APPROVED: 8 March 2018

doi:10.2903/sp.efsa.2018.EN-1396

EFSA Scientific Colloquium 23 – Joint European Food Safety Authority and Evidence-Based Toxicology Collaboration Colloquium

Evidence integration in risk assessment: the science of combining apples and oranges

25–26 October 2017
Lisbon, Portugal

European Food Safety Authority

Abstract

In evidence-based scientific assessments, evidence synthesis is the step that occurs after collecting the data relevant to a clearly formulated research question and appraising the validity of the studies selected for the assessment, according to structured and pre-defined approaches. When studies are readily comparable, evidence synthesis is usually carried out through meta-analysis. In hazard assessment in chemical risk assessment (CRA), the process for combining evidence, 'evidence integration', is a recognised challenge as the underlying evidence bases are very diverse and not readily comparable (owing e.g. to varying degrees of validity and precision, diverse data types, different populations and species, models, end-points, routes of exposure, and evidence streams - human observational studies, experimental animal studies, in vitro and computational models data). The European Food Safety Authority (EFSA) and the Evidence-Based Toxicology Collaboration (EBTC) organised a Colloquium to develop a multistakeholder understanding of the best practices, challenges and research needs for evidence integration in CRA, with a focus on hazard identification and on combining multiple studies and end-points for dose–response modelling. The methods discussed included: qualitative methods for integrating evidence within- and across evidence streams; bias-adjusted meta-analysis; quantitative approaches to combine evidence across evidence streams; and quantitative approaches for combining multiple end-points and multiple studies for dose–response modelling. All these methods showed advantages and needs for further development, testing, validation and effective implementation. Support to this could be provided by: more published primary toxicological and epidemiological data; optimisation of study design; a shared primary data repository; the establishment of a community of knowledge of toxicologists, epidemiologists and statisticians. Equally, to be conducted soundly, evidence integration in CRA should be undertaken by multidisciplinary groups (toxicologists and methodologists knowledgeable of the various integration techniques). EFSA and EBTC will continue the collaboration towards the development, testing and validation of best practices for evidence-based CRA.

Key words: Evidence synthesis, evidence integration, evidence-based, chemical risk assessment, toxicology, environmental health, hazard assessment, dose-response modelling

Question number: EFSA-Q-2017-00712

Correspondence: any enquires related to this output should be addressed to scientific.colloquia@efsa.europa.eu

Disclaimer: The views or positions expressed in this publication do not necessarily represent in legal terms the official position of the European Food Safety Authority (EFSA). EFSA assumes no responsibility or liability for any errors or inaccuracies that may appear.

Acknowledgements: EFSA wishes to thank Marc Aerts, Sofia Dias, Julian Higgins, Donald Rubin, Holger Schünemann, Wout Slob, Kristina Thayer, Stijn Vansteelandt and Matthew Wheeler, for their participation in the Colloquium as lecturers and/or discussion groups' chairs and for contributing to drafting this report. EFSA also wishes to thank its staff members Elisa Aiassa, Fulvio Barizzone, José Cortinas Abrahantes, Laura Martino and EBTC members Paul Whaley, Daniele Wikoff, Sebastian Hoffman and Rob de Vries for the preparatory work as the Organising Committee, as well as Katya Tsaioun (EBTC) and Didier Verloo (EFSA), for acting as overall chairs.

Suggested citation: EFSA (European Food Safety Authority) and EBTC (Evidence-Based Toxicology Collaboration) 2018. EFSA Scientific Colloquium 23: Evidence integration in risk assessment: the science of combining apples and oranges. EFSA Supporting publication 2018:16(3):EN-1396. 28 pp. doi:10.2903/sp.efsa.2018.EN-1396

ISSN: 2397-8325

Table of contents

Abstract..1
1. Introduction..4
2. EFSA and EBTC..4
3. Objectives of the Colloquium..5
4. Participants and format..5
5. Abstracts of speakers in opening plenary sessions..6
5.1. Lecture 1: A quantitative framework for evidence integration..6
5.2. Lecture 2: Integrating evidence within and across evidence streams using qualitative methods6
5.3. Lecture 3: Recent developments for combining evidence within evidence streams: bias-adjusted meta-analysis..7
5.4. Lecture 4: Quantitative approaches to combining evidence across evidence streams..7
5.5. Lecture 5: Introduction to benchmark dose estimation from multiple end-points and multiple studies: current practices and challenges..8
5.6. Lecture 6: Combining evidence on multiple end-points in dose–response assessments: multivariate models..8
5.7. Lecture 7: Other quantitative methods for combining multiple studies and end-points..8
6. Summary of discussion groups..9
6.1. DG1 – Qualitative methods for integrating evidence within and across evidence streams for hazard identification..9
6.1.1. DG1 background and introduction..9
6.1.2. Biological plausibility..10
6.1.3. Consistency..11
6.1.4. Sensitivity..11
6.1.5. DG1 Conclusions and future developments in the field..12
6.2. DG2 – Bias-adjusted meta-analysis..12
6.2.1. DG2 background and introduction..12
6.2.2. Going beyond traditional meta-analytic approaches: should we adjust for bias?..12
6.2.3. Advantages and limitations of currently available bias-adjustment methods..13
6.2.4. What are the requirements for additional evidence to inform bias-adjustment?..14
6.2.5. What are the additional skills required to implement bias-adjustment methods?..14
6.2.6. DG2 Conclusions and future developments in the field..14
6.3. DG3 – Quantitative approaches to combining evidence across evidence streams for hazard identification..15
6.3.1. DG3 background and introduction..15
6.3.2. What are the current practices to integrate heterogeneous evidence on hazard identification?15
6.3.3. What do we want to quantify in the context of integrating evidence for hazard identification?15
6.3.4. What are the current challenges in applying quantitative approaches in hazard identification?16
6.3.5. What are the proposed actions for the future?..16
6.4. DG4 – Using multiple end-points and multiple studies for dose–response modelling: quantitative approaches..17
6.4.1. DG4 background and introduction..17
6.4.2. Is there a way to better share information between statisticians and toxicologists?..17
6.4.3. Can a community of knowledge with available toxicological and statistical expertise be built? 18
6.4.4. What are the implications of empirical evidence shown?..18
6.4.5. Could all be looked from the risk scale viewpoint when combining end-points, going from continuous to quantal data pros and cons?..19
7. Overall conclusions and way forward..19
References..21
Appendix A – List of participants..24
Appendix B - Colloquium Programme..27

1. Introduction

Evidence-based scientific assessments involve applying structured and standardised approaches to minimise bias and random error and ensure transparency in the process for collecting, evaluating and combining evidence relevant to well formulated research questions, according to pre-defined protocols. These approaches are well established for healthcare intervention questions and their value has also been extensively acknowledged in chemical risk assessment, for which their application continues to be actively developing (Hoffmann and Hartung, 2006; Hartung, 2009; Stephens et al., 2013).

In evidence-based scientific assessments, evidence synthesis is the step that occurs after appraising the validity of the individual studies selected for the assessment. In evaluations of the efficacy of therapeutic interventions, this step is usually carried out through a meta-analysis, which encompasses statistical methods for combining data from similar, readily comparable studies.

In hazard identification and characterisation for chemical risk assessment, the underlying evidence bases are diverse and not readily comparable. Unlike in medicine, in this research field heterogeneity of evidence stems not only from varying degrees of validity and precision of studies and diverse data types (e.g. individual vs aggregated), but also from different populations and species, models, end-points, routes of exposure, and diverse evidence streams (human observational studies, experimental animal studies, *in vitro* and computational models data). As such, a process for combining evidence not only within – but also across – evidence streams is needed. This process is defined as 'evidence integration' and is particularly relevant for assessing the effects of exposure to a chemical substance (hazard identification), and for deriving health-based guidance values through dose–response modelling (hazard characterisation).

Evidence integration is a recognised challenge in evidence-based risk assessment for which different methods exist, ranging from approaches based on expert judgement, through structured qualitative methods, to complex quantitative methods.

The European Food Safety Authority and the Evidence-Based Toxicology Collaboration (Zurlo, 2011) housed at the Johns Hopkins Bloomberg School of Public Health (EBTC)1 organised a Colloquium to discuss the current state of the art of these methodological aspects, to bring together experts in the field, and to start addressing these challenges. The event, which took place in Lisbon on 25–26 October 2017, was the 23rd in the EFSA Colloquium series and the first to be jointly organised by EFSA and EBTC.

2. EFSA and EBTC

EFSA activity is focused on performing evidence-based scientific assessments in the field of food and feed safety. A central role is played by the development of guidance to support the implementation of sound methodology for using evidence. In recent years, a founding document was published that addressed the process for dealing with evidence and its guiding principles (EFSA, 2015 – first deliverable of EFSA PROMETHEUS project[2]) along with a set of horizontal guidance developed by the EFSA Scientific Committee, focusing on approaches for integrating the evidence (EFSA Scientific Committee et al., 2017a), while accounting for the uncertainty inherent in the data and the process (EFSA Scientific Committee et al., 2018) and properly considering the biological relevance of evidence and effects (EFSA Scientific Committee et al., 2017b).

Because of the common goal of bringing evidence-based methods into the field of toxicology and environmental health scientific assessments, EFSA and EBTC have recently started sharing views and scientific activities.

EBTC is a collaboration between academic, government, non-governmental and industry leaders located at the Johns Hopkins Bloomberg School of Public Health. A EBTC objective is to bring together stakeholders involved in safety assessments (governmental agencies – so far EFSA, the United States Environmental Protection Agency (US EPA), the United States Food and Drug Administration (US

[1] http://www.ebtox.org/

[2] http://www.efsa.europa.eu/en/methodology/evidence

FDA), the scientific community, industry and public representatives) to set out and facilitate wide acceptance and implementation of the new safety assessment paradigms.

EFSA acknowledges and supports the value that EBTC brings to the community of safety assessors and has supported this by participation in the EBTC Board of Trustees since 2016.

