrane declarations of interest are comprehensive, covering not only financial and professional interests but relevant publishing history and the specific contributions made by each contributor and member of the Review Team. This is to ensure conflicting interests do not distort the results of the review and allow the reader to put the conclusions of the reviewers into their full context.

3.2.3.1 Information about interests

The members of the Panel and BPA Working Group are listed in the Opinion. Organisational affiliations and declarations are not given in the Opinion but are instead available through the EFSA database of declarations of interest of active experts. For experts no longer serving with EFSA, declarations are available by email on request.

These declarations are not specific to the Opinion but instead the user needs to read lengthy documents and then interpret for themselves any potential conflicts, which for all Panel and Working Group members is a very lengthy task. The specific contribution made by each Panel and Working Group member to the development of the Opinion is not stated.

3.2.3.2 Conclusion

The declaration of interests is not consistent with a scientific standard for reviewing evidence. It is insufficiently complete, while for given information there is the practical difficulty for the user to construct an image of the interests of each member of the Panel and Working Group. Overall, the information presented is insufficient for reliably developing a clear and accurate picture of how those interests might have shaped the findings of the Opinion.

3.2.4 Systematic search method for capturing all evidence of potential relevance to the review

In order to make sure that the evidence surveyed in a review is representative of all of the available evidence (i.e. that sampling bias is avoided), a systematic search strategy should capture all research of potential relevance to the objective of the review, and should be reported in such a way as to be reproducible by a third party.

The Panel's search method was available on request rather than being presented as part of the Opinion. The Panel searched one on-line database, ISI Web of Knowledge v4.9 with the keywords "Topic= (bisphenol a)", refined by: Publication Years=(2009 OR 2008 OR 2007 OR 2010) AND Subject Areas=(TOXICOLOGY) AND Document Type=(ARTICLE).

This yielded 839 results of which 55 were discarded as duplicates and 5 for lacking an author's name, leaving a final list of 779 references. These are listed as part of the Panel's search methods (EFSA Working Group for BPA Opinion), also available on request.

Performing a simple search in a different database, PubMed, (search term: "bisphenol a"; filters "humans" and "other animals") for the same date range produces 1232 results. A number of these are dental studies with no toxicological or exposure data and unlikely to be of relevance; however, there are several studies in addition to those reported by the Panel which appear as if they might be relevant to EFSA's review, for example:

In utero exposure to diethylstilbestrol (DES) or bisphenol-A (BPA) increases EZH2 expression in the mammary gland: an epigenetic mechanism linking endocrine disruptors to breast cancer. Doherty LF, Bromer JG, Zhou Y, Aldad TS, Taylor HS. Horm Cancer. 2010 Jun;1(3):146-55. doi: 10.1007/s12672-010-0015-9. PMID: 21761357

Assessment of bisphenol A exposure in Korean pregnant women by physiologically based pharmacokinetic modeling. Shin BS, Hwang SW, Bulitta JB, Lee JB, Yang SD, Park JS, Kwon MC, Kim do J, Yoon HS, Yoo SD. J Toxicol Environ Health A. 2010;73(21-22):1586-98. doi: 10.1080/15287394.2010.511584. PMID: 20954083

Evidence to suggest glutamic acid involvement in Bisphenol A effect at the hypothalamic level in prepubertal male rats. Cardoso N, Pandolfi M, Ponzo O, Carbone S, Szwarcfarb B, Scacchi P, Reynoso R. Neuro Endocrinol Lett. 2010;31(4):512-6. PMID: 20802452

Estrogenicity of bisphenol a: a concentration-effect relationship on luteinizing hormone secretion in a sensitive model of prepubertal lamb. Collet SH, Picard-Hagen N, Viguié C, Lacroix MZ, Toutain PL, Gayrard V. Toxicol Sci. 2010 Sep;117(1):54-62. doi: 10.1093/toxsci/kfq186. Epub 2010 Jun 20. PMID: 20566471

Endocrine disrupting chemicals bind to a novel receptor, microtubule-associated protein 2, and positively and negatively regulate dendritic outgrowth in hippocampal neurons. Matsunaga H, Mizota K, Uchida H, Uchida T, Ueda H. J Neurochem. 2010 Sep 1;114(5):1333-43. doi: 10.1111/j.1471-4159.2010.06847.x. Epub 2010 Jun 7. PMID: 20534002

Each of these studies was published shortly before the date of EFSA's search and as such may not have propagated into Web of Knowledge. Given the large number of studies returned by the search strategy, it is difficult to know how many of the PubMed citations in total should have been retrieved by the Panel, nor what effect they might have had on the results of the review. It is therefore difficult to have full confidence in the search strategy, though equally it does not seem to be so weak as to greatly undermine confidence.

3.2.4.1 Conclusion

The search strategy seems consistent with a scientific standard for reviewing evidence. That it failed to retrieve some eligible studies could be of concern in relation to sampling bias, though in this instance appears unlikely to have materially affected the findings of the review. Reporting of the results of the search process could have been clearer.

3.2.5 Selection process for putting forward to analysis all relevant research from the citations yielded by the search strategy

In order to reduce the risk of selection bias in the use of relevant information yielded by the search process, clear inclusion and exclusion criteria should be stated for selecting from the results of the literature search the specific references relevant for answering the review Question. All data from each study included in the review which is relevant to the review objective should be included in the review.

3.2.5.1 The Panel's inclusion criteria for studies

The Panel's inclusion criteria were: "Full research papers published in peer-reviewed journals available in public domains since the EFSA 2006 opinion (2007 – July 2010)." They state that they included only "original data (no reviews, discussions or others)" and "human studies", excluding "pure biomonitoring studies".

Then: "For the animal toxicity studies the focus was on studies having the following experimental design:

- Developmental exposure, i.e. pre-, peri-, and/or early post-natal exposure
- oral route of exposure
- several tested doses (plus a control) including at least one dose level lower than the NOAEL of 5 mg/kg b.w./day."

From the final list of 779 references retrieved by the search method, 183 studies are cited in the bibliography of the relevant part of the Opinion (Part II).

There is no summary of the selection process for including evidence in the Opinion, so it is not clear which studies were considered by the Panel to meet the inclusion criteria, nor is there a transparent explanation for the user as to why some studies were excluded, leaving the possibility that some relevant studies were not included. These concerns are magnified by evidence of inconsistent use of the inclusion criteria and the selective use of papers from search results.

3.2.5.1.1 Evidence of inconsistent use of inclusion criteria

Some papers are cited even though they do not meet the inclusion criteria for the Opinion. These include but are not limited to: [EFSA] Sharpe (2010), cited as "evidence showing that Sprague-Dawley rats are as responsive to oestrogens as the other rat strains" when it is an opinion piece containing no original data; [EFSA] Diel et al. 2004 as showing the same when it was published outside the dates specified in the inclusion criteria; [EFSA] Alonso Magdalena et al. (2010) cited as evidence of "aggravated insulin resistance" but is a review containing no original data; and two further reviews containing no original data, [EFSA] Anderson (2005) and [EFSA] Hodge & Tracy (2007) are cited to support the statement "It is important to underline that during pregnancy in humans the glucuronidation pathway is induced as compared to the activity in non-pregnant women".

Some inclusion criteria appear to be applied inconsistently. A toxicity study in primates exposed subcutaneously to BPA, [EFSA] Leranth et al. (2008), is discussed in the Opinion but two other subcutaneous studies listed in the search results, [EFSA] Newbold et al. (2007) and [EFSA] Newbold et al. (2009), are not discussed in the Opinion. Oral route of exposure is not therefore consistently being used as an exclusion criterion.

3.2.5.1.2 Evidence of selective use of papers from search results

The Panel describes [EFSA] Mok-lin et al. (2010) as of "no relevance" to risk assessment, where the Panel "noted that there is not supporting evidence from animal studies on the biological plausibility of the relationship between BPA low-exposure and female fertility (e.g. Tyl et al., 2002, 2008, Ryan et al., 2010a)." The choice of the two Tyl and one Ryan papers as evidence for the claim appears

selective, since EFSA's own bibliography and list of retrieved papers include studies with at least superficial relevance to assessing the effect of BPA on female fertility. These include [EFSA] Adewale et al. (2009) and Fernandez et al. (2009) which are in the bibliography of the Opinion but not mentioned in this context; and also [EFSA] Newbold et al. (2007) and Newbold et al. (2009) which are in the list of retrieved studies but are not mentioned in the Opinion.

It is possible these are not in fact relevant – however, for the Panel to transparently conclude "there is not supporting evidence", an explanation is required.

3.2.5.2 Conclusion

The selection process used in the Opinion is not consistent with a scientific standard for reviewing evidence. There is evidence that studies which are apparently relevant to the Opinion and were retrieved by the search process were not included in the Opinion. At the same time, studies which did not meet the inclusion criteria were nonetheless included in the Opinion.

Although the Panel states that "several research studies not compliant with the inclusion criteria, but still useful for hazard identification and for support of biological plausibility e.g. in vitro studies or non-oral in vivo studies are also discussed in this opinion", this does not make the inclusion process more transparent nor less selective, and does not therefore reduce risk of selection bias in the review.

3.2.6 Assessment of the external validity of included studies

Judgments of external validity are concerned with the relevance of a study for answering a review question: can an observation in one study group, such as an epidemiological cohort or group of mice in an animal study, be taken as representative of effects in the general population?

3.2.6.1 Criteria used by the Panel for assessing the external validity of included studies

There are no explicit criteria in the Opinion for appraising the external validity of studies. Instead, criteria have to be inferred from discussion in the body text, at points at which study quality is described. In general, the external validity and internal validity of studies are run together to produce a judgment of the suitability of a study for risk assessment. This makes appraisal of the use of the criteria challenging.

3.2.6.2 Conclusion

That internal and external validity are not distinguished by the Panel in the discussion of study quality makes it difficult to discern if a study is being downgraded because it is not methodologically robust or because it is of limited direct relevance to the evaluation of the toxicity of BPA in humans.

That there is no clear scheme for weighting studies according to their external validity does not really help the user evaluate the Panel's conclusions; as such, the consistency of application of the criteria for external validity are not evaluable. This is something which can be improved upon in future.

