Published in final edited form as:

ALTEX. 2021 ; 38(2): 336–347. doi:10.14573/altex.2101211.

Applying Evidence-Based Methods to the Development and Use of Adverse Outcome Pathways

Rob B. M. de Vries[1,2], **Michelle Angrish**[3], **Patience Browne**[4], **Jan Brozek**[5], **Andrew A. Rooney**[6], **Daniele S. Wikoff**[7], **Paul Whaley**[1,8], **Stephen W. Edwards**[9], **Rebecca L. Morgan**[5], **Ingrid L. Druwe**[3], **Sebastian Hoffmann**[1,10], **Thomas Hartung**[11], **Kristina Thayer**[3], **Marc T. Avey**[12], **Brandiese E. J. Beverly**[6], **Maicon Falavigna**[5,13], **Catherine Gibbons**[3], **Katy Goyak**[14], **Andrew Kraft**[3], **Fernando Nampo**[15], **Amir Qaseem**[16], **Meg Sears**[17], **Jasvinder A. Singh**[18], **Catherine Willett**[19], **Erin Y. Yost**[3], **Holger Schünemann**[5,20], **Katya Tsaioun**[1]

[1]Evidence-Based Toxicology Collaboration (EBTC) at Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA;

[2]SYRCLE, Department for Health Evidence, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands;

[3]United States Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessments, Research Triangle Park, NC, USA;

[4]Test Guidelines Programme, Environmental Directorate, OECD, Paris, France;

[5]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada;

[6]Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA;

[7]ToxStrategies, Asheville, NC, USA;

[8]Lancaster Environment Centre, Lancaster University, Lancaster, UK;

[9]RTI International, Research Triangle Park, NC, USA;

[10]seh consulting + service, Paderborn, Germany;

[11]Center for Alternatives to Animal Testing (CAAT) at Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA;

[12]ICF, Durham, NC, USA;

[13]National Institute for Health Technology Assessment, UFRGS, Porto Alegre, Brazil;

[14]ExxonMobil Biomedical Sciences Inc., Annandale, NJ, USA;

[15]Evidence-Based Public Health Research Group, Latin-American Institute of Life and Nature Sciences, Federal University of Latin-American Integration, Foz do Iguassu, Parana, Brazil;

ktsaiou1@jhu.edu .

Conflict of interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this meeting report.

[16]Center for Evidence Reviews, The American College of Physicians, Philadelphia, PA, USA;

[17]Canadian Environmental Health Information Infrastructure, Ottawa Hospital Research Institute, Ottawa, ON, Canada;

[18]Medicine Service, VA Medical Center, Birmingham, AL, USA; Department of Medicine at the School of Medicine, University of Alabama at Birmingham (UAB), Birmingham, AL, USA; Department of Epidemiology at the UAB School of Public Health, Birmingham, AL, USA;

[19]Humane Society International, Washington, DC, USA;

[20]McMaster GRADE Centre and Michael G DeGroote Cochrane Canada Centre, McMaster University, Hamilton, ON, Canada

Abstract

The workshop “Application of evidence-based methods to construct mechanistic frameworks for the development and use of non-animal toxicity tests” was organized by the Evidence-based Toxicology Collaboration and hosted by the Grading of Recommendations Assessment, Development and Evaluation Working Group on June 12, 2019. The purpose of the workshop was to bring together international regulatory bodies, risk assessors, academic scientists, and industry to explore how systematic review methods and the adverse outcome pathway framework could be combined to develop and use mechanistic test methods for predicting the toxicity of chemical substances in an evidence-based manner. The meeting covered the history of biological frameworks, the way adverse outcome pathways are currently developed, the basic principles of systematic methodology, including systematic reviews and evidence maps, and assessment of certainty in models, and adverse outcome pathways in particular. Specific topics were discussed via case studies in small break-out groups. The group concluded that adverse outcome pathways provide an important framework to support mechanism-based assessment in environmental health. The process of their development has a few challenges that could be addressed with systematic methods and automation tools. Addressing these challenges will increase the transparency of the evidence behind adverse outcome pathways and the consistency with which they are defined; this in turn will increase their value for supporting public health decisions. It was suggested to explore the details of applying systematic methods to adverse outcome pathway development in a series of case studies and workshops.

1 Introduction

Two relatively new developments in the field of toxicology and environmental health are the development and use of adverse outcome pathways (AOPs) (Villeneuve et al., 2014) and the application of evidence-based approaches such as systematic reviews (Griesinger et al., 2008; Hoffmann et al., 2017; Stephens et al., 2016; Whaley et al., 2016) and systematic evidence maps[1] (Wolffe et al., 2019). An AOP describes a sequence of temporally and causally linked events at different levels of biological organization, which follows exposure to a stressor (e.g., chemical, physical, etc.) and leads to an adverse health effect in humans or

[1]Systematic reviews are designed to test specific hypotheses and are focused on narrowly defined questions, whereas systematic maps are more exploratory. Systematic maps are summaries of what the existing research is, not what it says.

wildlife. AOPs are used to organize mechanistic information and support the application of mechanistic data in chemical safety assessment (Groh et al., 2015). A systematic review can be defined as a process of systematically searching, selecting, appraising, and synthesizing evidence from existing research in such a way as to minimize the risk of bias and maximize transparency when summarizing what is already known in relation to answering a research question (Jadad et al., 1996). The methodology of systematic reviews was developed in the field of clinical research in humans (Scholten et al., 2005) but is increasingly being applied in the fields of toxicology and environmental health (Rooney et al., 2016; Woodruff and Sutton, 2014; Vandenberg et al., 2016) to collect data from the scientific literature for applications such as risk assessment in a more structured, transparent, and unbiased way.

So far, these two developments have taken place independently. However, although the concepts of AOPs and evidence-based methods are not directly linked, they may be mutually reinforcing. AOPs are a useful framework to organize and assess mechanistic evidence in risk assessments. For instance, a body of *in vitro* evidence may be considered less indirect or externally more valid (compared to evidence from animal studies) if that evidence is linked to an AOP. However, the process of assembling and evaluating the mechanistic information behind an AOP is reliant on expert knowledge and could be strengthened significantly with systematic methodologies as a foundation. Hence, systematic methods could be used here to an advantage to make the origination, development and use of AOPs more objective and reliant on the wide body of literature, while still relying on experts in the field for interpreting the data.

In order to explore if and how AOPs and evidence-based methods could work together, the Evidence-Based Toxicology Collaboration (EBTC) held a workshop, in collaboration with the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group on June 12, 2019 at McMaster University in Hamilton (Ontario), Canada. The workshop preceded the annual GRADE meeting on June 13–14.

The goal of the workshop was to discuss, via four related themes, how systematic review methods and AOP concepts can be combined to develop and use mechanistic evidence and new approach methodologies (NAMs) for predicting the toxicity of chemical substances in an evidence-based manner. More specifically, the workshop addressed these four questions:

1. How do scientists distinguish high quality *in vitro* studies from low quality ones?
2. What would a systematic approach to the development of AOPs look like?
3. How does one assess the certainty in AOPs, i.e., distinguish spurious AOPs from plausible ones?
4. How can AOPs inform the development of NAMs?

The workshop consisted of six expert plenary topic introductions[2], followed by break-out sessions per theme co-led by plenary speakers and EBTC researchers. Thirty-four researchers from academic institutes and regulatory agencies participated.