With this Colloquium, EBTC and EFSA started a new phase of their collaboration, working closer to bring together all stakeholders to work on concrete methodological challenges, form smaller groups focused on solving them, testing new tools and methods and bringing them to the community by publishing detailed documents such as guidance documents, by organising workshops and similar events that educate the community and by unify the methods of safety testing across different areas.

3. Objectives of the Colloquium

The aim of the Colloquium was to develop a multistakeholder understanding of the best practices, challenges and research needs for evidence integration in human risk assessment of chemicals, with a specific focus on hazard identification and on combining multiple studies and end-points for dose–response modelling in hazard characterisation.

Despite the focus on chemicals, the objective was to address these methodological aspects from a broad, cross-cutting point of view that is relevant to other research contexts (e.g. dietary reference values).

4. Participants and format

Eighty-one participants attended the Colloquium from 15 European countries, Canada, Qatar, Tunisia and the USA. They included EFSA staff and external experts from EFSA panels and working groups, 2 EBTC Board members, 4 EBTC staff, and representatives from 16 national authorities and 23 universities/research institutes. Representatives of international organisations, NGOs and private sector organisations also took part. The list of participants is available in Appendix 1.

The event consisted of an opening session with introductory keynote speeches, followed by a breakout groups session and a final plenary discussion (Appendix 2).

The breakout session was structured to engage small groups of participants in focused discussions on the topics introduced by the lecturers in the opening session, and in particular:

- Discussion Group 1 (DG1) explored qualitative methods for integrating evidence within- and across evidence streams for hazard identification.
- Discussion Group 2 (DG2) focused on bias-adjusted meta-analysis.
- Discussion Group 3 (DG3) looked at the possibility to apply, in the future, quantitative approaches to combine evidence across evidence streams for hazard identification.
- Discussion Group 4 (DG4) discussed the use of quantitative approaches for combining multiple end-points and multiple studies for dose–response modelling. It also focused on the information and infrastructure needed to assess the differences between the approaches discussed and on the methods/models that could be most useful when new sources of information are available (*in vitro* studies, 'omics type data, etc.).

In the final plenary session, the outcomes of the groups were presented and discussed: to draft the conclusions of the Colloquium and, as appropriate, the recommendations for next steps in addressing each challenge.

5. Abstracts of speakers in opening plenary sessions

5.1. Lecture 1: A quantitative framework for evidence integration

Donald Rubin/ *Harvard University | USA*

When trying to assess the causal effects on humans of various exposures or substances, such as air pollutants, vaccines or pharmaceuticals, it is common to rely on disparate sources of evidence because conducting randomised controlled trials (RCTs) on humans is either considered unethical or logistically too complex. The various sources can vary from: RCTs on animals (i.e. *in vivo*) or *in vitro* laboratory studies, observational (epidemiological) data on humans, sometimes RCTs on volunteers, etc. To combine these evidence streams into a coherent story about the benefit–risk trade-off to humans is certainly challenging, and is sometimes attacked from a non-quantitative perspective, relying on informal assessments of the evidence streams. This presentation argued for trying to make a quantitative assessment by taking a response surface perspective, which takes as input the descriptors of the various studies (their design features, e.g. Z; characteristics of the units being studied, e.g. X; and exposures under investigation, e.g. W) and attempts to model the outputs, e.g. Y, which are the causal effects of the various exposures; i.e. to create a mathematical relationship to estimate Y as a function of X, Z and W. The next step is to extrapolate this function to the limit at which the value of Z represents the perfect RCT with no unintended complications and the value of X indicates humans with various characteristics such as race, age, sex, etc. This framework was proposed by myself in a chapter entitled 'A New Perspective' (Wachter and Straf, 1990); also see its book review in JASA by Gene Glass (Glass, 1991), who coined the term in 1976. Trying for such formality often helps to clarify issues by revealing points of agreement or disagreement. Also embedding the entire enterprise within a Bayesian decision-theory framework can be similarly revealing.

5.2. Lecture 2: Integrating evidence within and across evidence streams using qualitative methods

Kristina Thayer / Environmental Protection Agency (EPA), Integrated Risk Information System (*IRIS*) Division / USA

There is high demand in environmental health for adoption of a structured process that evaluates and integrates evidence while making decisions transparent. The Grading of Recommendations Assessment, Development and Evaluation (GRADE)[3] framework holds promise to address this demand. For 17 years, GRADE has been applied successfully to areas of clinical medicine, public health and health policy, but experience with GRADE in environmental and occupational health is limited. Environmental and occupational health questions focus on understanding whether an exposure is a potential health hazard or risk, assessing the exposure to understand the extent and magnitude of risk, and exploring interventions to mitigate exposure or risk. Although GRADE offers many advantages, including its wide use (over 100+ organisations) and methodological rigour, there are features of the different sources of evidence used in environmental and occupational health that will require further consideration to assess the need for method refinement. An issue that requires particular attention is the evaluation and integration of evidence from human, animal, *in vitro*, and *in silico* (computer modelling) studies when determining whether an environmental factor represents a potential health hazard or risk. The objectives of this presentation were to provide an overview of how the GRADE framework overlaps with considerations (reliability, relevance and consistency) used by EFSA in its guidance on Weight of Evidence[4] and others, to identify priority areas for method assessment and development, and to discuss experience to date in applying GRADE to environmental health topics.

[3] http://www.gradeworkinggroup.org/

[4] http://www.efsa.europa.eu/it/efsajournal/pub/4971

5.3. Lecture 3: Recent developments for combining evidence within evidence streams: bias-adjusted meta-analysis

Julian Higgins / *University of Bristol / UK*

The lecture described approaches available for adjusting bias in the results of primary research studies when undertaking a statistical evidence synthesis using summary (aggregate) data. It began with a review of the approaches to assessing the risk that there is bias in a study result, including tools such as the recent ROBINS-I tool, the forthcoming ROBINS-E tool, and the Office of Health Assessment and Translation (OHAT) tool. In the past, a large proportion of evidence syntheses have made very little quantitative use of the results of these types of assessments and have, at best, commented on the limitations of the studies alongside presentation of the results of the evidence synthesis. The lecture provided an overview of the range of alternative strategies. These include: (i) stratification of studies, e.g. as part of a sensitivity analysis; (ii) incorporation of quality assessments or risk-of-bias assessments into statistical weights; (iii) meta-regression approaches that investigate the dependence of study results on study features or risk-of-bias assessments; (iv) direct 'corrections' for bias; (v) use of prior distributions for the extent of bias in a Bayesian framework; and (vi) triangulation approaches, in which bias can be both estimated and accounted for within a body of evidence.

5.4. Lecture 4: Quantitative approaches to combining evidence across evidence streams

Stijn Vansteelandt / *University of Ghent / BE and London School of Hygiene and Tropical Medicine / UK*

Standard methods for evidence synthesis combine estimates (e.g. log odds ratios) obtained from different populations; in extreme cases, this could be a combination of populations of humans as well as populations of animals. Two major, widely ignored concerns are: (a) that the summary effect obtained via such methods lacks interpretation, as it is unclear for which population the effect is described; and (b) that standard methods for evidence synthesis generally ignore the lack of similarity of baseline characteristics in the different study populations, which is nonetheless key for successful pooling of results from different populations. In this talk, these concerns were circumvented by using the results of each study to infer the adverse effect of an exposure (e.g. trihalomethanes) in a single, clearly defined target population (e.g. the population of people observed in one of the considered studies), so making pooling results from different studies possible. This was achieved using direct standardisation. In particular, the data from each of the separate studies were used to build a (separate) prediction model for the risk of adverse events in function of the exposure and observed baseline characteristics. These models enabled extrapolation of the results from each study to the people observed in the considered target study, by predicting what their risk of adverse events would be with and without exposure. By averaging these predictions, estimates for the risk of adverse events (with or without exposure) in the target study were obtained. These estimates were subsequently pooled across studies.

The proposed formalism made it clear that evidence synthesis involves extrapolation of the results from one study to another; the extent to which this can be successfully carried out depends on the similarity between the study populations in the different studies. Our formalism gave insight into the assumptions required to enable extrapolation of the results of a given study to a specific population. It, moreover, made it clear that the danger of extreme extrapolation can make evidence synthesis non-trivial when the considered studies include very different populations, e.g. when pooling results from one study in humans aged 20 to 30 years and another study in humans aged 20 to 60 years, or even more so when pooling the results from animal and human studies.

5.5. Lecture 5: Introduction to benchmark dose estimation from multiple end-points and multiple studies: current practices and challenges

Marc Aerts / Hasselt University / *BE*

In this introductory presentation, an overview was given of statistical methods and models, relevant for benchmark dose (BMD) estimation based on data from multiple end-points and/or from multiple studies. After introducing some illustrative examples, the guidelines regarding the current practice for a single end-point and a single study and their extension to multiple settings, as provided by the Update: EFSA Scientific Committee Guidance on the use of the benchmark dose approach in risk assessment (EFSA Scientific Committee et al., 2017c) and by US EPA Benchmark Dose Technical Guidance (US EPA, 2012), were briefly discussed. Next, statistical methods and models for dealing with multiple end-points were introduced, from simple and easy pragmatic approaches up to more advanced techniques, while describing initial pros and cons in the context of BMD estimation. Then, the extension to multiple studies and its combination with multiple end-points was briefly discussed. The presentation ended with a discussion on the major future challenges when applying these statistical methods and models and when shaping future guidance for researchers.