3.2.7 Assessment of the internal validity of the included studies

Internal validity is concerned with the credibility of a study, i.e. the risk of it being wrong either through random error (such as being statistically under-powered) or systematic error (bias). The criteria for judging the internal validity of animal and epidemiological studies are complex and therefore require careful planning and statement prior to conduct of the review, and care should be taken in ensuring each study is subjected to a fair test for internal validity.

These criteria should at least identify a material rather than hypothetical risk of bias or error, appraise the magnitude and direction of the risk of bias, and not treat these as equivalent limitations in the reporting and conduct of a study. The rationale for evaluating the internal validity of a study should be transparent, justified and consistently applied to all studies.

3.2.7.1 Criteria used by the Panel for assessing the internal validity of included studies

There is no detailed method specified for appraising the internal validity of the studies included in the Opinion. Instead, criteria have to be inferred from discussion in the body of the Opinion, at points at which study quality is described. The following are some examples of criteria apparently used by the Panel in appraising the credibility of studies included in the Opinion.

3.2.7.1.1 The presence of knowledge gaps as counting against study findings

The Panel appears to use the presence of knowledge gaps as a criterion for downgrading the credibility of study findings. For example, the Panel describes a lack of clarity about "the persistence of the BPA-mediated inhibition of

oestrogen-induced synaptogenesis" in [EFSA] Leranth et al. (2008). Absence of information about "the underlying mechanism of action" (p54) and a data gap left by "Effects being observed in humans at 2 years of age but not yet at 5 years of age" (p54) both count against the findings of [EFSA] Braun et al. (2009). The "uncertain clinical significance" of observed associations (p58) counts against [EFSA] Mendiola et al. (2010).

The first concern with this criterion is that it is always possible to identify research which has not been done, but which if it had would either reinforce or undermine an existing piece of research study. Given this criterion has wide scope for selective application and that some data gaps are going to be more important than others, a rationale for the application of this criterion should be articulated in order to ensure that it is being applied fairly to all studies included in a review. Although there is some evidence of a rationale, such as an epidemiological study not being admissible to risk assessment unless it is backed up with mechanistic evidence, there is no clear statement of when a data gap should count against the findings of a study and when it should not. This should be more clearly articulated.

The second concern is the criterion may blur the distinction between risk assessment and risk management. Correctly differentiating between study weaknesses and data gaps is an important matter because identifying a data gap is the job of risk assessors, while deciding what to do in the face of a data gap is the job of risk managers. Classifying a data gap as a study limitation seems to run the distinction together because it treats an issue which needs to be managed (the absence of data about risk) as if it is an issue which informs the level of risk itself (i.e. as if it is data about the absence of risk). This conflates risk assessment and risk management, which EFSA Opinions are supposed to avoid.

3.2.7.1.2 Possible confounding

Spot-sampling. The Panel observes that assessment of BPA exposure by single urine sample is not ideal, due to likely exposure measurement error from high within-individual temporal variability of BPA levels (p56). This confounder is described as a limitation undermining the value of [EFSA] Melzer et al. (2010) for risk assessment.

However, Melzer et al. state that this limitation would bias findings towards the null hypothesis: "The BPA measures in NHANES are based on single spot specimens, so misclassification from this single snapshot of body burden will have resulted in a smaller (diluted) estimate of the strength of association between BPA and the conditions of interest." Contrary to the Panel's analysis, this makes the finding more likely to be real rather than less, and serves as an example of confusion as to how risk of bias should affect the interpretation of the results of a study.

Genetic defect. With regarding to [EFSA] Salian et al. (2009), the Panel states (p65): "Only 8 dams per group were used and it may be plausible for a genetic defect in one pair to develop and magnify as subsequent generations are examined." While this is no doubt theoretically possible, there should be a statement of likelihood that this actually happened in order to evaluate whether or not this constitutes a material rather than hypothetical risk of bias.

Background contamination. Contamination by BPA is identified by the Panel as a potential confounder (p41). With regard to this, the Panel cites [EFSA] Doerge et al. (2010b) as showing that studies "reporting high levels of free BPA in serum, up to 20 μg/L, are very likely affected by sample contamination". While background contamination obviously needs to be controlled, it is not clear why a single study constitutes a definitive demonstration that high BPA measurements in another study are an error, rather than (for example) that the studies showing high levels of BPA means Doerge et al. might be under-reporting exposure. In this case, following up with the researchers seems appropriate.

Unmeasured confounding. In discussion of [EFSA] Braun et al. 2010, the Panel states that "unmeasured confounding, e.g. not adequately assessed parental psychopathology, alcohol or drug consumption, maternal behaviour toward the child, etc." may be confounding the findings of the study. While each of these factors may have confounded the study, some are more plausible than others (is it really credible that drug consumption would have differed enough between the two cohorts to significantly distort the results?), they are all hypothetical biases, and some are more hypothetical than others.

A similar point can be made for the discussion of [EFSA] Melzer et al. (2010), the Panel is concerned that confounding may arise from the fact that "Health outcome definitions were (5) based on self-reporting, including diabetes diagnosis, for which no laboratory test confirmation was available." Again, it is possible that this is the case – but is it really likely that people in the NHANES cohort were systematically mistaken or dishonest about their diagnoses?

3.2.7.1.3 Inadequate statistical power

Sufficient sample size is a criterion used by the Panel for judging the validity of studies. It is an explicitly-stated criterion used by the Panel for judging the validity of animal studies (p61), with the Panel describing a weakness of [EFSA] Yan et al. (2008) as using "small experimental groups of 3-4 animals and evaluations in males only" (p69). Sample size is also a concern for the Panel with [EFSA] Mendiola et al. (2010), an epidemiological study with "small" participation rates.

Although adequate sample size and statistical power is an essential feature of a study, the Panel never actually specifies what constitutes adequate size, merely

referring to studies in qualitative terms such as "small". Because statistical power is relative to study design (for example, pronounced effects do not need large cohorts in order to be detected) it is insufficient to simply dismiss a study as being in some sense "too small" without demonstrating that the study lacks sufficient statistical power to distinguish true effects from chance effects.

3.2.7.1.4 Incomplete reporting

Incomplete reporting is raised as an issue in several places. In discussion of [EFSA] Okabayashi and Watanabe (2010), the Panel states that "detail provided in the paper was limited". In discussion of [EFSA] Betancourt et al. (2010b), the Panel expresses concern that time of necropsy of individual animals was not exactly reported but given as "at 12 months of age or when tumour burden exceeded 10% of body weight". Of [EFSA] Yan et al. (2008) the Panel states: "due to the limited study reporting, no clear conclusions can be drawn from this study".

This treats inadequacies in reporting as if they are equivalent to inadequacies in conduct, which is an error. There is no evidence of attempts to contact study authors to resolve any outstanding questions left by incomplete reporting.

3.2.7.2 Confounders not taken into account by the Panel

The Panel fails to mention a number of important criteria for assessing the internal validity of studies. These include blinding of personnel and outcome evaluators and random allocation of animals to intervention groups. There is reason to believe these are very common, with at best one third of animal studies conducted for medical research implementing basic protections against bias, fewer than 1% reporting sample size calculations, and fewer still defining primary outcomes in advance (Macleod 2011). By not assessing these, the Panel could be overstating the internal validity of the included studies.

3.2.7.3 Conclusion

The assessment of the internal validity of studies included in the Opinion is not consistent with using a scientific method for reviewing evidence. The criteria for appraising internal validity are not fully and explicitly stated, so it is not possible to determine if the studies included in the review have been subjected to a fair test for credibility. There appear to be mistakes in applying the criteria which could lead to over- or underestimation of the internal validity of included studies.

3.2.8 Clarity of answer to the question for review

The answer to a review should be unambiguous and reflect what is stated by the research which has been evaluated and the level of confidence which can be had in those findings.

The answer given by the Panel is clear: that the TDI for BPA is still based on a multi-generation reproductive toxicity study in rats, because none of the studies reviewed by the Panel allow for a recalculation of the TDI. Although there are some studies which suggest the TDI might need to be revised, none of these actually permit such a revision to be made.

Any ambiguity in this answer, or the difference between the main Opinion and the Minority Opinion, likely derives from the ambiguity in the terms of reference of the Opinion.

3.2.8.1 Conclusion

The clarity of the answer to the question is consistent with a scientific standard for reviewing evidence, ambiguity in the question notwithstanding.

3.2.9 Overall conclusions about EFSA's 2010 Opinion on BPA

There are substantial differences between what is required of systematic reviews as conducted in evidence-based medicine and how evidence is appraised in the 2010 Opinion on BPA:

- The question which the Panel is answering is not clearly and unambiguously specified, so it is not clear if the answer is as useful to potential users as it might be.
- The Panel does not follow a pre-defined methodology for conducting its review, increasing the risk of bias from ad-hoc decision-making by the Panel.
- There is an insufficiently comprehensive and accessible declaration of interests, making it difficult to determine how the interests of the authors may have influenced the outcomes of the Opinion.
- Selection criteria for studies are unclear and seem to have been inconsistently applied – leaving the overall impression that data was used selectively in the review.
- Criteria for the internal validity of studies are only partially stated, it is not clear if they are consistently applied and there seem to be errors in the application of the criteria, making it less likely that the Opinion reliably distinguishes better research from worse.

There are two areas in which the Opinion appears to come much closer to meeting standards of best practice in systematic reviews:

- The search strategy is clearly-stated, sensitive and is probably comprehensive – although it would be helpful if this was given in the main document of the Opinion rather than available on request.
- The answer is clearly stated, though it suffers from the same ambiguities as the question asked.

Overall, the Opinion is therefore probably neither sufficiently transparent nor methodologically robust enough to constitute the sort of document which could adjudicate in a scientific dispute about the toxicity of BPA.

3.3 Case study: EFSA Opinion on BPA 2013

EFSA Panel on food contact materials, enzymes, flavourings and processing aids (CEF). DRAFT Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs – Part: exposure assessment. European Food Safety Authority (EFSA), Parma, Italy.