[2]The full recording of the workshop can be found on the YouTube channel of EBTC (https://youtu.be/NNa0r2qL4pI; accessed 28.01.2021).

2 Background on AOPs, evidence-based approaches and GRADE

2.1 EBTC and evidence-based approaches

Dr **Katya Tsaioun**, director of the Evidence-based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health, introduced EBTC as a collaboration of international stakeholders in toxicology and risk assessment, bound by the same vision of bringing evidence-based methods to the field and making them the standard used in risk assessment. Dr Tsaioun introduced the topic of the workshop, stressing that the AOP framework was a significant contribution to developing a transparent, reproducible framework for risk assessment of chemicals that allows for modernization of the regulatory testing paradigm. The AOP framework allows incorporation of advancements in the understanding of toxicological mechanisms and thus addresses some of the challenges and barriers that exist in integrating new science into toxicology and risk assessment. These challenges include uncertainty in assumptions about biological mechanisms and the tests that provide information on these mechanisms, and they are particularly important in the regulatory environment, where certainty in the results is paramount to make decisions about human and environmental health.

Dr Tsaioun then stressed that the mode of action (MoA) frameworks in the field of drug development offer an example for toxicologists. The concept of MoA (Fig. 1) has been in use in pharmacology for many decades and is at the start of every drug discovery program in which a molecular target is identified, validated and de-risked (Tsaioun, 2010). This mechanistic knowledge is then traced to higher levels of organization such as cellular effects, organ effects, and whole organism effects. In pharmacology, the aim is to find molecules with a desirable effect on the body to stop or modify a disease, whereas in toxicology the goal is to detect potential adverse effects of stressors such as environmental pollution, industrial chemicals, etc.

The AOP framework (Fig. 2) builds on the MoA concept (Meek, 2014). This framework consists of the molecular initiating event (MIE) and a series of cellular and tissue-level key events (KEs) that eventually lead to specific adverse outcomes (AO) on the organ and/or organism levels. This framework has been developed (Villeneuve et al., 2014) as a way to organize information, in particular for the regulatory environment, where certainty, rigor and reproducibility are necessary. However, Dr Tsaioun noted that our knowledge is constantly evolving and that there are large gaps in our current understanding of biology. Systematic methods, especially with the advent of machine learning (ML) and artificial intelligence (AI) techniques, have a potential to find such gaps and identify areas where scientific research may be directed, while doing so with transparency and minimal bias.

Systematic methodologies, including systematic maps and systematic reviews, aim to help assemble the current knowledge on a topic in an objective, comprehensive and transparent manner. The difference between a classical, narrative review and a systematic review lies in their structure. Narrative reviews are written by one or more experts, and are based on their opinions, expertise, and frequently an informal selection from the literature that they are familiar with. Systematic reviews, on the other hand, follow specific steps, starting with a specifically formulated research question (Morgan et al., 2018). Most systematic reviews

are structured around a PECO/PICO (Population, Exposure/Intervention, Comparator and Outcomes) question, which determines the inclusion and exclusion criteria and the strategy for the literature search.

The following are the main steps of the systematic review:

1. Identify a problem.
2. Formulate a specific PECO/PICO question.
3. Write and publish the protocol.
4. Search the evidence in the literature as broadly as possible.
5. Apply protocol-defined inclusion and exclusion criteria to screen the relevant literature.
6. Assess the risk of bias of the collected evidence.
7. Integrate the evidence found (qualitatively or quantitatively (as a meta-analysis)).

In short, systematic reviews are produced using a structured framework that assures transparency, objectivity and comprehensiveness. Such approaches are critical in regulatory settings, which is the reason for many agencies around the world, such as European Food Safety Authority (EFSA, 2010), National Toxicology Program at the National Institute of Environmental Health and Sciences (NIEHS) (Rooney et al., 2014; Thayer et al., 2014) and US Environmental Protection Agency (US EPA) (EPA, 2018), to adapt the systematic review frameworks for their assessments.

Systematic methodology could potentially also be used to develop and assess the certainty in an AOP and subsequently define and find the appropriate evidence that corresponds to specific MIE, KE, and especially the relationships between them (KER), in the process (Fig. 2). New AOPs are created in the AOP-Wiki, which serves as an international AOP repository. Currently, AOPs may be proposed by any researcher and may be listed as “under development: contributions and comments welcome” (OECD, 2017a). This method of soliciting expert knowledge into the AOP-Wiki, while making the initiation of a new AOP and comments on existing projects more democratic, could benefit from refinement to make sure that these initial contributions are evidence-based and are within the scope and feasibility of the AOP framework. Hence, work needs to be done to make it possible for the AOPs to be transparently and objectively updated as new biological KEs and pathways are discovered. Many of these questions have been wrestled with and solved in systematic frameworks developed in order to objectively summarize clinical research, and the AOP community can capitalize on these advancements to increase the objectivity, transparency and reproducibility of their work products.

Dr Tsaioun noted that there are many new ML- and AI-based tools that are currently being tested and validated for the most labor-intensive parts of systematic reviews such as literature search and screening (Howard et al., 2016; Tsafnat et al., 2013; Van der Mierden et al., 2019). The remaining challenge now is to transform the semi-automated processes into fully automated search and literature screening and data extraction tools. Eventually,

these tools will enable a systematic review process that can be updated in real time as new information becomes available. This will be highly valuable for applying evidence-based approaches to the development and evaluation of AOPs and mechanistic tests that measure KEs in the framework.

2.2 GRADE

Dr **Holger Schünemann**, chair and professor in the Department of Health Research Methods, Evidence and Impact at McMaster University, provided background on the GRADE framework and its applicability for assessing the certainty in AOPs. GRADE was established 20 years ago as an informal collaboration of people with an interest in addressing the shortcomings of assessment of confidence of various treatments in healthcare. GRADE has developed a practical and transparent approach to grading quality (or certainty) of evidence and strength of recommendations and is now considered the standard in the development of clinical practice guidelines. GRADE, which has 16 centers and networks around the globe to support users, is now used by more than 100 organizations around the world. GRADE is the method used in Cochrane[3] systematic reviews to assess certainty in the evidence that has been synthesized to answer a research question. GRADE is also used for the development of recommendations by an international group of diverse contributors from different backgrounds, including many public health professionals.

Although for many years the Bradford Hill (BH) criteria, developed in the 1960s (Hill, 1965) and then iterated by David Sackett and his colleagues (Guyatt et al., 1984; Tugwell et al., 1985), were considered the gold standard for assessing causality in epidemiology research, Dr Schünemann argued that over time this approach has become outdated. For example, publication bias (selective publishing of data, resulting in a skewed representation of a phenomenon in the published literature, e.g., a bias towards publishing positive toxicological outcomes) was not part of the criteria; and the criteria were not completely thought through with respect to association (rather than causation), the impact of interventions, and prognosis. These shortcomings have prevented the BH criteria from being used in exposure assessments. GRADE was developed to address these shortcomings (Schünemann et al., 2011). Additionally, GRADE has been extended to provide context for decision-making in the evidence-to-decision frameworks (Parmelli et al., 2017).

One important question during development of the GRADE framework was whether to focus on factors that make users more confident in a body of evidence or on factors that make them less confident. GRADE chose the latter approach, in which scientists look at scenarios in which confidence is high and then try to determine what factors could cause users to lose confidence in the evidence. This approach was selected as it was found to be easier in the majority of cases, as opposed to the BH considerations, where reasons to increase certainty are sought.