5.6. Lecture 6: Combining evidence on multiple end-points in dose–response assessments: multivariate models

Wout Slob */ RIVM / The Netherlands*

Toxicity studies usually involve multiple end-points. Risk assessors generally select the end-point that resulted in the lowest point of departure (PoD) as the most sensitive and hence critical end-point (for that study). However, PoDs are subject to uncertainties, and these uncertainties may differ among end-points. By calculating the BMD confidence intervals for all the end-points, this is made visible: the benchmark dose level (BMDL) for a particular end-point might be relatively low due to relatively large uncertainties in the data while, in reality, the end-point is not more sensitive than the other points. Furthermore, for continuous end-points, it may not be appropriate to use the same value for the benchmark response (BMR) (5% change in mean response, the default in EFSA guidance) for all end-points. A recent theory on effect size (Slob, 2017) showed that it would be more appropriate to scale the BMR to the maximum response or, as a surrogate, to the within-group variance. By doing so, the BMD confidence intervals for the various end-points in the same study tend to get much more similar. This raises the hypothesis that the BMD is in fact the same for all end-points that are affected in the same study. If this hypothesis can be further validated, one may derive one single BMD confidence interval (and one single PoD) for the whole study by using multivariate statistical techniques. Furthermore, various studies have shown that the observed differences in PoDs related to different species are of the same order of magnitude as study replication errors. This raises the hypothesis that species, in reality, hardly differ in sensitivity in the context of BMD estimation. Hypotheses such as these need to be further investigated in this specific field (BMD estimation from dose-response data), as it is important to know which factors do indeed have an impact on the potency of chemicals, and which factors do not. This knowledge is paramount in deciding how to deal with the multiple studies available for a particular chemical, as illustrated by a simple example.

5.7. Lecture 7: Other quantitative methods for combining multiple studies and end-points

Matthew Wheeler/*National Institute for Occupational Safety and Health (NIOSH), CDC/ USA*

Current risk assessment practice focuses on finding a critical effect end-point and estimating a PoD from a dose–response model for this effect. From the perspecticve of controlling risk, this practice has many problems that may not be health protective. Further, additional information on the same end-point may be available from other sources, which leads to the question 'how do we integrate all information to make more informed decisions?' To investigate this, we studied situations and problems encountered when attempting to amalgamate available information in a risk assessment. This talk presented a series of case studies trying to look at the varied situations in which this may occur. It covered routine cases in which one may have multiple end-points/studies for dose–response analysis to hypothetical models that relate *in vitro* high throughput data to an *in vivo* response. We examined data needs, statistical methodologies and current knowledge gaps. Rather than being a cookbook of recipes, the talk was designed to spur discussion among attendees about the possibilities, pitfalls and ways forward when integrating evidence from toxicity studies.

6. Summary of discussion groups

6.1. DG1 – Qualitative methods for integrating evidence within and across evidence streams for hazard identification

Chair: Holger Schünemann (McMaster University, Canada)

Rapporteurs from the organising committee: Paul Whaley (Lancaster University, UK) and Daniele Wikoff (ToxStrategies, Inc., USA)

Follow-up of Lecture 2 – Integrating evidence within and across evidence streams using qualitative methods. Kristina Thayer/*Environmental Protection Agency (EPA), Integrated Risk Information System (IRIS) Division|USA*

6.1.1. DG1 background and introduction

Although multiple methods have been developed for evidence integration, the GRADE methodology (Schünemann et al., 2003; Atkins et al., 2004) for assessing certainty in the body of evidence is the most well recognised and widely applied. The GRADE approach is generally qualitative but encourages quantitative judgements if they are well founded (Guyatt et al., 2017). Depending on the tool used to assess risk of bias, GRADE either requests risk of bias to be considered by rating down observational studies initially or use a rigorous tool such as ROBINS-I (Sterne et al., 2016) that compares the risk of bias in studies to randomised experiments. Further domains for downgrading or upgrading this rating based on predefined strengths and limitations of the overall evidence base are then applied. Specifically, the GRADE domains for downgrading confidence include: detailed risk of bias, inconsistency, indirectness, imprecision and publication bias. Other criteria, such as large effect, dose–response and plausible confounding can increase the confidence ratings.

Although many aspects of the ‘GRADE approach’ to evidence integration work well for hazard assessment, they need to be further explored for the field of environmental health (NAS, 2014; Morgan et al., 2016). Integration of heterogeneous data across evidence streams is one area that requires additional consideration for the application of GRADE in environmental health.

As such, the broad objective of this discussion group was to discuss the GRADE certainty of evidence framework and ways to put into operation consideration of evidence across streams within this framework. The following priorities were identified:

1) Is GRADE sufficient?, i.e. do the GRADE certainty-of-evidence domains consider all the factors that determine certainty about the presence of a hazard, association or effect?
2) Does GRADE satisfactorily address how different streams of evidence should be combined in the development of conclusions? If not, how do we preserve evidence-based principles in a rich integration process?
3) How do you best combine (‘integrate’) different evidence streams for hazard identification?
4) Are there other processes that should be included in the evidence integration process, and if so what are they?
5) How are the ratings for certainty integrated with the results of an evidence synthesis to develop/systematise conclusions? How can the ratings be used to evaluate contradictory data?

Recognising that well established methods for evidence integration in the fields of clinical medicine and nutrition (such as GRADE) overlap, but do not necessarily run parallel to those traditionally used in environmental health, the discussion focused on three concepts that are commonly identified by the toxicology community as important for reaching weight of evidence causality conclusions. These are:

(1) biological plausibility; (2) consistency between and across species/study types; and (3) sensitivity, or the ability of the study to detect the potential effect in question. As the environmental health community is beginning to utilise systematic review frameworks and integration techniques (such as GRADE), questions have been raised on how these concepts are considered within GRADE and whether they may be 'missing' or in need of refinement relative to their application in toxicology and risk assessment (Durrheim and Reingold, 2010; Hoffmann et al., 2017). GRADE provides guidance on how to put into operation the Hill considerations (Schünemann et al., 2011) but this may not be readily apparent without detailed familiarity and hands on experience with the GRADE framework. One challenge relates to differences in terminology used within GRADE to terminology used by the environmental health community, i.e. the same concept may be called something else in different scientific disciplines.

Before discussing the three principles, the group discussed reliance on GRADE methodology as the construct of focus. It was agreed that discussion could have been focused on other approaches (e.g. Table B.3 in EFSA Scientific Committee et al., 2017a). However, the group focused on GRADE because it is the most widely used framework in systematic reviews for chemical hazard assessment, for example having been implemented by NTP/OHAT (Rooney et al., 2014) and The Navigation Guide (Woodruff and Sutton, 2014) and proposed for use in endocrine disruptor identification (Vandenberg et al., 2016). GRADE seemingly covers aspects outlined by EFSA as important to WoE (Table B.3 in EFSA Scientific Committee et al., 2017a) and the GRADE Working Group is dedicated to providing guidance to help put into operation the application of WoE concepts that can promote reproducibility in the field (Morgan et al., 2016). In addition, feedback from the EFSA meeting on applicability of GRADE to environmental health would be considered by the GRADE Working Group as it continues to refine methods in this area.

The 25 participants in DG1 had diverse interests, many with backgrounds in risk assessment and food safety, representing government, non-governmental organisations and the private sector. Approximately half of the members of the discussion group were EBTC members or EFSA staff. In an effort to generate productive discussion, participants were provided with background information before the workshop. These included briefing notes, key publications that provided both a foundation for systematic review processes as implemented in environmental health evaluations, as well as foundation on how GRADE puts into operation the assessment of certainty of evidence for environmental health evaluations. Onsite, participants were provided with a worksheet that included descriptions of key terms and concepts that HS, KT, PW and DW developed to assist in facilitation.

Working in small group format, participants were tasked with defining and putting into operation the use of each of the principles of biological plausibility, consistency, and sensitivity in the context of evaluating the credibility of findings of an evidence synthesis (such as a hazard assessment). Each small group reported their initial conclusions in the form of a definition for each key concept. The relative merits of each definition were then discussed in the large group.

6.1.2. Biological plausibility

Although used repeatedly in chemical risk assessments, the consensus view of the DG was that the concept of 'biological plausibility' is not consistently or clearly defined. In environmental health the concept of 'biological plausibility' usually refers to consistency between data and biological theory or mechanism (EFSA Scientific Committee et al., 2017b), which best map to the Bradford Hill concepts of 'plausibility' and 'coherence' (Bradford Hill, 1965). However, precisely what puts into operation a judgement of biological plausibility is seldom made clear.

When attempting to define 'biological plausibility' and provide signalling questions on how to put it into operation, the small groups initially proposed a range of concepts that were, in fact, covered in current GRADE domains, including a strong study design (low risk of bias), consistency in findings between studies, strong association between exposure and effect, relevance (directness) of the data to the outcome of interest, and whether the observed association is plausible given current biological understanding, i.e. aspects of GRADE considerations for directness and consistency. What this discussion revealed was that while many facets of research conduct and results bear on biological plausibility, it is not the case that biological plausibility is itself an independent domain that affects certainty in the evidence. Rather, the extent to which a purported association between exposure and outcome is biologically plausible is a result of the evidence synthesis process overall.

As part of this discussion, participants were challenged to develop or cite examples in which aspects of biological plausibility would not be covered by the GRADE domains, in particular for strong association, indirectness, risk of bias and consistency. Situations put forward by the group included suggestions from epidemiological studies that high heels might be associated with breast cancer (there is a detectable association between the wearing of high heels and risk of breast cancer in women, but the absence of a plausible biological mechanism suggests the association is spurious) or pancreatic cancer being caused by tea consumption (risk of pancreatic cancer is associated with consumption of tea but the explanation is that early pancreatic cancer causes diabetic symptoms, which increases thirst and therefore increases tea consumption – a case of reverse causation). On discussion, however, the group reasoned that in these cases, the biological plausibility of these scenarios was accounted for either by the systematic review process or the GRADE framework, and therefore did not constitute an additional consideration. Regarding the spurious association, this issue would likely be addressed at the level of problem formulation and question development, whereas reverse causation would have been handled by GRADE via consideration of confounding under both risk of bias assessment and assessment of all plausible confounding.

So, the group appeared to converge on a conclusion that a determination of biological plausibility (high certainty of the evidence indicates high likelihood of biological plausibility) is derived from the results of the evidence synthesis, and is not a missing element from GRADE. That said, additional discussion is needed on how to put into operation consideration of mechanistic evidence in the context of biological plausibility. While mechanistic understanding is used to develop hypotheses for exploration via evidence synthesis and provides indirect evidence, it should not be required to test the results of a synthesis for credibility.