3.3.1 Clarity of question for review

Reviews should ask a clear, unambiguous question, the formulation and usefulness of which is justified by a presentation of the context in which the review is being conducted.

3.3.1.1 Clarity of context and question

The regulatory context and reasons for doing an exposure assessment are clear, describing the lack of an assessment since 2006 and the regulatory movements in various Member States as being the reasons for conducting a new review. The question being answered by the Panel is also clear, where they are to consider the exposure situations of the general population and vulnerable groups to BPA.

3.3.1.2 Conclusion

The clarity of the question is consistent with a scientific standard for reviewing evidence.

3.3.2 Use of a pre-published protocol

Cochrane reviews require Review Teams to develop and follow robust review protocols which are published prior to conducting a review, in order to avoid bias from subjective and ad-hoc decision-making in the conduct of the review.

3.3.2.1 Presence of a protocol

There was no pre-published protocol for the review. Presentation of methodology is scattered throughout the whole Opinion document.

There is something unclear about the emphasis on conservativeness in the method. The Panel states (line 1477): "In order to quantify the relative impact of each source, the assumptions made in the exposure assessments were aimed at obtaining a similar degree of conservativeness among the different sources."

There does not seem to be an explanation of precisely how different data sources are calibrated or adjusted in order obtain similar degrees of conservativeness, what that degree of conservativeness is, nor if this is a method which will give an accurate indicator of likely exposure.

Intuitively, it seems presenting an estimated exposure aimed at accuracy, with error margins and a judgment of the credibility and conservativeness of the estimate, would be sufficient. The advantage of aiming at conservativeness is therefore not entirely clear. Overall, more explanation of the merits and validity of this methodology seems necessary.

3.3.2.2 Conclusion

The absence of a pre-published protocol is not consistent with a scientific standard for reviewing evidence.

3.3.3 Comprehensive declaration of interests

Cochrane declarations of interest are comprehensive, covering not only financial and professional interests but relevant publishing history and the specific contributions made by each contributor and member of the Review Team. This is to ensure conflicting interests do not distort the results of the review and allow the reader to put the conclusions of the reviewers into their full context.

3.3.3.1 Information about interests

The members of the Panel and BPA Working Group are listed in the Opinion. Organisational affiliations and declarations are not given in the Opinion but are instead available through the EFSA DOI database and, for experts no longer serving with EFSA, by email on request. These declarations are not specific to the Opinion but instead the user needs to read lengthy documents and then interpret for themselves any potential conflicts, which for all Panel and Working Group members is a very lengthy task. The specific contributions made by each Panel and Working Group member is not stated.

3.3.3.2 Conclusion

The declaration of interests is not consistent with a scientific standard for reviewing evidence. It is insufficiently complete while for given information there is the practical difficulty for the user to construct an image of the interests of each member of the Panel and Working Group. Overall, the information presented is insufficient for reliably developing a clear and accurate picture of how those interests might have shaped the findings of the Opinion.

3.3.4 Systematic search method for capturing all evidence of potential relevance

In order to make sure that the evidence surveyed in a review is representative of all of the available evidence (i.e. that sampling bias is avoided), a systematic search strategy should capture all research of potential relevance to the objective of the review, and should be reported in such a way as to be reproducible by a third party.

3.3.4.1 Search method

The Panel used the term ["Bisphenol A" or "BPA"] to interrogate seven on-line research database, ISI Web of Knowledge - Web of Science (WoS), CAB Abstracts, American Chemical Society (ACS), EBSCOhost, Elsevier Science Direct, InformaWorld, SpringerLink. The search was done independently by two experts who compared results and discussed discrepancies. The search method itself seems to be satisfactory, however there is no apparent statement of the search results.

Regarding biomonitoring studies (line 2548) the Panel states: "Data on serum levels of unconjugated, conjugated, and total BPA in humans were retrieved from peer-reviewed scientific papers (published since 2006) which were identified by a systematic literature search." There is no information on how this search was carried out, unless the Panel means the method used was the same literature search as described on line 783. The results of the search are not presented.

The other element of the search strategy was a call for data by EFSA. Although the Panel offers a fairly detailed breakdown of the results of this call for data, the resulting data set may not be representative of all data on BPA exposure because the data is volunteered by Member States rather than being the product of a systematic search.

3.3.4.2 Missing studies

With no summary of which studies and data the Panel found in its search method, there is no way of knowing from the document if the Panel retrieved a sufficient proportion of all the data of possible relevance to the objective of the Opinion.

On analysis of the document it seems likely that data of potential relevance to the objective of the review is missing from the Opinion. For example, given that many epidemiological papers involve a biomonitoring component, there is a lot of data which seems potentially usable by the Panel but which does not feature in the Opinion. Such studies include:

- Stahlhut, R.W., Welshons, W.V., Swan, S.H., 2009. Bisphenol A data in NHANES suggest longer than expected half-life, substantial nonfood exposure, or both. Environ. Health Perspect. 117, 784–789.
- Spanier, A.J., Kahn, R.S., Kunselman, A.R., Hornung, R., Xu, Y., Calafat, A.M., Lanphear, B.P., 2012. Prenatal exposure to bisphenol A and child wheeze from birth to 3 years of age. Environ. Health Perspect. 120, 916–920.
- Perera, F., Vishnevetsky, J., Herbstman, J.B., Calafat, A.M., Xiong, W., Rauh, V., Wang, S., 2012. Prenatal bisphenol a exposure and child behavior in an inner-city cohort. Environ. Health Perspect. 120, 1190–1194.
- Braun, J.M., Yolton, K., Dietrich, K.N., Hornung, R., Ye, X., Calafat, A.M., Lanphear, B.P., 2009. Prenatal bisphenol A exposure and early childhood behavior. Environ. Health Perspect. 117, 1945–1952.
- Wolff, M.S., Engel S.M., Berkowitz G.S., Ye X., Silva M.J, Zhu C., Wetmur J., and Calafat A.M., 2008. Prenatal phenol and phthalate exposures and birth outcomes. Environmental health perspectives 116, 1092-1097
- Fénichel, P., Déchaux, H., Harthe, C., Gal, J., Ferrari, P., Pacini, P., Wagner-Mahler, K., Pugeat, M., Brucker-Davis, F., 2012. Unconjugated bisphenol A cord blood levels in boys with descended or undescended testes. Hum. Reprod. 27, 983–990.
- Engel, S.M., Levy, B., Liu, Z., Kaplan, D., Wolff, M.S., 2006. Xenobiotic phenols in early pregnancy amniotic fluid. Reprod. Toxicol. 21, 110–112.

From this list, Stahlhut et al. (2009) seems of particular relevance to estimating BPA exposure, given that it is concerned with BPA half-life in the blood, and Fénichel et al. (2012) seems relevant because of its cohort of French neonates.

3.3.4.3 Conclusion

The search method is not consistent with a scientific standard for reviewing evidence. Although the search method itself seems comprehensive, insufficient information is given about how many studies were retrieved by the strategy, making it difficult to judge if the data reviewed was representative of all possible data relevant to the objective of the Opinion. A parallel search for relevant data, it suggests the Opinion failed to retrieve large quantities of information relevant to the exposure assessment.

3.3.5 Selection process for putting forward to analysis all relevant research from the citations yielded by the search strategy

In order to reduce the risk of selection bias in the use of relevant information yielded by the search process, clear inclusion and exclusion criteria should be stated for selecting from the results of the literature search the specific references relevant for answering the review Question. All data from each study included in the review which is relevant to the review objective should be included in the review.

3.3.5.1 The Panel's inclusion and exclusion criteria

Information about inclusion and exclusion criteria for data in the Opinion is given at line 787: "Emphasis was put on migration studies on BPA, occurrence and intake levels of BPA from various dietary sources for the general population […]." Non-EU studies are generally excluded, though mechanistic studies and Japanese data are included, so this is not consistently applied as an exclusion criterion.

Because it is not explained what "emphasis" means in terms of how studies were treated in the Opinion, there is insufficient information to allow reproduction of the Panel's method for identifying studies for inclusion in the Opinion.

3.3.5.2 Inconsistent use of selection criteria

The Panel is deliberately selective in its use of its selection criteria, stating: "The CEF Panel noted that only very few data from Europe and/or obtained by a reliable analytical method were available and therefore decided to take into account data from Japan."

Making exceptions in this way risks bias because it puts data to analysis in an ad-hoc rather than systematic way. At the very least, explanation as to why the Japanese data is considered the most appropriate is needed, particularly considering its use of ELISA for measuring BPA exposure – a feature which would normally lead to a study being excluded from the analysis.

3.3.5.3 Conclusion

The process for selecting studies from the search results for review in the Opinion is not consistent with a scientific standard for reviewing evidence. The selection criteria are not clearly stated, so the method is not reproducible, and the results of the selection process not presented. Furthermore, the criteria are not applied consistently, with studies seeming to meet the inclusion criteria absent from the review, while studies which do not meet the criteria are included in the review.

3.3.6 External validity

Judgments of external validity are concerned with the relevance of a study for answering a review question: can an observation in one study group, such as an epidemiological cohort or group of mice in an animal study, be taken as representative of effects in the general population?

3.3.6.1 The Panel's approach to evaluating the external validity of data

There is no explicit discussion of how the external validity of data in the review is treated. Some of this discussion is interwoven with criteria for study quality. Some aspects of external validity (such as data from outside the EU region) are used as exclusion criteria. The issue of validity is also addressed in the uncertainty calculations, in which data generated in one region and therefore of unknown validity to another, is handled by introducing an uncertainty factor. There is also a question of whether it is even relevant to talk about validity of studies when the intent of the Opinion is to find the likely upper boundary of exposure of Europeans to BPA.

3.3.6.2 Conclusion

The appraisal of the external validity of data in the Opinion is not evaluable.

3.3.7 Internal validity

Internal validity is concerned with the credibility of a study, i.e. the risk of it being wrong either through random error (such as being statistically under-powered) or systematic error (bias). The criteria for judging the internal validity of animal and epidemiological studies are complex and therefore require careful planning and statement prior to conduct of the review, and care should be taken in ensuring each study is subjected to a fair test for internal validity.