[3]Cochrane is a global independent network of researchers, professionals, patients, carers and people interested in health formed to organise medical research findings to facilitate evidence-based choices about health interventions. A Cochrane Review is a systematic review of research in health care and health policy conducted according to the guidelines developed by Cochrane and these reviews are internationally recognized as the highest standard in evidence-based health care (https://www.cochranelibrary.com/about/about-cochrane-reviews).

Dr Schünemann said there is still work to be done to develop GRADE, specifically with regard to the indirectness domain, which is also an important consideration in relation to AOPs. The indirectness domain assesses the extent to which the evidence found reflects the population, intervention and outcomes of interest. GRADE has begun using AOPs to fill some of the gaps that exist in the indirectness domain as well as considerations regarding biological plausibility. The key take-home message in regard to AOPs, he said, is that AOPs mainly enable researchers to ask the right questions. For example, AOPs allow researchers to explore whether or not there is a direct connection between chemical exposure and adverse outcomes and support some judgments of indirectness. These provide the rationale to move to human studies and evaluate human evidence.

2.3 Modeled evidence

Dr **Jan Brożek**, associate professor in the Department of Health Research Methods, Evidence and Impact at McMaster University, proposed that the models used by researchers need to be formalized and become quantitative mathematical representations of reality. Models may be simple or much more complex, such as economic models and system dynamics in infectious diseases. Scientists and regulators alike need to be able to assess the certainty, or the quality, of the outputs.

“Certainty of evidence” refers to how much the results can be trusted and how much the results can help provide evidence in decision-making. There are three levels on which this may be assessed: (1) certainty in the inputs, (2) certainty in the model itself, and (3) certainty of the output from the model. “Certainty of modeled evidence” refers to the output and is the most important piece for decision-making. In order to assess this, it must be understood how certain researchers are about inputs and how certain they are about the model.

Given any piece of evidence, Dr Brożek said, there are factors that increase or decrease how much it can be trusted. The five factors used within GRADE that decrease certainty are: risk of bias, indirectness, inconsistency, imprecision and publication bias (Morgan et al., 2016b). The three factors that increase certainty are: large effect, dose response and opposing residual confounding.

The criteria that an ideal model needs to include should be generated first, and then existing models should be systematically searched to find one as close as possible. If such a model is not found, a novel model must be developed, or modeling must be forgone.

Choosing a model implies making certain assumptions, the most important of which is whether the new model can be assessed to be better than existing models. If there is an existing model, certainty of outputs must be addressed; this certainty is dependent on the inputs of the model and the model itself. It is due to these assumptions that GRADE suggests using systematic searching to find existing models instead of developing models. Creating a custom model with every interaction known included would be ideal, but toxicology as a field has many unknowns, which makes it difficult to create a model from scratch.

In summary, Dr Brożek proposed the following four steps in choosing a model:

- Perform a systematic literature search with specific search criteria.
- Choose the “best” model and use it as is.
- If this is not possible, adapt an existing model that you believe will be the best.
- If no existing model exists, then build your own model.

Further work is necessary to develop the tools to determine the certainty in a model (Brozek et al., 2021).

2.4 Systematic maps

Dr **Michelle Angrish**, a toxicologist at the US Environmental Protection Agency, provided an overview of “Systematic Maps and Literature-Driven AOP Development.” Dr Angrish stressed that systematic methods could be leveraged to support a more data-driven approach to mechanistic evidence integration using an AOP framework. Dr Angrish described the AOP analytical construct and how it is currently used to link a sequential chain of causally linked events at different levels of biological organization using the AOP components MIE, KE, KER, and AO, according to the Organisation for Economic Cooperation and Development (OECD) guidance (OECD, 2017a). The current AOP development approach is primarily an expert-driven process that could benefit from systematic methods, particularly with respect to increasing transparency and reproducibility. In particular, researchers could use evidence-based systematic methods to explore direct and indirect connections between outcomes of regulatory concern and the AOP analytical construct (i.e., KE-KER(s), etc.).

Dr Angrish next introduced systematic review methods in environmental health sciences and provided an example of how systematic review workflows and systematic mapping methods could be used to formulate a literature search, screen studies, and extract data (i.e., test system/species, exposure methods, experimental design, endpoints, chemicals, etc.). In the example, Dr Angrish demonstrated how the overall workflow could be formulated around a particular molecular target and outcome of interest. Experimental studies and extracted data organized into literature inventories can be mapped to the AOP schema, but a challenge lies in the semantics.

These inventories of environmental effect findings preserve the original written language as reported in the PDF documents, which presents a natural language processing challenge for several reasons. First, natural language is not machine-readable, making it difficult to make information automatically interoperable. Natural language requires humans to first sort out redundant or ambiguous (homonyms, polysemes, homophones, etc.) terms before labeling extracted information as the same. Second, humans must then train computers to digitize this information using controlled vocabularies (such as Universal Medical Language System). However, while these “digital” vocabularies include definitions and synonyms, they lack the relationships between terms found in ontologies that are needed to identify connections between exposure and effect findings that might have otherwise been missed. Dr Angrish concluded that mapping the digitized data using an AOP ontology model is

a possible solution (and an active area of research) with the added advantage that this approach combines expert- and data-driven evidence integration within an AOP framework.

2.5 Assessing risk of bias of *in vitro* studies

Dr **Andrew Rooney**, the acting director of the Office of Health Assessment and Translation (OHAT) at the National Toxicology Program (NTP), discussed the "Risk of Bias Appraisal of *In Vitro* Studies." His presentation focused on study quality or risk of bias appraisal for *in vitro* studies – a type of study that obviously forms an important part of the mechanistic studies that may be considered within an AOP.

There are four major aspects to be considered in the critical appraisal of any study:

- Internal validity (risk of bias): whether the study design and conduct may bias the results.
- External validity (applicability/generalizability): the extent to which the study addresses the research question or the review question.
- Reporting quality: the adequacy of reporting of the design, conduct and results of a study. This may be independently assessed or addressed as part of the assessment of internal validity.
- Sensitivity: whether or not the study design impacts the ability of that study to detect an effect (i.e., if only high doses were used and low doses are of interest, it will not be relevant to your research question).

In addition to the apical endpoint studies that need to be critically appraised, mechanistic data also need to be analyzed. Mechanistic data generally focus on upstream indicators and may be part of a larger study that investigated the health effect of chemical exposure, but the data are needed to illuminate the AOP. Mechanistic data can be found from a wide variety of study types: animal or human *in vivo* studies, *in vitro* studies of either animal or human cells, *in silico* or modeled data.

In terms of assessing risk of bias, existing approaches for *in vivo* study designs are effective. A number of tools have been designed to evaluate experimental animal studies including, but not limited to, the SYRCLE tool (Hooijmans et al., 2014), the Navigation Guide (Lam et al., 2014), and the NTP/OHAT tool (Rooney et al., 2014). As regulators are shifting towards using more *in vitro* studies, new tools and approaches are currently being developed as the approaches for *in vivo* studies are not fully applicable. Prominent approaches include the NTP/OHAT "use case"[4], Science in Risk Assessment and Policy (SciRAP) (currently in a pilot phase) (Molander et al., 2015), the EPA/IRIS Handbook and NTP (Rooney et al., 2014) approaches.