6.1.3. Consistency

Consistency (or inconsistency) of a body of evidence (across different research studies) is an element of GRADE, however it was recognised that there were features of consistency related to toxicological evidence that may need additional clarification. In GRADE, inconsistency refers to (unexplained) heterogeneity of study results; Bradford Hill describes consistent findings by different people and coherence between epidemiological and laboratory findings as increasing the likelihood of causality*.* With respect to toxicological evidence, the discussion group described consistency with respect to different streams of evidence and within streams of evidence. 'Consistency' was initially defined by the small groups as observation of the same pattern of response to an agent across animal and epidemiological evidence, species, different study designs within species, and different methods for measuring the same outcome.

Criteria for operationalising the assessment of consistency included identifying biologically plausible explanations for observed consistency based on mode-of-action arguments; presence of repeatable results from individual studies; and overlap of confidence intervals and direction of effect. Discussion of the criteria in the large group determined that all criteria proposed for determining consistency are already covered by GRADE: consistency across findings contributes to biological plausibility. Consensus in the larger group was that consistency, as understood by the group, was not a new concept which needs adding to the GRADE framework.

6.1.4. Sensitivity

Sensitivity was an element highlighted by the discussion group due to use of the term by the US EPA and EFSA. The US EPA refers to the 'sensitivity' of a study in terms of ability to detect the potential effect in question (Cooper et al., 2016). EFSA refers to sensitivity as one of several factors that need to be considered when assessing the reliability of a piece of evidence (EFSA Scientific Committee et al., 2017a) broadly as to whether the studies in question can detect the effect of interest at the concentration of concern. The small groups thought it would be useful for the GRADE Working Group to consider additional guidance on how sensitivity is considered in the framework.

Although not completely resolved, some options were suggested. For example, sensitivity could be considered in the context of relevance to the study question via the PECO (Population/Exposure/Comparison/Outcome) format (i.e. to what extent are the studies in question of appropriate design for measuring the effect in question?). In cases in which screening guidance for sensitivity cannot be articulated *a priori* as part of the PECO, then insensitive methods or model systems may be considered as part of directness, i.e. insensitive methods would be considered less direct. Another option is to consider study sensitivity as related to precision (i.e. whether a study design is sufficiently powered to detect an effect, or if study design parameters will result in the study being unable to generate confidence intervals that will not include null).

6.1.5. DG1 Conclusions and future developments in the field

While consistency, sensitivity and biological plausibility were agreed by the group to be covered by GRADE, there is a general need to provide more guidance, clear definitions, and explicitly address how the GRADE criteria are operationalised in the context of environmental health research and regulatory risk assessment. This is particularly true for biological plausibility, because there is a strong, intuitive sense that it is something additional to the GRADE framework. How biological plausibility maps onto GRADE therefore needs clear and explicit articulation. Experience through practical application and case studies will also be critical to advancing discussions and identifying refinements to evidence integration methodologies unique to chemical risk assessment.

Additionally, while not specified at the outset as discussion topics, several themes appeared that are not related to GRADE and might warrant clarification in the future. This included distinguishing between study quality (including risk of bias) and reporting of quality (how well are relevant items, including risk-of-bias items, reported in a study).

6.2. DG2 – Bias-adjusted meta-analysis

Chair: Sofia Dias (University of Bristol, UK)

Rapporteurs from the organising committee: Fulvio Barizzone and Elisa Aiassa (EFSA)

Follow-up of Lecture 3 – Recent developments for combining evidence within evidence streams: bias-adjusted meta-analysis. Julian Higgins/*University of Bristol/UK*

6.2.1. DG2 background and introduction

Evidence appraisal typically involves assessment of the internal validity or risk of bias (RoB) of each individual study. This appraisal is usually carried out using appraisal tools aimed at minimising subjectivity and increasing consistency and transparency in the process.

While the available tools do help to identify threats to validity of the results, they provide little or no guidance on how to assess the impact of threats to validity on the study results. For instance, most tools do not address the direction and magnitude of the internal biases identified within the tools. Studies are usually grouped according to different RoB categories (e.g. high, some concerns, low) and typically the result of study appraisal is addressed through sensitivity analyses or exploratory subgroup analyses.

It is also important to assess the relevance of study results to the research question at hand, for example in terms of populations studied and exposures measured. Differences in results that arise from these factors might be regarded as external biases. Again, subgroup analyses are often used to address these, but these tend to separate out the evidence rather than to integrate it.

Methods are available to synthesise evidence while accounting for internal and external biases and for the uncertainty about them. This approach is generally known as bias-adjusted meta-analysis, although it is rarely used in practice. Information on the biases may come from empirical evidence from an external collection of meta-analyses (Welton et al., 2009), expert knowledge elicitation (Turner et al., 2009) or a combination of the two (MRC Centre Cambridge 2017).

Before the Colloquium, the participants of this discussion group were provided with briefing notes and some reading material, to stimulate discussion on the points illustrated in the next sections.

6.2.2. Going beyond traditional meta-analytic approaches: should we adjust for bias?

A recap of the methods presented in the plenary lecture on bias adjustment was presented and the different approaches for bias adjustment discussed. Through a voting system aimed at encouraging participation, the group members agreed that there are advantages to performing bias-adjusted analyses following a qualitative assessment of RoB of included studies.

A bias-adjusted meta-analysis will produce an effect estimate that is adjusted for bias while also incorporating any additional uncertainty. This effect estimate can be used for better decision-making.

6.2.3. Advantages and limitations of currently available bias-adjustment methods

The advantages and limitations of several bias-adjustment methods were discussed.

Quality effect model

The quality effects model (Doi et al., 2015) uses weighting to adjust for study quality defined by a scale. Its main strength is that it can be used even with a limited number of studies and that no extra time would be required to collect extra data or elicit opinions. The adjustment process, being based on weights, does not require any measure of the direction or magnitude of bias that might actually be caused by the study features of concern. This is because it weights studies based on their quality-derived (i.e. relative to each other). It however does require a comprehensive tool for appraising the studies to be available, for which all factors are then assumed to have the same impact on the potential bias.

Regression-based methods

Regression-based methods for bias adjustment require enough studies to be available to estimate the regression coefficient associated with each bias factor. In practice only a limited number of studies may be available, making it impossible to use this method. In addition, it also requires defining relevant bias factors that should be included as covariates and extrapolation to the 'perfect' study, which may be hard to define in this context. This extrapolation may also be problematic if not many studies are of high quality, as this would require extrapolating very far beyond the available quality data. However, it is a simple method to apply, requiring only simple regression software that provides an informative adjusted effect size that does not need to rely on a single quality score.

Direct 'corrections' for bias

Methods for correcting each individual study for bias, before inclusion in the synthesis were thought to form part of an ideal practice and good statistical modelling principles, however they require tailoring to each situation and are, by definition, context dependent. They can also be extremely time consuming. There is also a requirement for information on which factors will contribute to bias and in which direction that bias will act.

The use of empirically based prior distributions to adjust each study for bias, while being a potentially reasonable and straight-forward approach, requires empirical data on the impact of different bias domains to be available, and particularly for the impact of multiple bias domains simultaneously. This is not yet available and would require a large research effort to collate. However, once the empirical evidence is available, it can be easily incorporated in multiple future analyses within a field.

The possibility of asking experts to produce an estimate of the direction and magnitude of bias for each study, which could then be used to correct the study effect before including in the analysis, was also discussed. This approach is explicit and transparent and shows exactly how each study was adjusted and its impact on the analysis. However, this approach is resource intensive as elicitation of bias from experts would take time and may require group elicitation methods to ensure both topic content and methodological skills are captured when expressing the beliefs about bias. There was also a concern that reproducibility of results would be limited, as even the same experts might suggest different adjustments at a different time. Psychological biases were also thought to pose a problem, although these are prevalent in all decision-making processes and methods are available to minimise their impact. Some members of the group pointed out that bias elicitation might be impossible because the absolute magnitude and direction of a bias induced by a quality deficiency in a particular study cannot be empirically confirmed. Nevertheless, examples of this approach do exist in the literature.

6.2.4. What are the requirements for additional evidence to inform bias-adjustment?

When using the quality-effects model, meta-epidemiological studies would ideally be available to inform the ranking of bias domains and components, since an assumption that each source of bias has equal impact is unlikely to hold. When using prior distributions for bias, there may be different requirements for external evidence, as discussed in the previous point. The key issue overall is to identify sources of evidence on which study features impact on bias and to what extent. The group agreed that this evidence already exists for some disciplines and specific problems, but has not yet been systematically explored, collected or analysed. Therefore, the evidence on bias domains and their impact required to carry out the adjustment is not yet available. The group also agreed that any time spent on bias adjustment (collecting evidence, implementing methods etc.) should be proportionate to the impact the potential adjustment is expected to have on the decision, but this is hard to quantify at the moment.

6.2.5. What are the additional skills required to implement bias-adjustment methods?

The group expressed the opinion that evidence integration teams should include expertise in both evidence appraisal and topic content and would therefore be able to inform assumptions about the bias direction and magnitude, if a bias elicitation process was used, or would be able to categorise studies as at risk or not at RoB, or according to any agreed scale. Therefore, given the methods available, it should be possible to implement bias-adjustment provided agreement could be reached on which are the most relevant bias domains and which quality scales or elicitation processes should be used.

However, some group members expressed concern that risk-of-bias assessment is not structured enough at the moment and remains very subjective, so further guidance on critical appraisal of studies might be needed.

6.2.6. DG2 Conclusions and future developments in the field

The group agreed that the main barriers for implementation of bias-adjustment methods were the lack of data on the impact of bias, a lack of expertise to conduct some of the more technically advanced methods, and a lack of time to collect evidence on bias for each specific problem. This situation could be improved by providing additional training on the methods, developing consistent measurements of bias domains for each field, and carrying out a systematic analysis of the association between bias features and effect size, which would provide valuable information for adjustment. The group noted the need for guidance and worked examples of different bias-adjustment methods used in meta-analyses in different disciplines.