These criteria should at least identify a material rather than hypothetical risk of bias or error, appraise the magnitude and direction of the risk of bias, and not treat as equivalent limitations in the reporting and conduct of a study. The rationale for evaluating the internal validity of a study should be transparent, justified and consistently applied to all studies.

3.3.7.1 The Panel's method for appraising the internal validity of studies

Appendix I is where the Panel presents their method for evaluating the quality of included studies. The Panel states: "The JRC guidelines on performance criteria and validation procedures of analytical methods used in controls of food contact materials were used as the basis to define the criteria for all methods considered in this opinion (JRC, 2009)."

These performance criteria relate to the analytical methods used to measure BPA in environmental samples. Those described by the Panel are recovery, repeatability, and limits of detection and quantification (LOD/LOQ). The Panel presents a table of the acceptable recovery values for analytical methods, states that analytical methods "should not exceed the level calculated by the Horwitz Equation", and for LOD/LOQ a method to correct for bias from samples reporting levels of BPA below the LOQ or LOD is presented.

There is also a LOD cut-off of 15 µg/kg and an LOQ cut-off of 50 µg/kg. For biomonitoring studies, "methods reporting LOD values greater than 0.4 µg/kg or LOQ values greater than 1.3 µg/kg were excluded from the exposure assessment".

The Panel also defines "supplementary criteria", though how these are applied is unclear. These include:

- "The selectivity of the method, i.e. whether or not interferences had been considered"
- "Whether or not measures had been taken to reduce or avoid background contamination"

- "Whether or not the method-performance data described have been derived for an appropriate matrix and at a concentration relevant to the levels measured in the samples"

The Opinion also presents a series of extraction and instrumental analysis techniques.

For biomonitoring data, the Panel states (line 2015): "The quality of each study was assessed on the basis of the criteria given in Appendix I." Appendix I is concerned with sampling techniques rather than all of the quality control issues involved with conducting biomonitoring studies. In addition, the panel states (line 2569): "For the assessment of reported serum BPA levels, the following aspects were specifically assessed:

- the proportion of detectable/quantifiable values in relation to the LOD/LOQ
- the proportion of unconjugated BPA in the total BPA serum concentration
- the average serum concentrations of unconjugated (U), conjugated (C) and
- total (T) BPA for studies reporting ≥50 % detectable values."

3.3.7.2 General comments on the Panel's method

Although the description of the criteria used in assessing data is lengthy, there does not appear to be an explanation of how these criteria are used. For example, it is not possible to determine if the quality criteria described are intended as inclusion criteria, presenting a positive list of techniques and cut-offs which a study has to meet in order to be considered, or if these are used as quality criteria. (As an additional point, few users of the Opinion will be experienced analytical chemists, so more explanation of the basic concepts presented here would likely be useful.)

Overall, there seems to be little discussion of the internal validity of studies, with all the included studies seeming to be treated as if they are equally valid. For example, migration data from PlasticsEurope (line 986) seems to be used uncritically for estimating water cooler exposure. This exposure estimate is much lower and produced by different experimental conditions to the other published data; an assessment of its relative credibility compared to the other studies would therefore seem to be appropriate but is absent.

There are other places in which it seems that the findings of single studies, interpreted by the Panel as offering the best estimates of BPA exposure, are taken as definitive. In Table 4, the "most extensive", "most reliable" or single available estimates are taken as definitive. For example, the data from [EFSA] Juberg et al. (2001) is taken as a definitive measure of BPA exposure from pacifiers with polycarbonate shields. The reasoning for this is not given.

Even if there is only one data source for an exposure estimate, the internal and external validity of the study should still be appraised – just because it is the only data available, it does not mean that any reliable conclusions can be drawn from it.

If several data sources are available, it is not necessarily appropriate to only use the single strongest data source. Excluding weaker data will reduce the precision of the exposure estimates and potentially reduces the utility of the review by treating data of potential value as if it has none at all. If a single study is to be preferred, then criteria for judging the preferred study to be of the highest quality of those available should be presented. Since this is not done in the Opinion, it is not possible to determine if the Panel's method for appraising data quality has led to the Opinion being based on the best available estimates of exposure.

3.3.7.2.1 Following up with authors about issues of methodological quality

There is evidence that issues of methodological quality were followed up with authors. For example, at line 2714 the Panel states: "Quality-control (QC) materials and standards were prepared from pooled human milk which derived from samples collected over several days from two donors (A. Cariot, pers. communication)". There are also "the box-percentile plots for unconjugated and total BPA (Figure 9, percentiles kindly provided by S. Duty)" (line 2784). This is the right thing to do but must be systematic so as not to introduce a bias by creating a sub-group of followed-up studies in the Opinion.

3.3.7.3 Conclusion

The assessment of the internal validity of studies in the Opinion is not consistent with a scientific standard for reviewing evidence because although criteria for study quality are presented, it is not clear how they are used. Furthermore, included studies seem to be treated as if they are all equally internally valid. Since this is unlikely to be true, there is a risk that the study either puts too much weight on smaller or less methodologically robust studies, or too little weight on larger or more robust studies.

3.7.8 Answer

The answer to a review should be unambiguous and reflect what is stated by the research which has been evaluated and the level of confidence which can be had in those findings.

3.3.8.1 Comparison with other assessments

It is very helpful to have in [EFSA] section 4.9.2 a comparison with other BPA exposure assessments. Comment on why the Panel methodology is superior would, however, help justify the Panel's methodological approach as compared to those employed by other expert groups. For example, the FAO/WHO Expert Meeting's decision to model BPA exposure according to the frequency of consumption of packaged food seems to have merit in estimating low and high levels of exposure, particularly in anticipating how exposure might vary between different socio-economic groups. Since this is information which might be of particular help to risk managers, justification of the Panel's choice of method would likely increase the utility of the Opinion.

3.3.8.2 Judgment of plausibility of the modeled exposure

At line 3265, the Panel states: "Overall the Panel concludes that all values covered by the combined uncertainty intervals for the two estimates remain plausible." It is helpful that the uncertainty estimates have been tested – however, it is not entirely surprising that the Panel finds its own results plausible, as it is unlikely that anyone who had gone to great effort to calculate a correct exposure would think their final calculation was incorrect (otherwise why would they have offered it as an answer in the first place?).

3.3.8.3 Discrepancies between the summaries and the body of the Opinion

The uncertainty charts show a range for exposure to BPA up to 1100 ng/kgbw/d, whereas the abstract only presents the result as "up to 857 ng/kg bw/day". Other uncertainties are not stated in the Abstract or Summaries, for example those relating to biomonitoring studies, such as on page 214 where the Panel states: "Biomonitoring studies may, therefore, not have captured high levels of exposure that may occur in specific geographic areas or specific population groups," and page 216: "The main sources of uncertainty in the estimation of high total exposure based on biomonitoring data is the sampling uncertainty due to limitations in the representativity of the available information on total BPA concentration in urine, the distribution uncertainty in the 95th percentile, and the uncertainty in the specific urinary output rate."

3.3.8.4 Conclusion

The answer of the Opinion is not consistent with a scientific standard for reviewing evidence, in that the Abstract and Summary seem to only

partially represent the overall findings of the review. Explanation of the superiority of the Panel's method over other assessments would likely be helpful to users.

3.3.9 Overall conclusions about the Opinion

There are substantial differences between best practice in systematic review as conducted in evidence-based medicine and in how research is appraised in the 2013 Opinion on BPA:

- The Panel does not follow a pre-defined methodology for conducting its review, increasing the risk of bias from ad-hoc decision-making by the Panel. Several points of method seem to risk introducing random or systematic error into the Review.
- There is an insufficiently comprehensive and accessible declaration of interests, making it difficult to determine if the interests of the authors have influenced the outcomes of the Opinion.
- Although the search strategy is clearly-stated, there is insufficient information about its results, making it impossible to evaluate if all evidence of potential relevance to the review had been located.
- The selection criteria for studies are unclear and seem to be inconsistently applied, with studies seeming to meet the inclusion criteria absent from the review, while studies which do not meet the criteria are included in the review. This gives the overall impression that data is used selectively.
- Criteria for the internal validity of studies are only partially stated, it is not clear how they are used, and there seems to be little appraisal the quality of data which is included in the review. Some studies are taken as definitive but without explanation as to why.
- The answer to the Opinion, as stated in the Summary and Abstract, does not seem to fully correlate with the results presented in the main body of the Opinion.

There is one area in which the Opinion appears to be relatively strong in comparison to best practice in systematic reviews:

- The question which the Panel is answering is clearly and unambiguously specified.

3.4 Overall conclusions from the case studies

In 2010, EFSA published a guidance document presenting a comprehensive account of systematic review techniques and how they might support EFSA's work in risk assessment in the area of food and feed safety assessments (European Food Safety Authority 2010), developed with input from the Cochrane Collaboration. In 2012, EFSA gave a presentation about how the next Scientific Opinion on BPA would draw on some of the systematic review techniques outlined in the 2010 guidance (Husøy 2012). In another presentation in 2013, EFSA described the training its staff have received in systematic review techniques, what has already been implemented, and outlines future plans for further training and for contracting out various elements of the systematic review process (Verloo 2013).

It was therefore hoped that, in putting together these two case studies, at least some elements of systematic review methods would discernible in the 2013 Opinion which were absent from the 2010 Opinion. Although there appears to have been some progress, such as in the formulation of the review question and the signs that the Panel is following-up with researchers to gather extra data missing from study reports (though this needs to be made systematic) there are also areas where there seems to have been regression, such as the search method in the 2013 Opinion appearing to be weaker than that of the 2010 Opinion, and the answer to the 2013 Opinion not seeming to be fully representative of actual findings.

Overall, progress seems to be haphazard, with the approach taken by the Panel in 2013 is more-or-less the same as that taken in 2010. An absence of protocol, insufficiently comprehensive declaration of interests, lack of transparency in search results, lack of clarity in method for evaluating study quality, and insufficient safeguards against selective use of data all militate against the capacity of either the 2010 or 2013 Opinions to represent a scientifically robust, state-of-the-art assessment of the toxicity of BPA. The inability of EFSA to demonstrate the robustness of its methods will weaken the Agency's ability to defend itself against accusations that improper influence has been exerted over the results of its Opinions.