The OHAT approach (originally established and published for human studies and laboratory animal toxicology studies) begins with a set of ten basic questions and an eleventh question addressing additional threats to validity:

[4]Protocol to evaluate the evidence for an association between perfluorooctanoic acid (PFOA) or perfluorooctane sulfonate (PFOS) exposure and immunotoxicity. https://ntp.niehs.nih.gov/ntp/ohat/pfoa_pfos/protocol_201506_508.pdf

1. Was administered dose or exposure level adequately randomized?
2. Was allocation to study groups adequately concealed?
3. Did selection of study participants result in the appropriate comparison groups?
4. Did study design or analysis account for important confounding and modifying variables?
5. Were experimental conditions identical across study groups?
6. Were the research personnel blinded to the study group during the study?
7. Were outcome data complete without attrition or exclusion from analysis?
8. Can we be confident in the exposure characterization?
9. Can we be confident in the outcome assessment?
10. Were all measured outcomes reported?
11. Were there no other potential threats to internal validity?

The study design determines which questions are applicable. The evaluation is endpoint-specific (so the questions are answered multiple times for studies with multiple endpoints), and the rating is on a 4-point scale.

In order to apply the OHAT risk of bias approach to *in vitro* studies, the criteria needed to be adapted[4]. The overall goal was to consider all of the different evidence streams in a parallel manner and adapt their concepts within a risk of bias approach. Following many iterations, NTP was able to develop a new model that mimicked the original OHAT 10+1 risk of bias questions.

Consider, for example, Question 1: “Was the administered dose randomized?” In human and animal studies, this question is easy to answer. In an *in vitro* study, researchers need to see whether the dose was given selectively to individual cells or tissues. Each cell needs to have an equal chance of being assigned to each study group. If a homogeneous cell suspension was used, then, by definition, there is no variation, and, as such, Question 1 will either not apply or indicate that there is a low risk of bias.

Even though the focus has been on risk of bias, there are other concerns to address with a study appraisal tool, including external validity. In environmental health, animal and human data need to come together in order to reach a hazard and risk decision. A combination of both bodies of evidence allows scientists to look at human observational studies to find a potential hazard and then use animal models to demonstrate causation. In the risk assessment context, it has been common to combine human and animal studies in this way.

In the OHAT method, mechanistic data, including *in vitro* studies, are considered together to inform the biological plausibility. Scientists may conclude that the chemical is a “presumed” hazard to humans, upgraded to being “known” if they are confident in the other evidence, or downgraded to “suspected” if they have less confidence.

EBTC has been working with OHAT and other stakeholders on harmonizing such agency-specific approaches based on the fundamental principles of systematic review, as summarized in the primer by Hoffmann et al. (2017). The authors of this primer represent major stakeholders in the field of toxicology, risk assessment and systematic review communities. There are a number of projects EBTC is coordinating and co-leading that focus on risk-of-bias of *in vitro* studies, done in collaboration with OHAT, US EPA, EFSA and other stakeholders.

2.6 Integrated approaches to testing and assessment

The plenary session was concluded by Dr **Patience Browne**, a policy analyst in the Test Guidelines Programme at the OECD, who discussed "IATA and Alternative Approaches Based on AOPs". Though, as Dr Browne indicated, AOPs are a relatively recent concept in the field of toxicology, they are an evolution of the preceding "mode of action" and "toxicity pathway" concepts useful for organizing diverse data in a codified way. The AOP framework also can be applied to map test guideline endpoints used to test the toxicity of a chemical to KEs in order to build predictive models, develop integrated approaches to testing and assessment (IATA), and guide logical next steps (or draw conclusions) for evaluating chemical safety (OECD, 2016).

IATAs range from highly flexible, informal approaches that rely heavily on expert judgement to prescriptive, structured approaches that rely on pre-defined rules to reach conclusions. Once the flexibility of an IATA has been removed and the information sources (e.g., *in silico*, *in chemico*, *in vitro*, *in vivo* methods) and data interpretation procedures are fixed, the IATA becomes a defined approach (DA) (OECD, 2017b). In a DA, expert judgement is removed and independent users of the DA should come to the same conclusion for the same chemical. A "complete" AOP (including a MIE, all KEs, all KERs, and an AO) is not necessary for IATA development. Rather, the appropriate degree of completeness of the AOP will vary highly with the problem formulation. Three examples were provided to illustrate how AOPs can be used to guide IATA development under different circumstances.

In the first example, Dr Browne discussed human skin sensitization, which is a relatively simple and well-understood toxicological process supported by a complete AOP, with multiple guideline methods that measure all KEs. There are also approximately 130 chemicals with data from multiple *in vitro* assays, *in vivo* mouse assays, and humans (Hoffmann et al., 2018). Collectively, this information has been used to develop DAs to predict with high sensitivity, specificity and accuracy the rodent and human skin sensitization response from *in silico* and *in vitro* data (Kleinstreuer et al., 2018). The non-animal methods combined in DAs are sufficiently predictive that results of DAs are accepted as replacements for animal data in some regulatory contexts (e.g., US EPA).

In the second example Dr Browne described, AOPs are used to help organize and integrate data for evaluating endocrine disruptors. International regulatory requirements for demonstrating the endocrine disrupting potential of a chemical vary by region, but all include linking an adverse effect to an endocrine MoA. As most available methods do not measure both of these types of data in a single assay, evaluating endocrine activity requires integrating evidence from multiple assays. While there are very few "complete"

AOPs for endocrine disruption, the MIEs and early KEs are well understood for sex steroid signaling pathways. The US EPA developed an estrogen receptor model based on data from 18 orthogonal *in vitro* assays that measure interactions with the estrogen receptor at various points in the signaling pathway using different technologies (Judson et al., 2015). The predictivity of the model has been evaluated using *in vitro* and *in vivo* reference chemicals and is sufficiently reliable to screen chemicals for *in vitro* activity, prioritize chemicals for further testing, and replace the requirement for the rodent uterotrophic bioassay data (Browne et al., 2015; Kleinstreuer et al., 2016).

The third example was an integrated approach for evaluating chemicals with respect to human carcinogenicity hazard. From a retrospective analysis of human carcinogens identified by the International Agency for Research on Cancer, 10 key characteristics were proposed for identifying and grouping mechanistic data to assess potential carcinogens (Smith et al., 2016). The key characteristics were not identified or organized using AOP frameworks, and, in fact, building “complete” AOPs or AOP networks to capture the various MIEs and KEs represented by the key characteristics would take years. Nonetheless, the key characteristics provide information on several levels of biological organization, and along with animal experimental and human epidemiological data, can be organized using an AOP construct without understanding all intermediate events or causal relationships. Such an approach has been proposed as a way to organize mechanistic data to predict therapeutic response of personalized cancer therapies (Morgan et al., 2016a).

These three examples range from a complete AOP, to a “partial” AOP with well-understood MIEs and early KEs, and a complex, multi-factorial pathway network of (in most cases) poorly understood events that lead to adverse outcomes. In these three examples, the ability to predict outcomes, replace the need for additional testing, or prioritize chemicals for additional testing vary. In general, the more complete the AOP or the greater the understanding of the underlying biology, the more confidence there is for predicting adverse responses from earlier KEs. However, an AOP framework can be useful for integrating evidence, even for circumstances in which the AOP is largely undescribed.

3 Break-out groups

Following the presentations, participants divided into four break-out groups to discuss the four workshop themes in more detail.