The group recommended that steps be taken to understand what evidence already exists on the impact of bias on effect sizes, moving towards meta-epidemiological studies of the impact of bias. Crucially the most important bias domains for each specific context should be identified and it should be decided if there is a systematic association with bias that is worth adjusting for, for each domain. This exercise may need to be carried out separately for different disciplines.

6.3. DG3 – Quantitative approaches to combining evidence across evidence streams for hazard identification

Chair: Donald Rubin (Harvard University, USA)

Rapporteurs from the organising committee: Laura Martino (EFSA) and Rob de Vries (EBTC)

Follow-up of:

- Lecture 1 – Introduction to evidence integration for HI: overview of qualitative and quantitative methods and challenges. *Donald Rubin/Harvard University/USA*.
- Lecture 4 – Quantitative approaches to combining evidence across evidence streams. Stijn Vansteelandt/University of Ghent/BE and London School of Hygiene and Tropical Medicine/UK.

6.3.1. DG3 background and introduction

The issue of determining the relationship between cause and effect is traditionally referred to in the literature as causality assessment or, when referring to statistical methodologies, causal inference. In recent years a suite of quantitative methods and approaches has been developed to address causal questions (e.g. Pearl, 2009; Imbens and Rubin, 2015; Greenland, 2017; Hernan and Robins, forthcoming). In hazard identification in human risk assessment of chemicals, the objective is to draw conclusions about the causal relationship between exposure to a chemical and possible adverse effects in humans, based on evidence from laboratory animals, *in vitro* and *in silico* studies and human observational studies. In the obvious absence of randomised clinical trials, the evidence available on adverse effects of chemicals suffers from uncertainties mainly stemming from the confounding factors affecting the validity of observational data and the external validity/biological relevance issues afflicting the use of animal, *in vitro* and *in silico* data. Accounting for these uncertainties and reducing the potential bias in the conclusions about causality represents one of the primary challenges in this context. A concept of mechanistic validation has been proposed previously (Hartung et al., 2013a).

6.3.2. What are the current practices to integrate heterogeneous evidence on hazard identification?

Several participants gave short presentations to set the scene and to show how evidence that is heterogeneous in some respect is integrated in the various domains of risk assessment. Although some participants were aware of some quantitative approaches, none of the presenters had used such an approach themselves for hazard identification of the type being considered at the conference. It was highlighted that one of the main challenges in chemical risk assessment is the need to combine experimental data usually on animals and observational data usually on humans together with *in vitro* and *in silico* data. Extrapolation was considered a crucial issue (Hartung, 2017), because frequently data from other species or other chemicals have to be used and integrated in the assessment to compensate for scarcity of evidence on the target population and target substance. It was acknowledged that getting perfectly relevant and valid evidence is unrealistic. In light of this, the concept of uncertainty and its role when performing evidence integration were also discussed.

From the discussion it became evident the participants had a keen interest in finding out how quantitative approaches could help to draw conclusions in hazard identification.

6.3.3. What do we want to quantify in the context of integrating evidence for hazard identification?

The objective of quantification in hazard identification was discussed. The following elements were considered possible targets of quantification:

- Strength of the associations between a series of end-points (indicating adverse health effects) and a potential hazard (of a given chemical substance) obtained by integrating various evidence streams in light of their relevance and validity. It was acknowledged, though, that this target is more in the scope of hazard characterisation than hazard identification because it implies consideration of the effect of size and the dose–response relationship.
- Measures of the contribution of each source of information to the conclusion reached

(influence analysis).
- Quantification of the confidence/certainty in the hazard conclusion. Probability judgements (i.e. probabilistic risk assessment) can be used to express the confidence of the experts in the conclusion that a chemical x is a hazard, based on the available evidence. This judgement can be carried out as a quantitative weight of evidence evaluation (Linkov et al., 2015).

The discussion also focused on the type and quality of evidence needed to conclude on hazard. The possible role of 'omics was disputed among other sources. As an example, a short-term animal test on metabolomics was mentioned (van Ravenzwaay et al., 2014). The potential contribution of the adverse outcome pathway (AOP) approach to evidence integration was briefly addressed without reaching firm conclusions. The OECD Integrated Approaches to Testing and Assessment (IATA)[5] were mentioned. This framework relies on an integrated analysis of existing information, coupled with the generation of new information using testing strategies. It was highlighted though that IATA focusses on hazard characterisation.

The chair underlined the importance of optimising study design. Suggestions were made to use factorial designs and fractional replication, which consider several factors at the same time.

6.3.4. What are the current challenges in applying quantitative approaches in hazard identification?

The potential value of moving to quantitative approaches in hazard identification was recognised by the group. Examples of quantitative methods for evidence integration used in industry applications were mentioned. Approaches based on Bayesian networks have been proposed, for instance, in line with an Integrated Testing approach (Hartung et al., 2013b; Rovida et al., 2015), which were recently accepted for skin sensitisation in the regulatory context by ECHA. The purpose of this approach is to: 1. assess the probability of toxicity from different test results; 2. determine the most valuable next test given previous test results and other information; 3. have a measure of model stability (e.g. confidence intervals) and robustness.

The potential of machine learning techniques for the classification of a chemical as a hazard based, on integrated approaches using alternative assays, was also highlighted (Hartung, 2016; Luechtefeld and Hartung, 2017).

These methodologies seem promising and their applicability to hazard identification should be better investigated.

From the discussion the need for a harmonised terminology as well as for a better mutual understanding between toxicologists and statisticians emerged, the current lack of which represents a partial barrier to the application of quantitative methods.

6.3.5. What are the proposed actions for the future?

The need to join forces to progress on the application of quantitative approaches in hazard identification was clearly recognised.

EBTC informed the group about the intention to set up a Working Group to tackle the challenging issue of evidence integration.

Attention was raised about the opportunities offered by existing methodologies already applied in other fields. It was proposed to consider the possibility to:

- formalise a loss function to account for the value of adding additional source of evidence (e.g. using Bayesian decision theory);
- use extrapolation methods beyond weighting, as frequently there is the need to move outside the combination of available evidence (when using animals for instance). Convex combinations do not work well in these instances;

[5] http://www.oecd.org/chemicalsafety/risk-assessment/iata-integrated-approaches-to-testing-and-assessment.htm

- optimise experimental design, for instance looking at many chemicals and species at the same time (varying multiple factors at one time). Similarly improve the design and use of human observational studies.

6.4. DG4 – Using multiple end-points and multiple studies for dose–response modelling: quantitative approaches

Chair: Marc Aerts (Hasselt University, BE)

Rapporteurs from the organising committee: Jose Cortinas Abrahantes (EFSA) and Sebastian Hoffmann (EBTC)

Follow-up of:

- Lecture 5 – Introduction to benchmark dose estimation from multiple end-points and multiple studies: current practices and challenges. Marc Aerts, Hasselt University, Belgium.
- Lecture 6 – Combining evidence on multiple end-points in dose–response assessments: multivariate models. Wout Slob, National Institute for Public Health and the Environment (RIVM), Ministry of Health, Welfare and Sport, the Netherlands.
- Lecture 7 – Other quantitative methods for combining multiple studies and end-points. Matthew Wheeler, The National Institute for Occupational Safety and Health (NIOSH), Centers for Disease Control and Prevention (CDC), USA.

6.4.1. DG4 background and introduction

Evidence integration, in the context of dose–response modelling when estimating reference points (RP) or PoD using benchmark dose modelling (EFSA Scientific Committee et al., 2017c), is the process of combining information on the hazard of interest coming from: (a) multiple end-points observed in a single study; as well as (b) one or multiple end-points of several studies. When modelling the data for estimation, information characteristics need to be carefully considered, accounting for different aspects such as study design, end-points measured, dependences, etc. Current practices of setting RPs/PoDs, often circumvent integration by focusing on the most critical study and the critical end-point. Advanced statistical models allow the incorporation of several end-points from a single study, among others by multivariate approaches, to derive values such as RP and PoD. Other simplified methods such as the analysis of each individual end-point studied could also incorporate evidence provided within streams in a more ad-hoc fashion, but not necessarily implying a loss of efficiency and precision. Bayesian models provide the framework to incorporate uncertainties and variabilities not only among end-points, but also among studies. In this context model uncertainty also plays an important role, which can be addressed by model averaging techniques that can out-perform any single model in terms of coverage of interval estimates of the parameters of interest.

Before the Colloquium, the participants of this discussion group were provided with briefing notes, links to the guidance documents from EFSA and EPA on benchmark dose modelling, as well as the presentations of the lecturers and the discussion points prepared by the speakers and rapporteurs to stimulate the discussion on the points illustrated in the next sections.

6.4.2. Is there a way to better share information between statisticians and toxicologists?

In the discussion group this question was raised based on the need for more data to better understand the general behaviour of toxicological dose–response relationships, the underlying processes and to be able to combine the evidence from multiple end-points, which could even come from multiple studies. The need for a **data repository** was emphasised, with sufficient detail on the study design used, the end-points measured, the aim of the study conducted, the compound under study and containing individual data with sufficient metadata. Apart from enabling a better understanding the general behaviour of dose–response relationships, such data could be used to explore and examine the performance of basic, as well as more advanced, innovative statistical tools and methods in this context. The need for a **science forum**, in which people could pose questions, share ideas, offer solutions, have discussions, etc., but on which at the same time scientific outputs could be shared to promote cross-fertilisation between the different fields (toxicology, epidemiology and statistics), was discussed as well. The idea of creating a harmonised format to share information in general in current times of 'big data' was supported by most participants. It was also pointed out during the discussion that there is a need for a curator role for such a repository to assess the adequacy of data and information including the quality of information uploaded. The repository concept was conceived as a publicly available source of information that everybody could access, but its usage should be carried out through an application process to safeguard correct usage of the information. In such an application process for data usage, the purpose of the application should be stated, while any results obtained from the data should be uploaded in the repository, in this way contributing to the knowledge sharing process. The discussion was also centred on how this could be achieved, who should coordinate this activity, and which existing activities are similar or comparable. It was also suggested to not only consider current and future studies, but to include historical studies in the repository as well. The idea of creating such repository was supported by all participants in the discussion group.