The situation at EFSA regarding recent reviews of BPA appears to be not all that different to that found by Cynthia Mulrow in medical research in 1987

The situation at EFSA regarding recent reviews of BPA appears to be not all that different to that found by Cynthia Mulrow in medical research in 1987, when she reviewed the state of the science of medical review articles and concluded that "medical reviews do not routinely use scientific methods to identify, assess and synthesize information" (Mulrow 1987).

As researchers such as Mulrow did in the late 1980s, we therefore conclude this section with the hypothesis that the use of systematic review methods will greatly enhance the ability of EFSA's Opinions to present the best possible statement of what is known about the toxicity of compounds such as BPA and thereby better position the agency for resolving disputes about the risk assessment of chemicals.

3.4.1 Table summarizing case study findings, with additional comments

Component of systematic review	EFSA 2010	EFSA 2013	General Comments
1. Clarity of question for review. *Reviews should ask a clear, unambiguous question, the formulation and usefulness of which is justified by a presentation of the context in which the review is being conducted.*	The objective is not consistent with a scientific standard for reviewing evidence because it is ambiguous, while insufficient justification is given in the text for the Panel's choice of operating a narrow interpretation of the terms of reference for the Opinion. The European Commission decided on the scope of the Opinion rather than scope being developed in a consultative process between reviewers, issue experts and users of the final review document.	The clarity of the question is consistent with a scientific standard for reviewing evidence. As for 2010, the European Commission decided on the scope of the Opinion.	There is improvement in the clarity and framing of the question for review. However, that the European Commission has primary responsibility for determining the scope of a Scientific Opinion may put limits on the relevance and usability of the final document, in comparison to the open expert-led consultative processes used by the Cochrane Collaboration. The EU Scientific Committees have been critical of how the scope of risk assessments and socio-economic analyses is currently generated within Commission Services "without reference to other stakeholders or the scientific committees", with scientific committees only being able to modify them so as to make them more answerable. The result, in the opinion of the Committees, is that EU experts have insufficient ownership of the problems to which they are supposed to provide answers (SCHER et al. 2013, p. 30).
2. Use of a pre-published protocol *Cochrane reviews require Review Teams to develop and follow robust review protocols which are published prior to conducting a review, in order to avoid bias from subjective and ad-hoc decision-making in the conduct of the review.*	The lack of pre-published protocol is not consistent with a scientific standard for reviewing evidence.	As before	Beginning the process of developing Scientific Opinions with a pre-published protocol would offer opportunities for tightening up methodological issues before an Opinion makes it all the way to draft and final publication and reduce risk of inconsistent application of selection criteria and inconsistent evaluation of the validity of included studies.
3. Comprehensive declaration of interests *Cochrane declarations of interest are comprehensive, covering not only financial and professional interests but relevant publishing history and the specific contributions made by each contributor and member of the Review Team. This is to ensure conflicting interests do not distort the results of the review and allow the reader to put the conclusions of the reviewers into their full context.*	The declaration of interests is not consistent with a scientific standard for reviewing evidence. It is insufficiently complete, while for given information there is the practical difficulty for the user to construct an image of the interests of each member of the Panel and Working Group. Overall, the information presented is insufficient for reliably developing a clear and accurate picture of how those interests might have shaped the findings of the Opinion.	As before	Although conflicts and declarations of interest have been a prominent issue around EFSA's work since 2010 (European Ombudsman 2011; Robinson et al. 2013; Butler 2012), there appears to have been little visible change at the point of use of Opinions. In neither case can a user readily discern how the interests of the members of the Panels and Working Groups might have influenced the development of the Opinion.
4. Systematic search method for capturing all evidence of potential relevance to the review. *In order to make sure that the evidence surveyed in a review is representative of all of the available evidence (i.e. that sampling bias is avoided), a systematic search strategy should capture all research of potential relevance to the objective of the review, and should be reported in such a way as to be reproducible by a third party*	The search strategy seems consistent with a scientific standard for reviewing evidence. That it failed to retrieve some eligible studies could be of concern in relation to sampling bias, though in this instance appears unlikely to have materially affected the findings of the review. Reporting of the results of the search process could have been clearer.	The search method is not consistent with a scientific standard for reviewing evidence. Although the search method itself seems comprehensive, insufficient information is given about how many studies were retrieved by the strategy, making it difficult to judge if the data reviewed was representative of all possible data relevant to the objective of the Opinion. In running a parallel search for relevant data, it seems the Opinion failed to retrieve large quantities of information relevant to the exposure assessment.	The 2013 Opinion represents a backward step in search method, where much less information about search results has been given, such that it is more difficult in 2013 than in 2010 to be confident that the search for data was comprehensive. Given that a full description of search methods and results is one of the more straightforward elements of systematic review to implement, and appears to have been achieved in 2010, it is surprising that this was not carried through to 2013.

Component of systematic review	EFSA 2010	EFSA 2013	General Comments
5. Selection process for putting forward to analysis all relevant research from the citations yielded by the search strategy *In order to reduce the risk of selection bias in the use of relevant information yielded by the search process, clear inclusion and exclusion criteria should be stated for selecting from the results of the literature search the specific references relevant for answering the review Question. All data from each study included in the review which is relevant to the review objective should be included in the review.*	The selection process used in the Opinion is not consistent with a scientific standard for reviewing evidence. There is evidence that studies which are apparently relevant to the Opinion and were retrieved by the search process were not included in the Opinion. At the same time, studies which did not meet the inclusion criteria were nonetheless included in the Opinion.	The process for selecting studies from the search results for review in the Opinion is not consistent with a scientific standard for reviewing evidence. The selection criteria are not clearly stated, so the method is not reproducible, and the results of the selection process not presented. Furthermore, the criteria are not applied consistently, with studies seeming to meet the inclusion criteria absent from the review, while studies which do not meet the criteria are included in the review.	Neither the 2010 nor 2013 Opinions offer enough information about the selection process to ensure that the results are reproducible. There is also cause for concern that the selection criteria are not being used consistently, increasing risk of selection bias in the data being put forward to analysis in the Opinions.
6. Assessment of the external validity of included studies. *Judgments of external validity are concerned with the relevance of a study for answering a review question: can an observation in one study group, such as an epidemiological cohort or group of mice in an animal study, be taken as representative of effects in the general population?*	That internal and external validity are not distinguished by the Panel in the discussion of study quality makes it difficult to discern if a study is being downgraded because it is not methodologically robust or because it is of limited direct relevance to the evaluation of the toxicity of BPA in humans. The methodological quality of appraisal of external validity is therefore not readily evaluable.	Similarly, the methodological quality of appraising of the external validity of data in the Opinion is not evaluable.	Introducing an explicit distinction between internal and external validity would be a new approach to discussion of study quality in Scientific Opinions, so it would not be fair to be overly critical of shortcomings in this aspect of reviews – except to say, this approach should be introduced as it will improve analytical quality of Opinions and the transparency and reproducibility of results.
7. Assessment of the internal validity of the included studies. *Internal validity is concerned with the credibility of a study, i.e. the risk of it being wrong either through random error (such as being statistically under-powered) or systematic error (bias). The criteria for judging the internal validity of animal and epidemiological studies are complex and therefore require careful planning and statement prior to conduct of the review, and care should be taken in ensuring each study is subjected to a fair test for internal validity.* *These criteria should at least identify a material rather than hypothetical risk of bias or error, appraise the magnitude and direction of the risk of bias, and not treat as equivalent limitations in the reporting and conduct of a study. The rationale for evaluating the internal validity of a study should be transparent, justified and consistently applied to all studies.*	The assessment of the internal validity of studies included in the Opinion is not consistent with using a scientific method for reviewing evidence. The criteria for appraising internal validity are not fully and explicitly stated, so it is not possible to determine if the studies included in the review have been subjected to a fair test for credibility. There appear to be mistakes in applying the criteria which could lead to over- or underestimation of the internal validity of included studies.	The assessment of the internal validity of studies in the Opinion is not consistent with a scientific standard for reviewing evidence. Although criteria for study quality are presented, it is not clear how they are used. Furthermore, included studies seem to be treated as if they are all equally internally valid. Since this is unlikely to be true, there is a risk that the study either puts too much weight on smaller or less methodologically robust studies, or too little weight on larger or more robust studies.	That each study included in the review may not have been subjected to the same fair test for methodological quality is a major concern, as it will result in selective use of data in the Opinion. In mitigation, validated tools for assessing the internal validity of toxicological studies have yet to be developed.
8. Clarity of answer to the question for review. The answer to a review should be unambiguous and reflect what is stated by the research which has been evaluated and the level of confidence which can be had in those findings	The clarity of the answer to the question is consistent with a scientific standard for reviewing evidence, ambiguity in the question notwithstanding.	The answer of the Opinion is not consistent with a scientific standard for reviewing evidence, in that the Abstract and Summary seem to only partially represent the overall findings of the review. Explanation of the superiority of the Panel's method over other assessments would likely be helpful to users.	It is surprising that the answer to the 2013 Opinion does not better represent the results of the body of the review – this should be easy to rectify.

Part Four.

Strategic Recommendations

"A systematic review uses a process to identify comprehensively all studies for a specific focused question (drawn from research and other sources), appraise the methods of the studies, summarize the results, present key findings, identify reasons for different results across studies, and cite limitations of current knowledge. In a systematic review, all decisions used to compile information are meant to be explicit, allowing the reader to gauge for him- or herself the quality of the review process and the potential for bias. In this way, systematic reviews tend to be more transparent than their narrative cousins, although they too can be biased if the selection or emphasis of certain primary studies is influenced by the preconceived notions of the authors or funding sources." (Garg, Hackam et al. 2008)

4 Strategic Recommendations

4.1 A premium on accessibility

EFSA Opinions are definitive documents with very strong regulatory impact and wide use; they are the documents everybody involved in food safety and chemicals policy relating to foodstuffs should be referring to when deciding what to do about the substances which fall under EFSA's purview. This gives them a very wide user base, including risk assessors and risk managers in EU institutions, MEPs and civil servants, non-government organisations, manufacturers and users of chemicals and food contact materials, industry associations, academics and researchers, the general public, journalists, and so on.