3.1 Application of risk of bias tools to a study

This group worked through a case study where they applied the OHAT Risk of Bias (RoB) tool[4] to a single study, also comparing it with the SciRAP tool (Molander et al., 2015; Beronius et al., 2018), the only other generalized tool currently available for assessing RoB of *in vitro* studies. Participants were provided with a copy of the study to be assessed and a copy of the OHAT RoB tool prior to the meeting. Many participants were already familiar with RoB in general, as well as with the challenges in applying RoB to *in vitro* studies. The practical experiences of the participants, including the use of SciRAP and the US EPA IRIS approach, were shared at the start of the break-out. Applying the individual criteria from the OHAT RoB tool often prompted more general discussions about, for example,

the relevance of these criteria for critical appraisal of *in vitro* studies, the role of *in vitro* studies in a risk assessment context and therefore the importance of performing a full RoB assessment, etc. A clear and recurring theme throughout the break-out was the difficulty in developing a "one-size-fits-all" type of approach that addresses all aspects of study validity for all types of *in vitro* studies and can be easily recognized/employed by both systematic review methodologists and bench scientists. The range of methodological aspects that could be associated with the potential for systematic error (risk of bias) was considered broad. Moreover, some methodological aspects were not regarded as internal validity, but rather as construct or external validity (noting that terminology is not well-defined or consistently recognized among practitioners). There was consensus, however, that multiple aspects were important to assessing overall validity – and trying to fit these into a RoB tool focused only on internal validity was difficult and likely to be insufficient for the needs of a practitioner. Furthermore, the need to assess aspects other than internal validity on an individual study basis introduces the need for refined workflows, as these aspects are not commonly accounted for (on an individual study basis) in a standard systematic review process. Various approaches were discussed to address this issue, including the use of inclusion/exclusion criteria, as well as signaling and prompting questions. The break-out group agreed, however, that flexibility to the topic, level of granularity, and the decision of when to do a RoB on *in vitro* data were important concepts to consider. There was also consensus on the importance of subject matter expertise in refining questions, conducting appraisals, and resolving conflict. In closing, the workgroup recognized that available tools have a fair bit of overlap within elements of internal validity, but more method development and practical experience is needed to refine and develop new tools and processes for assessing (all aspects of) the validity of *in vitro* studies in a systematic review.

3.2 Building evidence-based AOPs

The discussion focused on building AOPs in an evidence-based way. To that end, the participants were presented with the AOP concept and schema and then discussed what sort of information should be sought if AOPs were to be constructed using a literature-driven (rather than expert-driven) approach. Major challenges to applying GRADE practitioners' experience in literature-driven approaches to the AOP context include the use of a different technical vocabulary by AOP developers (e.g., MIE, KE, essentiality) and GRADE practitioners (e.g., evidence profile, inconsistency, indirectness), and the processes by which an AOP is formalized. AOP development involves the use of specialized vocabulary that is not necessarily intuitive to an external audience, partly because it relates to the technical governance processes by which AOPs are reviewed and approved. Also, AOP review and approval involves a process that is not clearly transparent to the GRADE practitioner, because it does not involve methods that are obviously reproducible and explicitly developed to minimize the potential RoB in assembling and interpreting the evidence. The key concepts from the AOP technical vocabulary (e.g., KER, essentiality and biological plausibility) as well as the key steps in the development and use of AOPs (e.g., submission to and assessment in the OECD AOP-Wiki) were discussed. Since indirectness is one of the GRADE domains for assessment of certainty, there was an intuition in the group that the GRADE framework might be applicable in the AOP space; however, developing a

comprehensive and precise understanding of the AOP development process, concepts and vocabulary seems essential if progress is to be made.

A key conclusion from the group discussion was related to the fundamental unit of an AOP – the KER. Since relationships between KEs should be evaluable in the same way as an exposure-outcome pair (both are putatively causally-related pairs of events), an AOP should be evaluable in the same way as the exposure-outcome relationships from conventional systematic reviews. If so, then putative AOPs could be treated as a series of systematic reviews of KERs, with certainty evaluated using a suitably adapted GRADE approach for each event-event link. Certainty in the overall AOP would be a function of certainty in each individual KER (illustrated in Fig. 3; cf. Collier et al., 2016).

The AOP development process does not currently require use of systematic review methods, nor have any systematic review approaches been noted in the AOPs submitted for review thus far. The group observed that the benefits of systematic review and GRADE criteria for assessing certainty in an AOP seem intuitively applicable, could be of significant benefit, and therefore should be further explored. In addition to AOP development, the OECD approval process for AOPs could potentially benefit from the transparency and consistent operationalization of the GRADE approach as well.

3.3 Certainty in AOPs

The discussion's purpose was to examine criteria on which to base the assessment of the certainty of evidence used in AOPs. It was noted that in order to comprehensively understand one's certainty in the MoA depicted in an AOP framework, certainty in the framework depends on the underlying evidence used to establish each KER. Using example AOP frameworks related to diabetes and bladder cancer, the group discussed processes to provide structure and transparency to the presentation of AOPs, and to qualify the uncertainty of the evidence for KERs within an AOP. Participants considered two approaches for structuring the AOP development and evaluation process: a top-down approach and a bottom-up approach. Both processes consider the overarching research question, identifying the population, exposure, comparison, and outcomes of interest (PECO), as well as any available evidence that can be used to inform the framework (e.g., peer-reviewed publications, meta-data, public databases, AOP-Wiki, etc.) at the KER level or across several KERs. Participants emphasized the importance of transparency throughout the process for evaluating certainty within AOPs, especially when there is limited or no evidence between KE pairs.

When understanding the certainty of the underlying evidence, the perspective of the end-user should be taken into account: Does the certainty relate to the entire AOP model in order to inform questions about an exposure/population outcome relationship or to elements within the AOP to understand the uncertainty of a relationship between KEs? When considering the relationship between KEs, participants established the following elements for assessment: study limitations (RoB), inconsistency (unexplained heterogeneity between individual studies), indirectness (how well the evidence reflects the question asked), imprecision, and publication bias. No additional domains were identified. The identified factors strongly suggest that current evidence assessment domains used within the GRADE

framework (Guyatt et al., 2011) would be sufficient for the evidence assessment within AOPs (content validation).

3.4 AOPs to inform NAM development

The group discussed how AOPs can inform the development of NAMs, especially non-animal experimental studies. As a starting point for the discussion, the participants agreed and stressed that AOPs are a simplified depiction of complex biology. As a consequence, AOPs may leave out KEs, other molecular mechanisms leading to the same apical response, or other adverse responses that are a consequence of the same MIE. Thus, the confidence in a conclusion that a chemical is “toxic” is higher than in a conclusion that a chemical is “non-toxic”.

One important application of an AOP is to support the development of IATAs (OECD, 2016). An IATA can be developed for different purposes, such as chemical prioritization or replacement of animal tests. If AOPs are used to support the development of IATAs intended to substitute for *in vivo* data, the nature of the AOP and the KEs measured in the IATA must be considered. For example, MIEs may be easier to measure, but they are further removed from (and may be less predictive of) the AO, while events close to the AO are likely more predictive but may be difficult to measure with non-animal methods. Furthermore, if two KEs are highly dependent, the question arises if both need to be addressed experimentally (see, e.g., van Vliet et al., 2018). While complete understanding of all events in an AOP is a goal, organizing evidence in an AOP is informative in itself, as it may illustrate gaps in knowledge (Leist et al., 2017). A systematic assessment of the relevance of KEs and KERs – using evidence-based approaches including a comprehensive literature search, critical appraisal of individual studies, and a systematic integration of all relevant information – can be used to evaluate the certainty/confidence in the understanding of the biology linking the mechanism to intermediate responses and apical effects. A question raised in this regard was how certainty/confidence in the final decision can be expressed as a function of confidence in KE, KER, and the NAMs modelling these elements.