6.4.3. Can a community of knowledge with available toxicological and statistical expertise be built?

The point here discussed was linked to the fact that when combining evidence, guidance on how this task should be performed is necessary. It was pointed out during the plenary presentation that reference on how to deal with such problems is very scarce in the available guidance documents. In both the updated guidance from EFSA (EFSA Scientific Committee et al., 2017c) and on the one published by EPA (US EPA, 2012), limited information is provided on what to do in such cases. Limited pragmatic solutions are discussed in both guidance documents, but how they should be performed is not explicitly defined. This aspect could be enhanced if a community of knowledge is created in which knowledge from both fields (toxicology and statistics) could be created, allowing for interaction between the different disciplines, creating opportunities to work on subjects that are key within the toxicology domain that could be dealt with by recent developments within the statistical field. It was also discussed that, very likely in the future, there will be a shift in the availability of types of data from current practice in which animal experiments are conducted to other types of data, such as *in vitro* or 'omics, etc. When considering other types of information, mechanistic models, AOPs and other methods are expected to become more prominent and investing in a community that could work on such topics could help prepare for that moment when it arrives. The idea of creating a community of knowledge could stimulate sharing methods and models and could boost their usage and further developments. Existing initiatives such as EFSA's Knowledge Junction repository were mentioned as examples of potential starting points. A community of knowledge could further provide indications as to which newly developed methods are often used or are of general interest and therefore could be implemented and offered through a user-friendly interface. It was pointed out that active use of such a repository is of extreme importance to maximally exploit such community. How active use could be stimulated was extensively discussed, but current practices do not support freely accessible knowledge sharing, although some participants thought that it could be feasible if the European Union community perceives the benefit of such practices, for instance, usage and testing of newly developed methods that could potentially lead to improvements in the methods proposed. The participants also highlighted the need for training, facilitating the usage of new methodology as a guided and supervised process, evidencing the importance of cooperation between fields when working on methodological developments.

6.4.4. What are the implications of empirical evidence shown?

The participants discussed that once a data repository is available and a community of knowledge is created, then information and methods could be developed to study interesting hypotheses in the context of BMD estimation, such as the ones presented in the plenary session. Hypotheses discussed were:

- end-points could be seen as all equally sensitive,
- inter-species differences are negligible.

Such hypotheses could imply major changes associated with risk assessment practice. Its implications were noted and discussed, i.e. inter-species differences defined as minor or negligible, how this would be considered when combining evidence from experiments in which the compound under study was assessed using different animal species. Could information from these experiments be simply combined and assessed as a compound effect and extrapolation factors from animals to human might render unnecessary? What are the implications for previous assessments? These concerns were raised during the discussion, but of course, before trying to answer how to deal with the issues put forward, evidence supporting such hypotheses needs to be provided.

6.4.5. Could all be looked from the risk scale viewpoint when combining end-points, going from continuous to quantal data pros and cons?

The last point discussed was the possibility to transform different scales from various end-points to a common scale defined by the risk of observing the undesired effect of interest. The benefit of such approach is that all end-points are comparable, and methods to combine end-points when dealing with binary outcomes are readily available in the statistical field. It was also pointed out that in such cases there is no need to define the size of the effect for different scales, as all end-points are measured on the same scale. However, the objection was made that an effect to be considered as undesirable depends on the type of effect, and, in most cases, on the size of the effect (one of the examples discussed was that of comparing malignant tumours, moderate liver lesions and 10% change in liver enzymes). It was also discussed that information lost might be of concern when using such an approach and this might conflict with other methodological developments within the field of toxicology. It was concluded that further research is needed to evaluate if this idea is applicable.

7. Overall conclusions and way forward

Evidence integration in chemical risk assessment is a challenge and further methodological developments are needed to support the production of evidence-based hazard and risk conclusions.

Among the structured qualitative approaches, the GRADE is a promising framework for qualifying the certainty in a body of evidence. This method, whose use has been recently extended from the healthcare research to the field of chemical risk assessment for hazard identification, incorporates the Bradford Hill criteria for causality, including fundamental aspects like consistency, sensitivity and biological plausibility. Therefore, GRADE is fully applicable to the multistream context of environmental health research and regulatory risk assessment. However, more guidance, clear definitions and explicit operationalisation rules, as well as testing and validation, are needed to support its implementation in this research field.

Alternatively, quantitative methods can be used to address limitations in the evidence and, in general, sources of uncertainty and variability that can affect conclusions on hazard identification and characterisation. Quantitative approaches provide the decision makers with conclusions that are less prone to subjective interpretation and more explicit as far as the level of conservativism. Conversely, they request application of more complex methods and sometimes complementary information with respect to the one traditionally collected.

Among the quantitative methods, bias-adjusted meta-analyses include a suite of techniques that allows accounting for the direction and the magnitude of bias in the effect estimate. Nevertheless bias adjustment must be informed by evidence on the impact of bias on the effect estimates, which may not always be available.

Quantitative methods are also available to support conclusions on causal inference and establishment of dose–response relationships. Bayesian networks offer powerful tools to combine evidence from heterogeneous sources and identify possible causal relationships even when there are complex multivariate associations. These approaches still require validation in the context of hazard identification.

In the context of *in vivo* studies, several methods considering the nature of the end-point using the frequentist as well as the Bayesian paradigm can be used to model dose–response data. Methods that consider the possibility of combining end-points by converting them into a common scale were shown.

These methods could even include random effects terms to account for study heterogeneity when pooling results from several experiments. In the context of big data, the need for other types of model might be more relevant when other type of data (*in vitro*, ʼomics, etc.) become more common. However, quantitative methods and their underlying hypotheses (e.g. for dose–response modelling, equal sensitivity for all end-points or magnitude of inter-species differences) require testing, for which data are required.

Overall, to allow further development, testing, validation and effective implementation of methods for evidence integration (both qualitative and quantitative), the following needs should be addressed by the relevant scientific community:

- **Need for more published primary toxicological and epidemiological data**, to explore and examine the performance of basic and more advanced statistical tools, as well as qualitative approaches. It is important to note that the data need to include all available types, as in toxicology the data are shifting from animal to other models, such as *in vitro*, ʼomics, computational models etc.).
- **Need to optimise experimental design**, for instance looking at many chemicals and species at the same time (varying multiple factors at one time). Similarly, **need to improve the design and exploitation of human observational studies**.
- **Need for a shared primary data repository**. Data should be shared in a publicly available data repository, whose sustainability and correct usage should be guaranteed by a sustainable funding and business model that allows for consistent monitoring of the data quality, including access, use and sharing.
- Need for **a community of knowledge of toxicologists, epidemiologists and statisticians**, to allow across-discipline interaction, to facilitate mutual understanding, and to create opportunities for work on key domain subjects (e.g. AOP) in light of the most recent developments within the respective fields. This community should promote the development of **harmonised terminology** and could be supported by a **science forum**, in which people could pose questions, share ideas, offer solutions, have discussions, etc., but in which at the same time scientific outputs could be shared to promote cross-fertilisation between different fields. Initiatives such as EFSA's Knowledge Junction repository are a potential starting point. This community should also promote **training** opportunities on the different methods and **regular exchange** through scientific conferences and workshops.

Equally, to be conducted soundly, evidence integration should be undertaken by **multidisciplinary groups** of assessors, including both experts from the specific chemical field, toxicologists and methodologists knowledgeable of the various integration techniques.

EFSA and EBTC will continue the collaboration to provide a platform for the multidisciplinary interaction between scientists with the overarching goal to develop, test and validate best practices in safety assessment. Colloquia, such as the one on evidence integration, break barriers and silos between the scientific disciplines, facilitate the development of a common vocabulary and provide room for free scientific discussion and argument, which ultimately leads to advancements in science and to the development of new methodologies for risk assessment.

References

Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer T, Varonen H, Vist GE, Williams JW, Jr., Zaza S and Group GW, 2004. Grading quality of evidence and strength of recommendations. BMJ, 328, 1490. doi:10.1136/bmj.328.7454.1490

Bradford Hill A, 1965. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 58, 295–300.

Cooper GS, Lunn RM, Agerstrand M, Glenn BS, Kraft AD, Luke AM and Ratcliffe JM, 2016. Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures. Environment International, 92–93, 605–610. doi:10.1016/j.envint.2016.03.017

Doi SA, Barendregt JJ, Khan S, Thalib L and Williams GM, 2015. Advances in the meta-analysis of heterogeneous clinical trials II: The quality effects model. Contemp Clin Trials, 45(Pt A):123-9 doi: 10.1016/j.cct.2015.05.010. Epub 2015 May 21.

