Organisation for Economic Co-operation and Development

ENV/CBC/MONO(2022)24/ANN2

Unclassified

English - Or. English

2 September 2022

ENVIRONMENT DIRECTORATE
CHEMICALS AND BIOTECHNOLOGY COMMITTEE

Appendix B1. Report on the statistical analyses performed to assess developmental neurotoxicity (DNT) of deltamethrin and flufenacet and a stressor-based AOP for DNT

Series on Testing and Assessment
No. 362

JT03501735

OECD Environment, Health and Safety Publications

SERIES ON TESTING AND ASSESSMENT

NO. 362

Appendix B1. Report on the statistical analyses performed to assess developmental neurotoxicity (DNT) of deltamethrin and flufenacet and a stressor-based AOP for DNT

IOMC

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT

Paris 2022

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 38 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in twelve different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides**; **Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; Safety of Manufactured Nanomaterials;** and **Adverse Outcome Pathways.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

- **Also published in the Series on Testing and Assessment: link**

For this and many other Environment, Health and Safety publications, consult the OECD's World Wide Web site (www.oecd.org/chemicalsafety/)

or contact:

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

E-mail: ehscont@oecd.org

Table of contents

FIGURES

TABLES

Abbreviations

AOP	Adverse Outcome Pathway
BN	Bayesian Network
CFD	Conditional Probability Distribution

Abstract

This report aims at describing the methodological approach adopted to collect, appraise, analyse and integrate the evidence for assessing developmental neurotoxicity of deltamethrin and flufenacet and to develop an AOP for DNT taking into account the identified uncertainties. Overall, it can be defined as an 'evidence-based AOP approach', i.e. an AOP derived combining the systematic retrieval, screening and appraisal of the evidence from the scientific literature with a tailored primary data collection of invitro data and advanced methodological approaches aiming at delivering sound, transparent and accessible scientific advice in support to decision making. The report describes the results of the systematic review as for the number of selected relevant studies, their main characteristics and a summary of the appraisal of their internal validity. Central in the methodological framework is the role of the uncertainty analysis. Tailored uncertainty tables have been used to screen endpoints/KEs that might be relevant for a deltamethrin-based DNT-AOP. A Bayesian network approach, supported by expert knowledge elicitation, has been adopted to analyse and express the uncertainty in the Key Event Relationships and overall in the pathways. The methodological framework is quite innovative in the context of the AOP approach and therefore might require additional validation in the future. it represents a step forwards in setting up quantitative AOPs models, the lack of which is considered one the main obstacles to successfully implement the AOP framework in the regulatory context.

1 Methodological approach

The assessment carried out to set up this stressor-based AOP, using deltamethrin as prototypical chemical, represents a step foreword in terms of the methodological approach taken to collect, appraise, analyse and integrate the evidence taking into account the identified uncertainties. Overall, it can be defined as an 'evidence-based AOP', i.e. an AOP derived combining the systematic retrieval, screening and appraisal of the available evidence from the scientific literature with a tailored primary data collection of *in vitro* data (Masjosthusmann et al., 2020) and advanced methodological approaches aiming at delivering sound, transparent and accessible scientific advice in support to decision making.

Two main methodological aspects can be considered innovative in the context of the AOP framework. The first is the recourse to the systematic review process for the retrieval, the selection, the appraisal and the synthesis of the available evidence. The second, the use of a probabilistic approach, to analyse and express the uncertainty that is inherent in all scientific assessments both in relation to the retrieved evidence and the methods.

1.1. The systematic review process

1.1.1. The protocol (Appendix A)

Clarification of the question and planning the methods for addressing the question and its sub-questions is a well established practice in primary research and in the systematic review context (Chandler et al., 2020). Acknowledging the advantages of this approach, in recent times EFSA has started to regularly adopt this approach for all the scientific assessments that are not related to an application for regulatory products (EFSA, 2020). Following this line, a plan for the systematic review component of this mandate has been drafted (see protocol Appendix A). The protocol includes:

- translation of the mandate into scientifically answerable questions and sub-questions (including specification of the population, intervention/exposure, control, outcomes);
- list of literature databases to search;
- search strings to retrieve the studies;
- eligibility criteria and procedures for screening studies for relevance;
- data model and procedures for the data extraction;
- critical appraisal tools (CATs) and procedures for assessing the risk of bias;
- tools used to performed the activities above.