The group concluded that AOPs as a representation of knowledge may be used to guide development of NAMs to (a) strengthen the evidence base for the relevance of KEs, (b) develop batteries of assays to measure a specific type of toxicity or non-specific toxicity, (c) target KEs (i.e., nodes) common to several AOP, or (d) target specific mechanisms. Development of NAMs for use in a regulatory context should consider how resulting information is to be used. For example, it may not be necessary to develop AOPs that include all possible AOs to conclude a chemical may have adverse effects. However, to be confident in a negative identification, the AOP network coverage would need to include all relevant biological space.

4 Plenary concluding discussion

Due to the requirements of the regulatory environment and for historical reasons, laboratory animal models continue to serve as the standard in toxicology in order to extrapolate to human outcomes in the absence of direct human data. But it remains an indirect approach and, thus, offers a less than perfect prediction of human outcomes. Since neither *in vivo*

animal nor *in vitro* mechanistic data are considered purely predictive of human outcomes at this point, various types of mechanistic evidence should be combined within an assessment framework to provide robust, high-confidence outcomes. AOPs provide such an integrated framework but have traditionally been defined through an expert review process that can be inconsistent and incomplete. Applying systematic methodologies to AOP development would result in higher certainty in the evidence evaluation underpinning each AOP, which would then increase the utility of that AOP for supporting regulatory decisions.

The AOP framework was designed to provide a transparent structure to organize knowledge/data and can function as the scientific basis for toxicity extrapolations via the underlying mechanisms of toxicity. AOPs can serve as a guide for new assay development to fill important gaps and add information about measurable KEs leading to AOs. Retrospective systematic literature reviews of *in vivo* and *in vitro* mechanistic studies promise a way to provide a bridge to human outcomes as well as increase the transparency, consistency and objectivity in the development and assessment of AOPs as well as the resulting toxicity predictions.

AOPs combine *in vitro* or *in vivo* data by mapping the available data to the associated KE within the biological mechanism. Therefore, the framework naturally integrates different types of data, but it does not assess the individual types of data or provide guidance on how the information from different types of data should be combined. Systematic methodologies can fill this gap within current AOP development practices. The first ideas on how to assess the RoB in *in vitro* (mechanistic) studies and how to apply (the domains from) GRADE to AOPs, notably to KERs, are being developed, but further research is necessary to establish the details of the required tools (Box 1).

As many individuals and organizations do not have the resources to delve into large and growing literature databases, the development of broad systematic evidence maps has the potential to allow far more researchers to use the scientific literature effectively and to formulate specific questions that can be answered by systematic reviews (Wolffe et al., 2019). In order to construct such systematic evidence maps, the scientific community needs to restructure the way they conduct, report and publish results. At the moment, the primary raw data are stripped from experiments as they are prepared for publication. These data need to be accessible to anyone and kept as machine-readable meta-data, rather than being lost. Systematic evidence maps built on such transparent and complete data would allow researchers to find gaps in the knowledge and notice novel associations that were not apparent before. The beauty of the AOP paradigm is that it provides a framework for capturing the extracted information, but the challenge is identifying and ascertaining the relationships between KEs in order to place them within the framework. Systematic evidence maps and systematic reviews would be an invaluable tool in this process.

In conclusion, combining systematic review methods and AOP concepts seems to be a logical next step in further developing and modernizing the regulatory testing paradigm with new science, including NAMs. However, further conceptual and methodological research and technical developments are necessary to fully realize this potential. The workshop participants agreed to continue the discussions in a series of workshops and case studies

that would explore ways to use systematic methods in the process of building and certainty assessment of AOPs and associated test methods.

Acknowledgements

We would like to thank Elisa Aiassa (European Food Safety Authority) for co-leading the break-out group on risk of bias in *in vitro* studies and Miroslav Klugar and Jitka Klugarova (visiting scholars McMaster) for their active participation in the break-out groups.

Abbreviations

AI	artificial intelligence
AO	adverse outcome
DA	defined approach
EBTC	Evidence-Based Toxicology Collaboration
EFSA	European Food Safety Authority
GRADE	Grading of Recommendations Assessment, Development and Evaluation
IATA	integrated approaches to testing and assessment
KE	key event
KER	key event relationship
MIE	molecular initiating event
ML	machine learning
MoA	mode of action
NIEHS	National Institute of Environmental Health and Sciences
NTP	National Toxicology Program
OECD	Organisation for Economic Cooperation and Development
OHAT	Office of Health Assessment
PECO/PICO	Population, Exposure/Intervention, Comparator and Outcomes
RoB	risk of bias
SciRAP	Science in Risk Assessment and Policy
US EPA	US Environmental Protection Agency

References

Beronius A, Molander L, Zilliacus J et al. (2018). Testing and refining the science in risk assessment and policy (SciRAP) web-based platform for evaluating the reliability and relevance of in vivo toxicity studies. J Appl Toxicol 38, 1460–1470. doi:10.1002/jat.3648 [PubMed: 29806706]

Browne P, Judson RS, Casey WM et al. (2015). Screening chemicals for estrogen receptor bioactivity using a computational model. Environ Sci Technol 49, 8804–8814. doi:10.1021/acs.est.5b02641 [PubMed: 26066997]

Brozek JL, Canelo-Aybar C, Akl EA et al. (2021). GRADE guidelines 30: The GRADE approach to assessing the certainty of modeled evidence-an overview in the context of health decision-making. J Clin Epidemiol 129, 138–150. doi:10.1016/j.jclinepi.2020.09.018 [PubMed: 32980429]

Collier ZA, Gust KA, Gonzalez-Morales B et al. (2016). A weight of evidence assessment approach for adverse outcome pathways. Regul Toxicol Pharmacol 75, 46–57. doi:10.1016/j.yrtph.2015.12.014 [PubMed: 26724267]

EFSA (2010). Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA J 8, 1637. doi:10.2903/j.efsa.2010.1637

EPA (2018). Application of Systematic Review in TSCA Risk Evaluations. https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/application-systematic-review-tsca-risk-evaluations

Griesinger C, Hoffmann S, Kinsner A et al. (2008). The emerging concept of evidence-based toxicology (EBT) – Results of the 1st international forum towards EBT. ALTEX 25, Suppl 1. https://www.altex.org/index.php/altex/article/view/2211/2150

Groh KJ, Carvalho RN, Chipman JK et al. (2015). Development and application of the adverse outcome pathway framework for understanding and predicting chronic toxicity: I. Challenges and research needs in ecotoxicology. Chemo-sphere 120, 764–777. doi:10.1016/j.chemosphere.2014.09.068

Guyatt G, Oxman AD, Akl EA et al. (2011). Grade guidelines: 1. Introduction-grade evidence profiles and summary of findings tables. J Clin Epidemiol 64, 383–394. doi:10.1016/j.jclinepi.2010.04.026 [PubMed: 21195583]

Guyatt GH, Webber CE, Mewa AA et al. (1984). Determining causation – A case study: Adrenocorticosteroids and osteoporosis: Should the fear of inducing clinically important osteoporosis influence the decision to prescribe adrenocorticosteroids? J Chronic Dis 37, 343–352. doi:10.1016/0021-9681(84)90100-0 [PubMed: 6371037]