Durrheim DN and Reingold A, 2010. Modifying the GRADE framework could benefit public health. Journal of Epidemiology and Community Health, 64, 387–387. doi:10.1136/jech.2009.103226

EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Benfenati E, Chaudhry QM, Craig P, Frampton G, Greiner M, Hart A, Hogstrand C, Lambre C, Luttik R, Makowski D, Siani A, Wahlstroem H, Aguilera J, Dorne J-L, Fernandez Dumont A, Hempen M, Valtueña Martínez S, Martino L, Smeraldi C, Terron A, Georgiadis N and Younes M, 2017a. Guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal 2017a;15(8):4971. doi:10.2903/j.efsa.2017.4971

EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Younes M, Bresson J-L, Griffin J, Hougaard Benekou S, van Loveren H, Luttik R, Messean A, Penninks A, Ru G, Stegeman JA, van der Werf W, Westendorf J, Woutersen RA, Barizzone F, Bottex B, Lanzoni A, Georgiadis N and Alexander J, 2017b. Guidance on the assessment of the biological relevance of data in scientific assessments. EFSA Journal, 15, 4970. doi:10.2903/j.efsa.2017.4970

EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen KH, More S, Mortensen A, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Silano V, Solecki R, Turck D, Aerts M, Bodin L, Davis A, Edler L, Gundert-Remy U, Sand S, Slob W, Bottex B, Abrahantes JC, Marques DC, Kass G and Schlatter JR, 2017c. Update: use of the benchmark dose approach in risk assessment. EFSA Journal, 15, 4658. doi:10.2903/j.efsa.2017.4658

EFSA Scientific Committee, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Younes M, Craig P, Hart A, Von Goetz N, Koutsoumanis K, Mortensen A, Ossendorp B, Martino L, Merten C, Mosbach-Schulz O and Hardy A, 2018. Guidance on uncertainty analysis in scientific assessments. EFSA Journal, 16, 5123. doi:10.2903/j.efsa.2018.5123

Glass GV, 1991. The future of meta-analysis. by Kenneth W. Wachter, Miron L. Straf. Journal of the American Statistical Association, 86, 1141–1142. doi:10.2307/2290539

Greenland S, 2017. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. European Journal of Epidemiology, 32, 3–20. doi:10.1007/s10654-017-0230-6

Guyatt GH, Ebrahim S, Alonso-Coello P, Johnston BC, Mathioudakis AG, Briel M, Mustafa RA, Sun X, Walter SD, Heels-Ansdell D, Neumann I, Kahale LA, Iorio A, Meerpohl J, Schunemann HJ and Akl EA, 2017. GRADE guidelines 17: assessing the risk of bias associated with missing participant outcome data in a body of evidence. Journal of Clinical Epidemiology, 87, 14–22. doi:10.1016/j.jclinepi.2017.05.005

Hartung T, 2009. Food for thought…on evidence-based toxicology. ALTEX, 26, 75–82. doi:10.14573/altex.2009.2.75

Hartung T, 2016. Making big sense from big data in toxicology by read-across. ALTEX, 33, 83–93. doi:10.14573/altex.1603091

Hartung T, 2017. Perspectives on *in vitro* to *in vivo* extrapolations. Applied In Vitro Toxicology. doi:10.1089/aivt.2016.0026

Hartung T, Hoffmann S and Stephens M, 2013a. Mechanistic validation. ALTEX, 30, 119–130. doi:10.14573/altex.2013.2.119

Hartung T, Luechtefeld T, Maertens A and Kleensang A, 2013b. Integrated testing strategies for safety assessments. ALTEX, 30, 3–18. doi:10.14573/altex.2013.1.003

Hernan MA and Robins JM, forthcoming. Causal inference. Chapman & Hall/CRC, Boca Raton.

Hoffmann S and Hartung T, 2006. Toward an evidence-based toxicology. Human & Experimental Toxicology, 25, 497–513. doi:10.1191/0960327106het648oa

Hoffmann S, de Vries RBM, Stephens ML, Beck NB, Dirven H, Fowle JR, 3rd, Goodman JE, Hartung T, Kimber I, Lalu MM, Thayer K, Whaley P, Wikoff D and Tsaioun K, 2017. A primer on systematic reviews in toxicology. Archives of Toxicology, 91, 2551–2575. doi:10.1007/s00204-017-1980-3

Imbens GW and Rubin DB, 2015. Causal inference for statistics, social, and biomedical sciences: an introduction, Cambridge University Press, Cambridge, 677 pp.

Linkov I, Massey O, Keisler J, Rusyn I and Hartung T, 2015. From "weight of evidence" to quantitative data integration using multicriteria decision analysis and bayesian methods. ALTEX, 32, 3–8. doi:10.14573/altex.1412231

Luechtefeld T and Hartung T, 2017. Computational approaches to chemical hazard assessment. ALTEX, 34, 459–478. doi:10.14573/altex.1710141

Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Hooijmans C, Langendam M, Mandrioli D, Mustafa RA, Rehfuess EA, Rooney AA, Shea B, Silbergeld EK, Sutton P, Wolfe MS, Woodruff TJ, Verbeek JH, Holloway AC, Santesso N and Schunemann HJ, 2016. GRADE: Assessing the quality of evidence in environmental and occupational health. Environment International, 92–93, 611–616. doi:10.1016/j.envint.2016.01.004

MRC Centre Cambridge 2017. Development of a method for adjusting trial results for biases in meta-analysis: combining generic evidence on bias with detailed trial assessment. Research Councils UK. Available online: http://gtr.rcuk.ac.uk/projects?ref=MC_EX_MR%2FK014587%2F1

NAS (National Academy of Sciences), 2014. Review of EPA's Integrated Risk Information System (IRIS) process. 978-0-309-30414-6, Washington, DC, 170 pp. Available online: https://www.nap.edu/catalog/18764/review-of-epas-integrated-risk-information-system-iris-process

Pearl J, 2009. Causality: models, reasoning, and inference. 2nd ed, Cambridge University Press, Cambridge, 484 pp.

Rooney AA, Boyles AL, Wolfe MS, Bucher JR and Thayer KA, 2014. Systematic review and evidence integration for literature-based environmental health science assessments. Environmental Health Perspectives, 122, 711–718. doi:10.1289/ehp.1307972

Rovida C, Alepee N, Api AM, Basketter DA, Bois FY, Caloni F, Corsini E, Daneshian M, Eskes C, Ezendam J, Fuchs H, Hayden P, Hegele-Hartung C, Hoffmann S, Hubesch B, Jacobs MN, Jaworska J, Kleensang A, Kleinstreuer N, Lalko J, Landsiedel R, Lebreux F, Luechtefeld T, Locatelli M, Mehling A, Natsch A, Pitchford JW, Prater D, Prieto P, Schepky A, Schuurmann G, Smirnova L, Toole C, van Vliet E, Weisensee D and Hartung T, 2015. Integrated Testing Strategies (ITS) for safety assessment. ALTEX, 32, 25–40. doi:10.14573/altex.1411011

Schünemann HJ, Best D, Vist G, Oxman AD and Group GW, 2003. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ, 169, 677–680.

Schünemann H, Hill S, Guyatt G, Akl EA and Ahmed F, 2011. The GRADE approach and Bradford Hill's criteria for causation. Journal of Epidemiology and Community Health, 65, 392–395. doi:10.1136/jech.2010.119933

Slob W, 2017. A general theory of effect size, and its consequences for defining the benchmark response (BMR) for continuous end-points. Critical Reviews in Toxicology, 47, 342–351. doi:10.1080/10408444.2016.1241756

Stephens ML, Andersen M, Becker RA, Betts K, Boekelheide K, Carney E, Chapin R, Devlin D, Fitzpatrick S, Fowle JR, 3rd, Harlow P, Hartung T, Hoffmann S, Holsapple M, Jacobs A, Judson R, Naidenko O, Pastoor T, Patlewicz G, Rowan A, Scherer R, Shaikh R, Simon T, Wolf D and Zurlo J, 2013. Evidence-based toxicology for the 21st century: opportunities and challenges. ALTEX, 30, 74–103.

Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hrobjartsson A, Kirkham J, Juni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schunemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF and Higgins JP, 2016. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, i4919. doi:10.1136/bmj.i4919

Turner RM, Spiegelhalter DJ, Smith GC and Thompson SG, 2009. Bias modelling in evidence synthesis. J R Stat Soc Ser A Stat Soc, 172, 21–47. doi:10.1111/j.1467–985X.2008.00547.x

US EPA (US Environmental Protection Agency), 2012. Benchmark dose technical guidance. Washington, DC, 87 pp. Available online: https://www.epa.gov/sites/production/files/2015–01/documents/benchmark_dose_guidance.pdf

van Ravenzwaay B, Montoya GA, Fabian E, Herold M, Krennrich G, Looser R, Mellert W, Peter E, Strauss V, Walk T and Kamp H, 2014. The sensitivity of metabolomics versus classical regulatory toxicology from a NOAEL perspective. Toxicology Letters, 227, 20–28.

Vandenberg LN, Agerstrand M, Beronius A, Beausoleil C, Bergman A, Bero LA, Bornehag CG, Boyer CS, Cooper GS, Cotgreave I, Gee D, Grandjean P, Guyton KZ, Hass U, Heindel JJ, Jobling S, Kidd KA, Kortenkamp A, Macleod MR, Martin OV, Norinder U, Scheringer M, Thayer KA, Toppari J, Whaley P, Woodruff TJ and Ruden C, 2016. A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. Environmental Health, 15, 74. doi:10.1186/s12940-016-0156-6

Wachter KW and Straf ML, 1990. The future of meta-analysis. Russell Sage Foundation, New York, 238 pp.

Welton NJ, Ades AE, Carlin JB, Altman DG and Sterne JAC, 2009. Models for potentially biased evidence in meta-analysis using empirically based priors. Journal of the Royal Statistical Society Series a-Statistics in Society, 172, 119–136. doi:10.1111/j.1467–985X.2008.00548.x

Woodruff TJ and Sutton P, 2014. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environmental Health Perspectives, 122, 1007–1014. doi:10.1289/ehp.1307175

Zurlo J, 2011. Evidence-based Toxicology Collaboration Kick-off Meeting. ALTEX, 28, 152.