There is therefore a premium on making these Opinions accessible to educated non-specialists who have relatively limited time but need relatively quick assurance that the Opinion is the result of a comprehensive and objective process – because if the documents are insufficiently understood or trusted, they will not form the basis of policy decisions.

The challenge, however, is that the two case studies in this report suggest that there is much which needs to be done in order to put risk assessment on a solid, transparent evidential base.

The following recommendations should help introduce systematic review techniques into the risk assessment process. They cover changes to the Opinion documents themselves, the processes by which Opinions are generated, and the research which needs to be done to address unresolved methodological issues relating to the use of systematic review techniques in chemical risk assessment.

This is supported by an overview of current research initiatives which could provide research capacity for informing these changes. Finally, there is a "next steps" section in which a series of short- and medium-term goals are defined which will lead to greater use of systematic review techniques in EU risk assessments.

4.2 Strengthening Scientific Opinions

The following recommendations concern the documentary means for strengthening Scientific Opinions.

4.2.1 Publish protocols in advance of conducting Opinions

Advance publication of a protocol prior to conduct of a review reduces the risk of bias in the review by discouraging ad-hoc decision-making by the reviewers.

Advance publication also provides an extra opportunity for user feedback, which helps ensure methodological robustness and maximum utility of a review.

Although there are many outstanding methodological issues surrounding the systematic review of toxicological data for the purpose of developing Scientific Opinions, it should already be possible to describe and consult on some elements of a protocol prior to conduct of an Opinion. These at least include the question to be asked, the search strategy and the inclusion and exclusion criteria for selecting studies for review.

Opening up the review process through advance publication of and consultation on a protocol would be one way to increase ownership of the scientific problems addressed by EU experts. As will be discussed later in this section, developing a system whereby Opinion objectives can be generated in a bottom-up process would further help with this problem.

4.2.2 Issue guidance on the structure and writing of Opinions

There is work to be done in assessing the extent to which Opinions are found by their users to be understandable and how modifications to structure (such as introducing a standard structure), length (it would be fair to say that expert Opinions are not generally noteworthy for their concision) and readability (reducing the required expertise for the informed reader) might improve usability.

4.2.2.1 Structuring Scientific Opinions

The structure of both the 2010 and 2013 Opinions is unorthodox for scientific publications, which typically consist of an introduction, method, results, discussion and conclusions. This may lead to unnecessary confusion for readers, as information relevant to understanding the Opinion may be difficult to find.

For example, in the 2013 Draft Opinion there is a section titled "Handling of Data", which discusses the method the Panel uses for adjusting for left-censored data. At the same hierarchical level in the document there is a section concerned with data about BPA migration from food contact materials into food simulants. This is in spite of the former being methodological while the latter presents results. This mixing of method and results happens again when "General Assumptions for Calculation" (a methodological consideration) are presented alongside "Exposure Estimation from Dietary Sources" (presenting results) in section 4.6.

As a very large document the structuring requirements of an Opinion may need to be complex, and it is not reasonable to expect a Scientific Opinion to be accessible to everybody. Nonetheless there should still be value in adhering to a rigorous information hierarchy which minimises the risk that a user simply misunderstands elements of an Opinion because the structure of the document has undermined their ability to process it.

EFSA should therefore develop guidance on the structure of Scientific Opinions, aimed at achieving maximum ease-of-use of Opinions for users and ensuring consistency in presentation of information between different Opinions.

4.2.2.2 Use of summaries, tables and charts

In general, both the 2010 and 2013 Scientific Opinions favour narrative text over tables for summarising the methods and results of included studies.

The rationale for presenting data in tables is simple: since the data extracted from studies are what constitute the results of a review, and results are normally best presented in tabular format, it makes sense to present the extracted data in tables. This also encourages consistency of presentation and makes it easier for the reader to see the results.

The 2013 Opinion makes better use of tables and charts than the 2010 Opinion, particularly with its large study quality tables in Appendix IX and clear summaries of biomonitoring data in Figures 2, 5 and 8. However, EFSA could go much further with its use of tables, presenting the key methodological and results data it is extracting from the studies included in the review in a format similar to that used by the Cochrane Collaboration.

4.2.3 Tighten up controls on the interests of the authors of Opinions

4.2.3.1 Avoid direct financial conflicts of interest, carefully manage indirect and research interests

Any person with an interest which could be perceived as a direct financial stake in the outcome, such as recent, current or future employment, ownership of patents and shares, and so forth, should be disbarred from conducting an Opinion. Additionally, authors of Opinions reviewing their own research should be discouraged. If it cannot be avoided, then the controls in place to limit subsequent risk of bias should be laid out in the Protocol.

Indirect financial interests are a much more difficult issue: some people

secure funding on the basis of making progress towards restrictions being placed on the use of certain chemicals; others secure funding by arguing against such restrictions; in general, academics find it easier to secure funding by discovering the presence of problems rather than the absence of them.

Unpicking such interests is difficult and it is beyond the scope of this report to make general recommendations as to how to deal with them, except to say that even the appearance that interests have prejudiced a review can do major damage to its credibility and must therefore be very carefully managed.

4.2.3.2 Include comprehensive declarations of interest in Opinions

In addition to restrictions on direct financial interests, all other relevant interests should be stated in a comprehensive declaration of interests, along with a description of the contribution made by each person involved in the review process. This should cover not only employment but extend to anything might be perceived by readers as capable of influencing an author's judgments, including political, academic and other interests.

This is not to prejudice the work of the reviewer but to allow the reader to put the conclusions of the review in their full context.

4.2.4 Be systematic and transparent in the evaluation of included research

4.2.4.1 Comment on all studies found and included in the literature review

It is not clear from either the 2010 or 2013 studies just how comprehensive the literature searches were, nor which studies were considered by the Panels to meet the inclusion criteria for the Opinions, nor why some studies appeared to be excluded from review.

To address this, EFSA should issue guidance for a transparent, systematic approach to reporting the results of literature searches and the selection process for including studies in the analysis in Scientific Opinions. In Cochrane Reviews, diagrams showing the results of the search process, lists of excluded studies, summaries of key data from included studies and so forth reassure the user that all the possible sources of data have been considered and that all relevant data has been included.

4.2.4.2 Distinguish between quality of reporting and quality of conduct of research

Both the 2013 and 2010 Opinions conflate quality of reporting with the quality of conduct of research. This unnecessarily downgrades evidence which is stronger than study reports might suggest. The 2013 Opinion shows evidence of follow-up with researchers to resolve questions about study methods; this is the correct thing to do but in future must be systematic.

4.2.4.3 Treat separately internal and external validity in assessing methodological quality

In risk assessment the concept of methodological quality of research tends to be absorbed by concepts such as "reliable", "reliable for risk assessment" and "research quality", and incorporates concepts relating both to the internal and external validity of a study. Running the two concepts together in the analysis makes it unclear if a study is downgraded because it is of lower methodological quality or of limited external validity. This is an important distinction which should be addressed separately in Scientific Opinions.

4.2.4.4 Be systematic about evaluating study quality, focusing on material risk of bias

Evaluation of study quality is presented in the form of narrative text in both the 2010 and 2013 Opinions. The presentation is not systematic, with different problems being presented for each study (not all of which squarely address material risk of bias) while at times critique does not seem to be wholly consistent. Overall, it is not clear if all the included studies are being subjected to a fair test of methodological quality.

EFSA should issue guidance on the consistent and systematic evaluation and presentation of the assessment of methodological quality of studies included in an Opinion.

4.2.5 Institute a peer-review process for revising and accepting Opinions

To ensure that all Opinions conform as far as possible to standards of best practice in their writing and conduct, a peer-review process should be instituted for approval of Opinions before publication as drafts and final documents.

4.3 Reorganise the processes by which Scientific Opinions are produced

Developing sufficient capacity to meet demand for useful, high-quality reviews while operating under tight resource constraints is a major challenge: the Cochrane Collaboration is made up of 31,000 people operating in more than 100 countries, 70% of whom are authors of Cochrane Reviews (Cochrane Collaboration 2013). Reproducing this capacity requires a fundamental re-think of the processes by which Scientific Opinions are generated.

The following premises should inform consideration of what would be a step-change in the conduct of Scientific Opinions in the EU:

1. So long as a Scientific Opinion is produced according to a sufficient standard, it does not matter who authors it (so long as there are adequate controls on conflicts of interest).

2. The demand for Scientific Opinions in the risk assessment community can be used to define Opinion objectives.

3. Scientific Opinions only need to be conducted by small teams of reviewers with some logistical and editorial support, with two primary authors often likely to be sufficient.

4. Authors will volunteer to produce Scientific Opinions, so long as they are scientific publications with the same caché as papers published in journals, advancing the careers of those who publish them.

5. Research funders will support the production of Scientific Opinions, so long as they are of sufficient quality to make significant contributions to advancing knowledge in the environmental health sciences.

If each of these holds true, it should be possible to increase the number of academics and the amount of financial support available for producing high-quality Opinions, without any EU institution necessarily having to pay more to produce each Opinion.

It also changes the function of EU institutions from being organisations which directly author Scientific Opinions to ones which concentrate on setting and enforcing the standards for Scientific Opinions, providing logistical support for their production, making sure the right topics are being addressed in the right way, and in peer-reviewing protocols and draft Opinions.

The structure of such an organisation could consist of equivalents to the following elements, as constitute the Cochrane Collaboration (The Cochrane Collaboration 2013):

- **Review Groups**, to deal with specific topics in chemical risk assessment and manage the production of reviews, including the assembly of review teams for conducting specific reviews;

- **Advisory Groups**, to assist Review Groups and Review Teams in developing protocols for reviews;

- **Methods Groups** to deal with issues such as guidance for assessing the internal validity and external validity of research, judging the overall quality of a body of evidence, statistical techniques and so forth;

- **An Operations Unit** responsible for overseeing the mechanisms that ensure systematic reviews are of maximum objectivity and utility;

In addition, an independent auditing body should be created to ensure that reviews being published under such a structure are of unimpeachable quality.