Hill AB (1965). The environment and disease: Association or causation? Proc R Soc Med 58, 295–300. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1898525/pdf/procrsmed00196-0010.pdf [PubMed: 14283879]

Hoffmann S, de Vries RBM, Stephens ML et al. (2017). A primer on systematic reviews in toxicology. Arch Toxicol 91, 2551–2575. doi:10.1007/s00204-017-1980-3 [PubMed: 28501917]

Hoffmann S, Kleinstreuer N, Alepee N et al. (2018). Non-animal methods to predict skin sensitization (I): The Cosmetics Europe database. Crit Rev Toxicol 48, 344–358. doi:10.1080/10408444.2018.1429385 [PubMed: 29474128]

Hooijmans CR, Rovers MM, de Vries RBM et al. (2014). Syrcle's risk of bias tool for animal studies. BMC Med Res Methodol 14, 43. doi:10.1186/1471-2288-14-43 [PubMed: 24667063]

Howard BE, Phillips J, Miller K et al. (2016). Swift-review: A text-mining workbench for systematic review. Syst Rev 5, 87. doi:10.1186/s13643-016-0263-z [PubMed: 27216467]

Jadad AR, Moore RA, Carroll D et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clinl Trials 17, 1–12. doi:10.1016/0197-2456(95)00134-4

Judson RS, Magpantay FM, Chickarmane V et al. (2015). Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. Toxicol Sci 148, 137–154. doi:10.1093/toxsci/kfv168 [PubMed: 26272952]

Kleinstreuer NC, Ceger PC, Allen DG et al. (2016). A curated database of rodent uterotrophic bioactivity. Environ Health Perspect 124, 556–562. doi:10.1289/ehp.1510183 [PubMed: 26431337]

Kleinstreuer NC, Hoffmann S, Alepee N et al. (2018). Non-animal methods to predict skin sensitization (II): An assessment of defined approaches*. Crit Rev Toxicol 48, 359–374. doi:10.1080/10408444.2018.1429386 [PubMed: 29474122]

Lam J, Koustas E, Sutton P et al. (2014). The navigation guide – Evidence-based medicine meets environmental health: Integration of animal and human evidence for PFOA effects on fetal growth. Environ Health Perspect 122, 1040–1051. doi:10.1289/ehp.1307923 [PubMed: 24968389]

Leist M, Ghallab A, Graepel R et al. (2017). Adverse outcome pathways: Opportunities, limitations and open questions. Arch Toxicol 91, 3477–3505. doi:10.1007/s00204-017-2045-3 [PubMed: 29051992]

Meek ME (2014). Evolution of mode of action/adverse outcome pathway analyses. Toxicol Lett 229, S10. doi:10.1016/j.toxlet.2014.06.063

Molander L, Ågerstrand M, Beronius A et al. (2015). Science in risk assessment and policy (SciRAP): An online resource for evaluating and reporting in vivo (eco)toxicity studies. Hum Ecol Risk Assess 21, 753–762. doi:10.1080/10807039.2014.928104

Morgan MM, Johnson BP, Livingston MK et al. (2016a). Personalized in vitro cancer models to predict therapeutic response: Challenges and a framework for improvement. Pharmacol Ther 165, 79–92. doi:10.1016/j.pharmthera.2016.05.007 [PubMed: 27218886]

Morgan RL, Thayer KA, Bero L et al. (2016b). GRADE: Assessing the quality of evidence in environmental and occupational health. Environ Int 92–93, 611–616. doi:10.1016/j.envint.2016.01.004

Morgan RL, Whaley P, Thayer KA et al. (2018). Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. Environ Int 121, 1027–1031. doi:10.1016/j.envint.2018.07.015 [PubMed: 30166065]

OECD (2016). Guidance Document for the Use of Adverse Outcome Pathways in Developing Integrated Approaches to Testing and Assessment (IATA). OECD Series on Testing and Assessment, No. 260. OECD Publishing, Paris. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2016)67&doclanguage=en

OECD (2017a). Revised Guidance Document on Developing and Assessing Adverse Outcome Pathways. OECD Series on Testing and Assessment, No. 184. OECD Publishing, Paris. https://bit.ly/2ZzVSw0

OECD (2017b). Guidance Document on the Reporting of Defined Approaches to be Used Within Integrated Approaches to Testing and Assessment. OECD Series on Testing and Assessment, No. 255. OECD Publishing, Paris. doi:10.1787/9789264274822-en

Parmelli E, Amato L, Oxman AD et al. (2017). GRADE evidence to decision (EtD) framework for coverage decisions. Int J Technol Assess Health Care 33, 176–182. doi:10.1017/s0266462317000447 [PubMed: 28655365]

Rooney AA, Boyles AL, Wolfe MS et al. (2014). Systematic review and evidence integration for literature-based environmental health science assessments. Environ Health Perspect 122, 711–718. doi:10.1289/ehp.1307972 [PubMed: 24755067]

Rooney AA, Cooper GS, Jahnke GD et al. (2016). How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. Environ Int 92–93, 617–629. doi:10.1016/j.envint.2016.01.005

Scholten RJ, Clarke M and Hetherington J (2005). The Cochrane Collaboration. Eur J Clin Nutr 59, Suppl 1, S147–149; discussion S195–146. doi:10.1038/sj.ejcn.1602188 [PubMed: 16052183]

Schünemann H, Hill S, Guyatt G et al. (2011). The GRADE approach and Bradford Hill's criteria for causation. J Epidemiol Community Health 65, 392–395. doi:10.1136/jech.2010.119933 [PubMed: 20947872]

Smith MT, Guyton KZ, Gibbons CF et al. (2016). Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. Environ Health Perspect 124, 713–721. doi:10.1289/ehp.1509912 [PubMed: 26600562]

Stephens ML, Betts K, Beck NB et al. (2016). The emergence of systematic review in toxicology. Toxicol Sci 152, 10–16. doi:10.1093/toxsci/kfw059 [PubMed: 27208075]

Thayer KA, Wolfe MS, Rooney AA et al. (2014). Intersection of systematic review methodology with the NIH reproducibility initiative. Environ Health Perspect 122, A176–177. doi:10.1289/ehp.1408671 [PubMed: 24984224]

Tsafnat G, Dunn A, Glasziou P et al. (2013). The automation of systematic reviews. BMJ 346, f139. doi:10.1136/bmj.f139 [PubMed: 23305843]

Tsaioun K, K. SA (2010). ADMET for Medicinal Chemists: A Practical Guide. John Wiley & Sons, Inc. doi:10.1002/9780470915110

Tugwell P, Bennett KJ, Sackett DL et al. (1985). The measurement iterative loop: A framework for the critical appraisal of need, benefits and costs of health interventions. J Chronic Dis 38, 339–351. doi:10.1016/0021-9681(85)90080-3 [PubMed: 3923014]

Van der Mierden S, Tsaioun K, Bleich A et al. (2019). Software tools for literature screening in systematic reviews in biomedical research. ALTEX 36, 508–517. doi:10.14573/altex.1902131 [PubMed: 31113000]

van Vliet E, Kuhnl J, Goebel C et al. (2018). State-of-the-art and new options to assess T cell activation by skin sensitizers: Cosmetics Europe workshop. ALTEX 35, 179–192. doi:10.14573/altex.1709011 [PubMed: 28968481]

Vandenberg LN, Ågerstrand M, Beronius A et al. (2016). A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. Environ Health 15, 74. doi:10.1186/s12940-016-0156-6 [PubMed: 27412149]