Appendix A – List of participants

Name	Affiliation	Country	DG
Marc AERTS	Hasselt University	BE	4
Elisa AIASSA	European Food Safety Authority (EFSA)	IT	2
Ardelean ADRIAN IOAN	Sanitary Veterinary and Food Safety Directorate Cluj	RO	2
Fulvio BARIZZONE	European Food Safety Authority (EFSA)	IT	2
Federica BARRUCCI	European Food Safety Authority (EFSA)	IT	4
Maria BASTOS	University of Porto – Faculty of Pharmacy	PT	1
Andrea BAU	European Food Safety Authority (EFSA)	IT	1
Anna BERONIUS	Karolinska Institutet	SE	3
Irena BOGOEVA	Risk Assessment Centre on Food Chain, Ministry of Agriculture, Food and Forestry	BG	3
Monia BOUKTIF ZARROUK	National agency of sanitary and environmental control of the products	TU	1
Jan BROZEK	McMaster University	CA	4
Paulo CARMONA	Economic and Food Safety Authority of Portugal (ASAE)	PT	3
Anna Federica CASTOLDI	European Food Safety Authority (EFSA)	IT	1
Laura CICCOLALLO	European Food Safety Authority (EFSA)	IT	2
Wendie CLAEYS	Belgian Federal Agency for the Safety of the Food Chain	BE	4
Jose CORTINAS ABRAHANTES	European Food Safety Authority (EFSA)	IT	4
Mafalda COSTA	Economic and Food Safety Authority of Portugal (ASAE)	PT	3
Federica CRIVELLENTE	European Food Safety Authority (EFSA)	IT	1
Cristina CROERA	European Food Safety Authority (EFSA)	IT	1
Alie DE BOER	Maastricht University	NL	3
Agnès DE SESMAISONS	European Food Safety Authority (EFSA)	IT	3
Rob DE VRIES	SYRCLE, Department for Health Evidence, Radboudumc & Evidence-Based Toxicology Collaboration	NL	3
Amélia DELGADO	MeditBio – University of Algarve	PT	1
Sofia DIAS	University of Bristol	UK	2
Suhail DOI	Australian National University and Qatar University	QA	2
Jack FOWLE	Science to Inform, LLC	US	1
Geoff FRAMPTON	University of Southampton	UK	2
James FREEMAN	ExxonMobil Biomedical Sciences, Inc.	US	1
Nikolaos GEORGIARIS	European Food Safety Authority (EFSA)	IT	1
Marios GEORGIARIS	European Food Safety Authority (EFSA)	IT	2
David GOTT	Food Standards Agency	UK	3
Matthias GREINER	German Federal Institute for Risk Assessment (BfR)	DE	2
Ksenia GROH	Food Packaging Forum Foundation	CH	4
Annika HANBERG	Karolinska Institutet, Institute of Environmental Medicine	SE	3
Andy HART	Fera Science Ltd	UK	2
Thomas HARTUNG	John Hopkins University	US	3
Antonio F. HERNANDEZ-	University of Granada	ES	2

Name	Affiliation	Country	DG
JEREZ			
Julian HIGGINS	University of Bristol	UK	2
Sebastian HOFFMANN	SEH Consulting + Services	DE	4
Francis KRUSZEWSKI	American Cleaning Institute	US	1
Miranda LANGENDAM	University of Amsterdam	NL	1
Therese LILLEGARD	Norwegian Scientific Committee for Food Safety	NO	2
Paoloa MANINI	European Food Safety Authority (EFSA)		3
MARREILHA DOS SANTOS	Faculdade de Farmácia Universidade de Lisboa	PT	3
Olwenn MARTIN	Brunel University London	UK	3
Laura MARTINO	European Food Safety Authority (EFSA)	IT	3
Filipa MELO DE VASCONCELOS	Economic and Food Safety Authority of Portugal (ASAE)	PT	2
Hans MIELKE	German Federal Institute for Risk Assessment (BfR)	DE	4
Sarogini MONTEIRO	Economic and Food Safety Authority of Portugal	PT	1
Alicja MORTENSEN	National Research Institute on Working Environment (NRCWE)	DK	1
Andrei MOT	University of Agronomic Sciences and Veterinary Medicine of Bucharest	RO	3
Pedro NABAIS	Economic and Food Safety Authority of Portugal (ASAE)	PT	1
Emer O'REILLY	Food Safety Authority of Ireland (FSAI)	IE	4
Rebecca RAM	Safer Medicines Trust	UK	1
Gilles RIVIERE	French Agency for food environment and occupational health and safety (ANSES)	FR	4
Donald Bruce RUBIN	Harvard University	US	3
Maria SANTOS	Autoridade de Segurança Alimentar e Económica (ASAE)	PT	4
Kerstin SCHMIDT	BioMath GmbH	DE	2
Holger SCHÜNEMANN	McMaster University	CA	1
Dick SIJM	Netherlands Food and Consumer Product Safety Authority (NVWA)	NL	3
Lea SLETTING JAKOBSEN	Technical University of Denmark (DTU)	DK	4
Wout SLOB	Netherlands National Institute for Public Health and the Environment (RIVM)	NL	4
Inger-Lise STEFFENSEN	Norwegian Institute of Public Health/Norwegian Scientific Committee for Food Safety	NO	3
Martin STEPHENS	Johns Hopkins University	US	1
Kristina THAYER	Environmental Protection Agency (EPA)	US	1
Sofie Theresa THOMSEN	National Food Institute, Technical University of Denmark (DTU Food)	DK	3
Daniela TOMCIKOVA	European Food Safety Authority (EFSA)	IT	2
Katya TSAIOUN	Evidence-Based Toxicology Collaboration (EBTC)	US	4
Karin VAN EDE	Key Toxicology	NL	3
Stijn VANSTEELANDT	Ghent University and the London School of Hygiene and Tropical Medicine	BE	3
Didier VERLOO	European Food Safety Authority (EFSA)	IT	2
Susana VIEGAS	ESTeSL-IPL	PT	2

Name	Affiliation	Country	DG
Misha VROLIJK	Maastricht University and Netherlands Food and Consumer Product Safety Authority (NVWA)	NL	4
Ine WAALKENS-BERENDSEN		NL	2
Paul WHALEY	Lancaster Environment Centre	UK	1
Matthew WHEELER	National Institute for Occupational Safety and Health	US	4
Daniele WIKOFF	ToxStrategies	US	1
Rudolf WOUTERSEN	Former employer TNO Innovation for Life	NL	1
Ami YAMADA	Danone	FR	3
Maged YOUNES	EFSA Panel Member	DE	1
Johanna ZILLIACUS	Institute of Environmental Medicine, Karolinska Institutet	SE	4

Appendix B - Colloquium Programme

DAY 1 Wednesday, 25 October 2017		
08:00	Registration	
09:00	Welcome and introduction to the event	Didier Verloo, *EFSA, Assessment and Methodological Support unit*
09:05	Objectives of the Colloquium	Katya Tsaioun, *EBTC at Johns Hopkins Bloomberg School of Public Health (USA)*
SESSION 1 – INTEGRATING EVIDENCE FOR HAZARD IDENTIFICATION (HI) Chair: Katya Tsaioun, EBTC at Johns Hopkins Bloomberg School of Public Health (USA)		
09:15	Lecture 1 – Introduction to evidence integration for HI: overview of qualitative and quantitative methods and challenges *Questions*	Donald Rubin, *Harvard University (USA)*
09:45	Lecture 2 – Integrating evidence within and across evidence streams using qualitative methods *Questions*	Kristina Thayer*, Environmental Protection Agency (EPA), Integrated Risk Information System (IRIS) Division (USA)*
10:15	Lecture 3 – Recent developments for combining evidence within evidence streams: bias-adjusted meta-analysis *Questions*	Julian Higgins, *University of Bristol (UK)*
10:45	Lecture 4 – Quantitative approaches to combining evidence across evidence streams *Questions*	Stijn Vansteelandt, *University of Ghent (BE)*
11:15	Coffee/Tea break	
SESSION 2 – INTEGRATING EVIDENCE FOR DOSE–RESPONSE MODELLING Chair: Didier Verloo, EFSA, Assessment and Methodological Support unit		
11:45	Lecture 5 – Introduction to dose–response modelling to derive health-based guidance values: current practice and challenges *Questions*	Marc Aerts, *Hasselt University (BE)*
12:15	Lecture 6 – Combining evidence on multiple end-points in dose–response assessments: multivariate models *Questions*	Wout Slob, *National Institute for Public Health and the Environment (RIVM), Ministry of Health, Welfare and Sport (The Netherlands)*
12:45	Lecture 7 – Other quantitative methods for combining multiple studies and end-points *Questions*	Matthew Wheeler, *The National Institute for Occupational Safety and Health (NIOSH), Centers for Disease Control and Prevention (CDC) (USA)*
13:15	Introduction to discussion groups	Didier Verloo and Katya Tsaioun
13:20	Lunch break	

SESSION 3 – DISCUSSION GROUPS (DG)		
14:30	DG1: Qualitative methods for integrating evidence within- and across evidence streams for HI	Chair Holger Schünemann, *McMaster University (Canada)* Rapporteurs Paul Whaley (EBTC) and Daniele Wikoff (EBTC)
	DG2: Bias-adjusted meta-analysis	Chair Sofia Dias, *University of Bristol (UK)* Rapporteurs Fulvio Barizzone (EFSA) and Elisa Aiassa (EFSA/EBTC)
	DG3: Quantitative approaches to combining evidence across evidence streams for HI	Chair: Donald Rubin, *Harvard University (USA)* Rapporteurs Laura Martino (EFSA) and Rob de Vries (EBTC)
	DG4: Using multiple end-points and multiple studies for dose–response modelling: quantitative approaches	Chair Marc Aerts, *Hasselt University (BE)* Rapporteurs Jose Cortinas Abrahantes (EFSA) and Sebastian Hoffmann (EBTC)
16:30	Coffee/Tea break	
17:00	Discussion groups continue	
18:30	Adjourn	
19:00	Networking cocktail	

DAY 2 Thursday, 26 October 2017 am		
SESSION 4 – CONTINUATION OF DISCUSSION GROUPS		
09:00	Focus on summarising challenges, guidance needs and related outcomes of the discussion groups and the production of reports to the plenary session	
10:00	Coffee/Tea break	
SESSION 5 – FINAL PLENARY SESSION Co-chairs: Didier Verloo and Katya Tsaioun		
10:30	Report back from DG1 and discussion	Holger Schünemann, *McMaster University (Canada)*
11:10	Report back from DG2 and discussion	Sofia Dias, *University of Bristol (UK)*
11:50	Report back from DG3 and discussion	Donald Rubin, *Harvard University (USA)*
12:30	Report back from DG4 and discussion	Marc Aerts, *Hasselt University (BE)*
13:10	Take-home messages	Daniele Wikoff, *ToxStrategies, Inc. and EBTC (USA)*
13:30	COLLOQUIUM ADJOURNS	