4.4 Research goals

The overall goal is to develop a tool which combines the judgments of the overall directness and quality of a body of evidence into a single judgment of the overall strength of evidence supporting the answer to a question posed in a Scientific Opinion.

In medicine this already exists as the GRADE framework; however, in chemical risk assessment Cochrane structures and procedures are not yet established for many elements of the types of questions faced by regulatory risk assessors.

Research goals therefore break down into the following three components:

1. **A tool for appraising the internal validity of included studies**, and the synthesis of these judgments into a general statement of the overall quality of the body of evidence.

2. **A tool for appraising the external validity of included studies**, and the synthesis of these judgments into a general statement of the overall directness of the body of evidence.

3. **A tool for combining the quality and directness of evidence** into an overall statement of the strength of the body of evidence.

Any such tools would need to be validated before being introduced into general use.

4.4.1 Current research initiatives

The following is a selection of research initiatives concerned with developing various elements of what would be a systematic review procedure for evaluating data on chemical toxicity. It should be noted that not all of the projects will have equal merit, nor will they necessarily have been conceived with the development of systematic review methods in mind. The citations are examples of publications by the research organisations; the names are individuals associated with the work.

4.4.1.1 Full systematic review approaches

The Navigation Guide. Dr Patrice Sutton, Dr Tracey Woodruff. Applying medicine's GRADE methodology for appraising the overall strength of a body of evidence to the safety assessment and risk management of chemicals. Probably the most fully worked-up approach to applying systematic review techniques to syntheses of toxicological data. (Woodruff, Sutton 2011a, 2011b)

The Evidence-Based Toxicology Collaboration. Dr Sebastian Hoffmann, Prof Thomas Hartung. A long-standing and recently reinvigorated international network of chemicals policy stakeholders concerned with introducing a range of concepts from evidence-based medicine into chemical risk assessment. Based at Johns Hopkins University in the United States. (Hoffmann, Hartung 2006)

European Food Safety Authority Scientific Assessment Support Unit. Dr Didier Verloo. Application of systematic review methods to food and fee safety assessments – a comprehensive exploration of the issues within a specific regulatory context. (European Food Safety Authority 2010)

US National Toxicology Panel. Dr Kristina Thayer. The development of systematic review techniques for evaluating the toxicity of chemicals. Currently revising two draft systematic review protocols after a public consultation (Birnbaum et al. 2013) and has conducted workshop reviews using a range of systematic review techniques (Maull et al. 2012).

UK Joint Water Evidence Group. Dr Deborah Coughlin (Imperial College London). An exploration of rapid, resource-light approaches to systematic review of ecotoxicity data. (Coughlin)

4.4.1.2 Methodological components of systematic review

Stockholm University. Dr Marlene Ågerstrand, Prof Christina Rudén. Criteria for the reporting and evaluation of research, to contribute to the transparent use of toxicity data in environmental risk assessment. (Ågerstrand et al. 2011a)

American Chemistry Council, Conrad Law and Policy Counsel. Dr Richard Becker, Dr James Conrad. Criteria for assessing the credibility of a scientific study. (Conrad, Becker 2011)

European Centre for the Validation of Alternative Methods. ToxRTool, a spreadsheet-based toolkit for evaluating the quality of a study. (Schneider et al. 2009)

Klimisch Ring Test. Dr Robert Kase. An attempt to develop and validate an improved scheme for evaluating study quality based on the widely-used Klimisch criteria. Unpublished at the time of writing. (Kase 2013)

Cochrane Non-Randomised Studies Methods Group. Prof Barnaby Reeves. The development of tools for the appraisal of non-randomised studies, such as epidemiological research. (Deeks et al. 2003)

CAMARADES Collaboration. Prof Malcolm Macleod (University of Edinburgh). The development of techniques for appraising the quality and conducting meta-analyses of animal studies. (Macleod 2011)

Louvain Center for Toxicology and Applied Pharmacology, Université Catholique de Louvain. Prof Geneviève Van Maele-Fabry. Several systematic reviews of epidemiological data exploring associations between exposure to environmental chemicals and their effects on human health. (van Maele-Fabry et al. 2012)

4.5 Next steps

4.5.1 Long-term goal

In chemical risk assessment, we should be looking to establish a gold standard for evidence review. As yet, nobody is supplying one. To achieve that, we need to establish a body which manages the production of systematic reviews of toxicological research, while systematic review techniques should be used as much as practicably possible in the conduct of Scientific Opinions.

- **Establish a Cochrane Collaboration-like group** for managing the production of systematic reviews of chemical toxicity.

4.5.2 Short-term goals (1 year)

Every future Scientific Opinion represents an opportunity to inch towards the full implementation of systematic review methods. All imminent Scientific Opinions, including the next Scientific Opinion by EFSA on BPA (the hazard component of the overall risk assessment) should at least be

structured to maximize ease of understanding, include a comprehensive declaration of interests, present the full results of the evidence search and selection processes, and have a clear description of the methods used for appraising and synthesizing the studies included in the Opinions. To support this programme, EFSA scientific staff and experts should receive training in systematic review techniques. In addition, researchers should be piloting more systematic reviews to identify research challenges.

In the short term, the following needs to happen:

- **Resolve the issues with documenting the methods and results of Scientific Opinions.** This should be an easy win for EFSA as it simply requires editorial attention to the structure of Opinions and to extending the use of tables and charts. This should begin with the forthcoming hazard assessment component of the BPA Scientific Opinion currently under development.

- **Set out new requirements for presentation of comprehensive declarations of interest in Opinions.** This should be another easy win as it only requires relocating information from the expert interests database to the Opinion document, while the scope of the declaration can be adopted from a well-established convention in systematic reviews in medicine.

- **Establish a training programme in systematic review techniques for EU experts and scientific staff.** This is can be introduced as part of the existing training programme and draw on trainings EFSA has already conducted in-house.

- **Begin conducting pilot systematic reviews.** This is something which should be prioritized for support by funders of research and carried out by the research community.

4.5.3 Medium-term (1-5 years)

In the medium-term, Scientific Opinions which are planned but have not yet begun to be conducted can have protocols published for consultation. Future Opinions should be developed in a bottom-up process. Funding should also be made available for education and research in systematic review methods and facilitating the development of research networks, to accelerate the development of research capacity and community knowledge of systematic review methods.

- **Begin pre-publishing protocols for Scientific Opinions.** Although this report is directed at EFSA, this should apply to any EU expert committee which is reviewing evidence.

- **Provide funding for research into systematic review methods and establishing networks of researchers developing systematic review methods.** This should be made available by EFSA in the course of outsourcing research on systematic review techniques and also by research funders.

- **Change the process by which the scope and topic of risk assessments is decided, to make it user-led.** This will be down the Commission, while there should be a consultation on what these changes should be and how they should be implemented.

- **Institute a general educational programme for risk assessment stakeholders in systematic review.** This is necessary for developing a common understanding of what constitutes good practice in review across the whole chemical policy community.

4.6 Final comments

We have focused our conclusions on EFSA because we don't want to reach beyond the evidence we have analysed in this report. However, the situation at EFSA is likely the same for any organisation conducting a review of the available evidence – so the lessons should hold for any expert committee, any group of researchers, and even a company assembling a REACH dossier.

It is, of course, going to be very difficult to implement systematic review techniques in risk assessment. Medicine already has substantial difficulty with synthesising data from randomised controlled trials; in toxicological research we have no such thing. Here, we are dependent on epidemiology, animal models and in vitro research. Developing tools for appraising study validity and synthesising data from such diverse sources is a major challenge.

The challenge is worth it because, ultimately, the endeavour is scientific – it would be a shame if our method for reviewing studies was uneven and unpredictable, instead of being at least as methodical, objective and scientific as the studies we are reviewing. And without such methods, it is hard to see how risk assessment can ever become a truly evidence-based enterprise.

Bibliography

Ågerstrand, M.; Küster, A.; Bachmann, J.; Breitholtz, M.; Ebert, I.; Rechenberg, B.; Rudén, C. (2011a): Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals. *Environmental Pollution* 159 (10), pp. 2487–2492.

Ågerstrand, Marlene; Breitholtz, Magnus; Rudén, Christina (2011b): Comparison of four different methods for reliability evaluation of ecotoxicity data: a case study of non-standard test data used in environmental risk assessments of pharmaceutical substances. *Environ Sci Eur* 23 (1), p. 17.

Antman, E. M.; Lau, J.; Kupelnick, B.; Mosteller, F.; Chalmers, T. C. (1992): A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA* 268 (2), pp. 240–248.

Atkins, David; Best, Dana; Briss, Peter A.; Eccles, Martin; Falck-Ytter, Yngve; Flottorp, Signe et al. (2004): Grading quality of evidence and strength of recommendations. *BMJ* 328 (7454), p. 1490.

Berlin, J. A.; Rennie, D. (1999): Measuring the quality of trials: the quality of quality scales. *JAMA 282* (11), pp. 1083–1085.

Bero, L. A.; Grilli, R.; Grimshaw, J. M.; Harvey, E.; Oxman, A. D.; Thomson, M. A. (1998): Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. *BMJ* 317 (7156), pp. 465–468.

Birnbaum, Linda S.; Thayer, Kristina A.; Bucher, John R.; Wolfe, Mary S. (2013): Implementing Systematic Review at the National Toxicology Program: Status and Next Steps. *Environ. Health Perspect.* 121 *(4)*, pp. a108–a109.

Bjelakovic, Goran; Nikolova, Dimitrinka; Gluud, Lise Lotte; Simonetti, Rosa G.; Gluud, Christian (2012): Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. *Cochrane Database Syst Rev 3*, pp. CD007176.

Butler, D. (2012): EU agencies accused of conflicts of interest. European Parliament reprimands food advisory body for industry links. *Nature,* 5/15/2012.

Chalmers, Iain; Hedges, Larry V.; Cooper, Harris (2002): A brief history of research synthesis. *Eval Health Prof* 25 (1), pp. 12–37.

Cochrane Collaboration (2013): About Us. UK. Available online at http://www.cochrane.org/about-us, checked on 10/16/2013.