Villeneuve DL, Crump D, Garcia-Reyero N et al. (2014). Adverse outcome pathway development II: Best practices. Toxicol Sci 142, 321–330. doi:10.1093/toxsci/kfu200 [PubMed: 25466379]

Whaley P, Halsall C, Ågerstrand M et al. (2016). Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations. Environ Int 92–93, 556–564. doi:10.1016/j.envint.2015.11.002

Wolffe TAM, Whaley P, Halsall C et al. (2019). Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. Environ Int 130, 104871. doi:10.1016/j.envint.2019.05.065 [PubMed: 31254867]

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. Environ Health Perspect 122, 1007–1014. doi:10.1289/ehp.1307175 [PubMed: 24968373]

Box 1:

Key messages

The *AOP framework* represents a significant development in how mechanistic information is assembled for safety assessment in toxicology and environmental health:

- AOPs allow the integration of human data, animal data, and data from new approach methods (NAMs) within a single framework.
- AOPs are developed using a transparent framework. To support this, the AOP KnowledgeBase was developed, which includes two modules: the AOP-Wiki and Effectopedia.
- The problem for which an AOP is used will determine the level of detail required when describing the molecular initiating event, key events, adverse outcomes, and the relationships among them. In every case, however, systematically and transparently assembled literature data could support the assessment of the certainty in AOPs.

A systematic review is a transparent, traceable, reproducible summary of the results from primary studies. Specific methods and tools that could facilitate AOP development were identified:

- Automated text mining and machine learning tools allow for systematic literature mapping to occur at a much faster pace and will soon allow for real-time updates.
- Individual studies assessment: The original Office of Health Assessment and Translation (OHAT) risk-of-bias tool for *in vivo* studies has been adapted in order to assess a potential for bias in *in vitro* primary studies.
- Bodies of evidence assessments: GRADE (Grading of Recommendations, Assessment, Development and Evaluations) framework, which was developed to assess certainty in clinical evidence and is now being adapted for environmental health, operationalized the Bradford Hill (BH) criteria in the validated GRADE framework.

The workshop identified a key barrier to communication between the two fields: terminology. Workshops like this will help harmonize the terminology, resolve misunderstandings and inspire collaboration to collectively advance the fields of AOP development and systematic review.

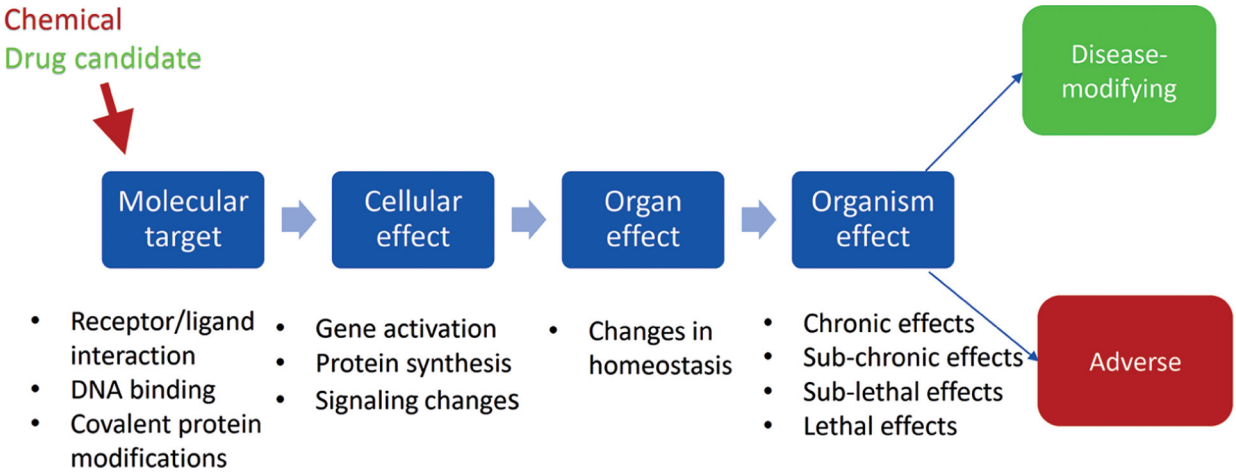

Fig. 1:
Mode of action (MoA) pathway

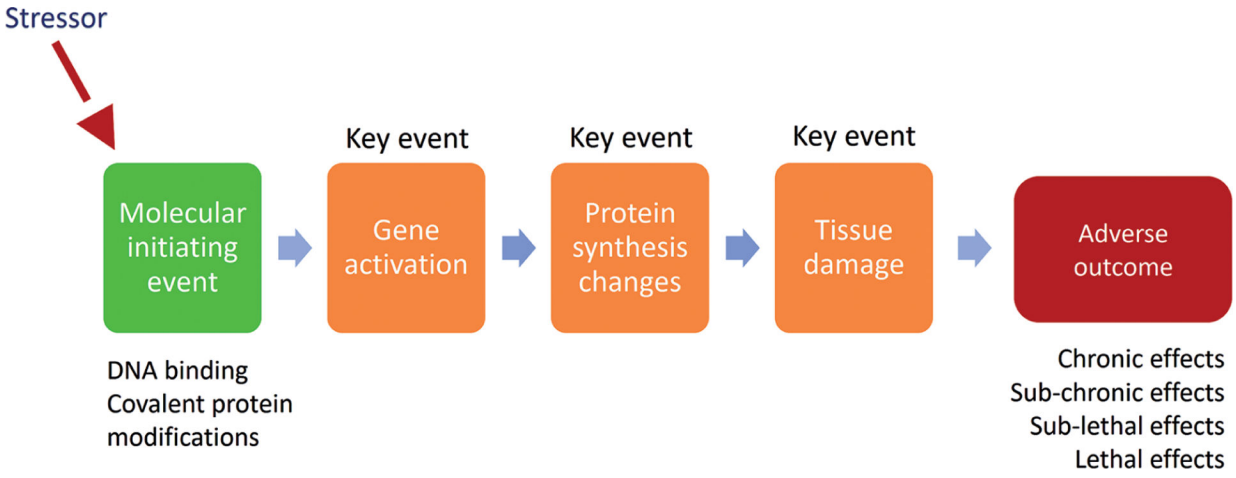

Fig. 2:
Adverse outcome pathway (AOP) framework

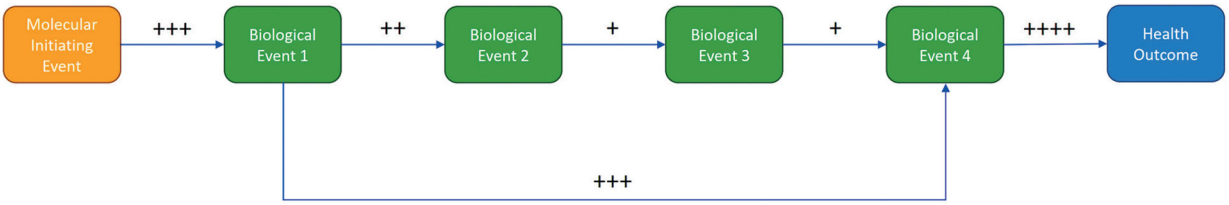

Fig. 3:

Schematic putative AOP (proposed by Fernando Nampo) displaying estimated certainty in key event relationships (from the molecular initiating event, MIE, through key events, to adverse outcome, AO), evaluated for certainty using the GRADE framework (each + symbol representing level of certainty, with + being lowest and ++++ being highest)