Conrad, James W.; Becker, Richard A. (2011): Enhancing credibility of chemical safety studies: emerging consensus on key assessment criteria. Environ. *Health Perspect*. 119 (6), pp. 757–764.

Coughlin, Deborah: JWEG Evidence Review Guidance - Beta Test Version (100713). UK Environment Agency, checked on 7/23/2013.

Deeks, J. J.; Dinnes, J.; D'Amico, R.; Sowden, A. J.; Sakarovitch, C.; Song, F. et al. (2003): Evaluating non-randomised intervention studies. *Health Technol Assess 7* (27), pp. iii-x, 1-173.

EFSA Panel on Food Contact Materials, Enzymes Flavourings and Processing Aids (CEF) (2010): Scientific Opinion on Bisphenol A: evaluation of a study investigating its neurodevelopmental toxicity, review of recent scientific literature on its toxicity and advice on the Danish risk assessment of Bisphenol A.

EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids (CEF) (2013): Draft Scientific Opinion on the risks to public health related to the presence of bisphenol A in foodstuffs-part exposure assessment. Public Consultation. European Food Safety Authority, checked on 7/26/2013.

EFSA Working Group for BPA Opinion: Full list of Scientific Articles on Bisphenol A retrieved on 22nd June 2010. European Food Safety Authority, checked on 7/23/2013.

European Chemicals Agency (2010): Practical guide 2: How to report weight of evidence.
European Food Safety Authority (2010): Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Guidance for those carrying out systematic reviews. *EFSA Journal* 8(6), p. 1637.

European Food Safety Authority (2013): About EFSA. Available online at http://www.efsa.europa.eu/en/aboutefsa.htm, checked on 29.7.13.

European Ombudsman (2011): Ombudsman: EFSA should strengthen procedures to avoid potential conflicts of interest in 'revolving door' cases.

Evans, Imogen; Thornton, Hazel; Chalmers, Iain; Glasziou, Paul (2011): Testing Treatments. Better research for better healthcare. 2nd ed. London: Pinter & Martin Ltd.

Garg, Amit X.; Hackam, Dan; Tonelli, Marcello (2008): Systematic review and meta-analysis: when one study is just not enough. *Clin J Am Soc Nephrol* 3 (1), pp. 253–260. Available online at http://cjasn.asnjournals.org/content/3/1/253.full.pdf#page=1&view=FitH.

Glasziou, Paul; Vandenbroucke, Jan; Chalmers, Iain (2004): Assessing the quality of research. *BMJ* 328 (3 January), pp. 39–41.

Goldacre, Ben (2012): Bad pharma. How drug companies mislead doctors and harm patients. London: Fourth Estate.

Gøtzsche, Peter C.; Nielsen, Margrethe (2011): Screening for breast cancer with mammography. *Cochrane Database Syst Rev* (1), pp. CD001877.

GRADE Working Group (2013): Grading the quality of evidence and the strength of recommendations.

Greenhalgh, Trisha (2010): How to read a paper. The basics of evidence-based medicine. 4th ed. Chichester, West Sussex, UK, Hoboken, NJ: Wiley-Blackwell.

Hartung, Thomas (2009): Food for thought… on evidence-based toxicology. ALTEX 26 (2), pp. 75–82.

Hemilä, Harri; Chalker, Elizabeth (2013): Vitamin C for preventing and treating the common cold. *Cochrane Database Syst Rev 1*, pp. CD000980.

Higgins, J. P. T.; Altman, D. G.; Gotzsche, P. C.; Juni, P.; Moher, D.; Oxman, A. D. et al. (2011): The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 343 (oct18 2), pp. d5928.

Higgins, Julian P. T.; Green, Sally (Eds.) (2008): Cochrane handbook for systematic reviews of interventions. Chichester, England, Hoboken, NJ: Wiley-Blackwell.

Hoffmann, S.; Hartung, T. (2006): Toward an evidence-based toxicology. *Hum Exp Toxicol* 25 (9), pp. 497–513.

Husøy, T. (2012): EFSA's approach to hazard assessment. Changes from EFSA's approach to BPA in 2010. Available online at http://www.efsa.europa.eu/en/121029/docs/121029-p04.pdf.

Jefferson, Tom; Rivetti, Alessandro; Di Pietrantonj, Carlo; Demicheli, Vittorio; Ferroni, Eliana (2012): Vaccines for preventing influenza in healthy children. *Cochrane Database Syst Rev* 8, pp. CD004879.

Jüni, P.; Witschi, A.; Bloch, R.; Egger, M. (1999): The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 282 (11), pp. 1054–1060.

Kane, R. L. (1995): Creating practice guidelines: the dangers of over-reliance on expert judgment. *J Law Med Ethics* 23 (1), pp. 62–64.

Kase, R. (2013): Klimisch 2.0 - More Transparency and Quality in Risk Assessment. Switzerland. Available online at http://www.oekotoxzentrum.ch/projekte/klimisch/index_EN, updated on 5/28/2013, checked on 10/12/2013.

Klimisch, H. J.; Andreae, M.; Tillmann, U. (1997): A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul. Toxicol. Pharmacol.* 25 (1), pp. 1–5.

Krauth, David; Woodruff, Tracey J.; Bero, Lisa (2013): Instruments for Assessing Risk of Bias and Other Methodological Criteria of Published Animal Studies: A Systematic Review. Environ Health Perspect, checked on 6/24/2013.

Liberati, Alessandro; Altman, Douglas G.; Tetzlaff, Jennifer; Mulrow, Cynthia; Gøtzsche, Peter C.; Ioannidis, John P. A. et al. (2009): The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Med* 6 (7), pp. e1000100.

Lind, J. (1753): A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the subject. Edinburgh.

Macleod, Malcolm (2011): Why animal research needs to improve. *Nature* 477 (7366), p. 511.

Maull, Elizabeth A.; Ahsan, Habibul; Edwards, Joshua; Longnecker, Matthew P.; Navas-Acien, Ana; Pi, Jingbo et al. (2012): Evaluation of the association between arsenic and diabetes: a National Toxicology Program workshop review. *Environ. Health Perspect.* 120 (12), pp. 1658–1670.

Mulrow, C. D. (1987): The medical review article: state of the science. *Ann. Intern. Med.* 106 (3), pp. 485–488.

Mulrow, C. D. (1994): Rationale for systematic reviews. BMJ 309 (6954), pp. 597–599.
Mulrow, C. D.; Cook, D. J.; Davidoff, F. (1997): Systematic reviews: critical links in the great chain of evidence. *Ann. Intern. Med.* 126 (5), pp. 389–391.

Patsopoulos, Nikolaos A.; Analatos, Apostolos A.; Ioannidis, John P A (2005): Relative citation impact of various study designs in the health sciences. *JAMA* 293 (19), pp. 2362–2366.

Rennie, Drummond; Chalmers, Iain (2009): Assessing authority. *JAMA* 301 (17), pp. 1819–1821.

Reynolds LA, Tansey EM, editors (2005): Prenatal corticosteroids for reducing morbidity and mortality after preterm birth. London: Wellcome Trust Centre for the History of Medicine

Rimer, Jane; Dwan, Kerry; Lawlor, Debbie A.; Greig, Carolyn A.; McMurdo, Marion; Morley, Wendy; Mead, Gillian E. (2012): Exercise for depression. *Cochrane Database Syst Rev* 7, pp. CD004366.

Robinson, Claire; Holland, Nina; Leloup, David; Muilerman, Hans (2013): Conflicts of interest at the European Food Safety Authority erode public confidence. *J Epidemiol Community Health* 67 (9), pp. 717–720.

SCHER; SCENIHR; SCCS (2013): Making Risk Assessment More Relevant for Risk Management. Opinion of the Scientific Committees.

Schneider, Klaus; Schwarz, Markus; Burkholder, Iris; Kopp-Schneider, Annette; Edler, Lutz; Kinsner-Ovaskainen, Agnieszka et al. (2009): "ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicol. Lett*. 189 (2), pp. 138–144.

Schulz, K. F.; Chalmers, I.; Hayes, R. J.; Altman, D. G. (1995): Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 273 (5), pp. 408–412.

Sharpe, Richard M. (2010): Is it time to end concerns over the estrogenic effects of bisphenol A? *Toxicol. Sci.* 114 (1), pp. 1–4.

The Cochrane Collaboration (2013): The Cochrane Policy Manual [updated 12 July 2013]. Available online at http://www.cochrane.org/organisational-policy-manual, checked on 7/31/2013.

The Cochrane Collaboration (2002): Open Learning Material. Assessing Quality of Studies. Using information about validity in your review. The Cochrane Collaboration. Available online at http://www.cochrane-net.org/openlearning/html/mod9-4.htm, checked on 7/1/2013.

The PLoS Medicine Editors (2011): Best Practice in Systematic Reviews: The Importance of Protocols and Registration. *PLoS Med* 8 (2), pp. e1001009.

van Maele-Fabry, Geneviève; Hoet, Perrine; Vilain, Fabienne; Lison, Dominique (2012): Occupational exposure to pesticides and Parkinson's disease: A systematic review and meta-analysis of cohort studies. E*nvironment International* 46, pp. 30–43.

Verloo, D. (2013): Systematic review (SR) in EFSA. EFSA Scientific Assessment Support Unit.

Veronesi U, Cascinelli N, Mariani L, et al (2002): Twenty-year follow up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. *New England Journal of Medicine. 3*47:1227–32

Woodruff, Tracey J.; Sutton, Patrice (2011a): An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. *Health Aff (Millwood)* 30 (5), pp. 931–937.

Woodruff, Tracey J.; Sutton, Patrice (2011b): Appendix: Navigation Guide Methodology, checked on 10/22/2012.

Woolf, S. H. (2000): Evidence-based medicine and practice guidelines: an overview. *Cancer Control 7* (4), pp. 362–367.

Young, Charles; Horton, Richard (2005): Putting clinical trials into context. *Lancet* 366 (9480), pp. 107–108.

Design: Miriam Sturdee
hello@linemaker.co.uk

This report was produced as part of the Policy from Science Project
www.policyfromscience.com