Evidence synthesis and integration including uncertainty analysis has been included only partially in the protocol since, due to the novelty of the application of a probabilistic approach in the AOP framework, it was difficult to anticipate the methods to use at the time of the protocol draft. Therefore the methods to address the Key Event Relationships (KERs) and the related uncertainty have been defined only once the data extraction has been completed. It was possible instead to anticipate the methods to screen the evidence to select MIEs/KEs/AOs to be included in the putative AOP (i.e. uncertainty tables).

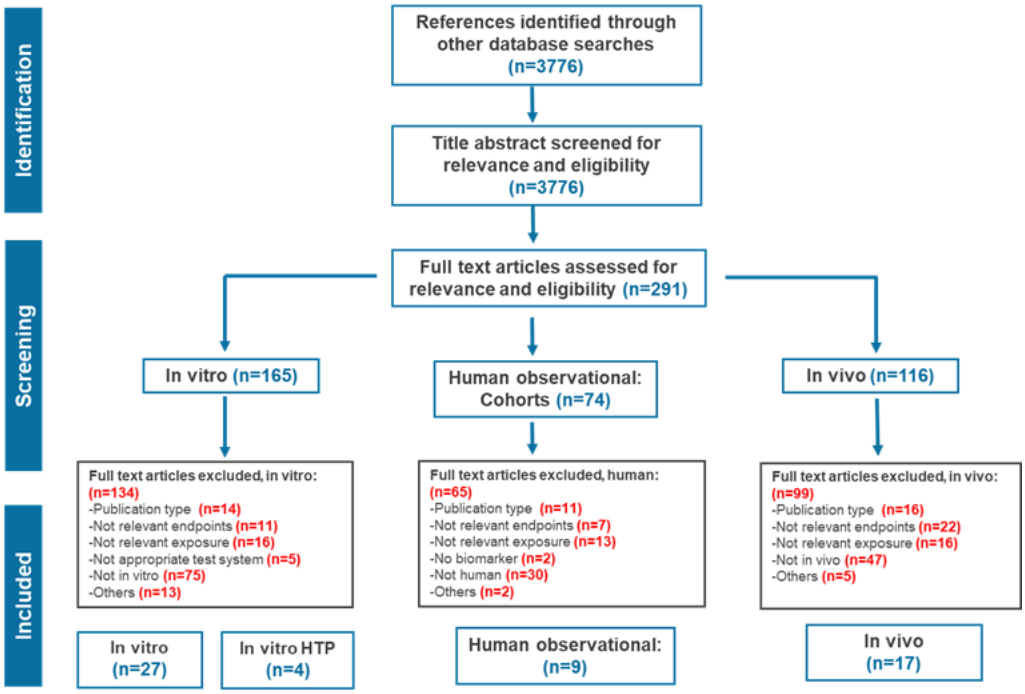
1.1.2. Systematic review results

The literature searches were conducted in three electronic bibliographic databases (PubMed, Web of Science, TOXNET) and three resources indexing PhD theses (DART, EBSCO and PQDR) in July 2020, and updated on 23 of November 2020 (deltamethrin) and on 7 December 2020 (flufenacet) by an information specialist. Search strings are described in the protocol (Appendix A). Terms for the exposure were combined with relevant terms for DNT outcomes (human and *in vivo* studies) or methods (*in vitro* studies) and a specific search string was designed to identify studies applying high-throughput methods to evaluate potential developmental neurotoxicity without terms of exposure (Appendix A). The DNT outcomes were predefined by a series of toxicological *in vivo* and *in vitro* endpoints and measurements in human observational studies and categorised in endpoint categories translated into keywords for the searches (see Appendix A Section 2.1.3).

1.1.3. Deltamethrin systematic review results

For deltamethrin, two independent reviewers screened the literature identified through the searches; 3,776 unique references were identified after removing duplicates (see PRISMA Chart, Figure 1). The evidence was clustered as *in vivo* (containing *in vivo* experimental studies), *in vitro* (containing *in vitro* mechanistic studies and behavioural studies with exposure conducted in zebrafish up to 120 hours post fertilisation) or human (containing human observational studies) during the title and abstract screening. The title and abstract screening left 291 relevant articles that underwent a full-text review, of those, 165 were classified as *in vitro*, 74 as human evidence and 116 as *in vivo*. For *in vivo*, 99 publications were excluded and 17 included providing relevant data on 58 endpoints. For human, 65 publications were excluded and eight included providing relevant data on 12 endpoints. For *in vitro*, 134 were excluded and 31 publications were included on 64 endpoints (see full list of references included and excluded and reasons for exclusion, Appendix B2.1).

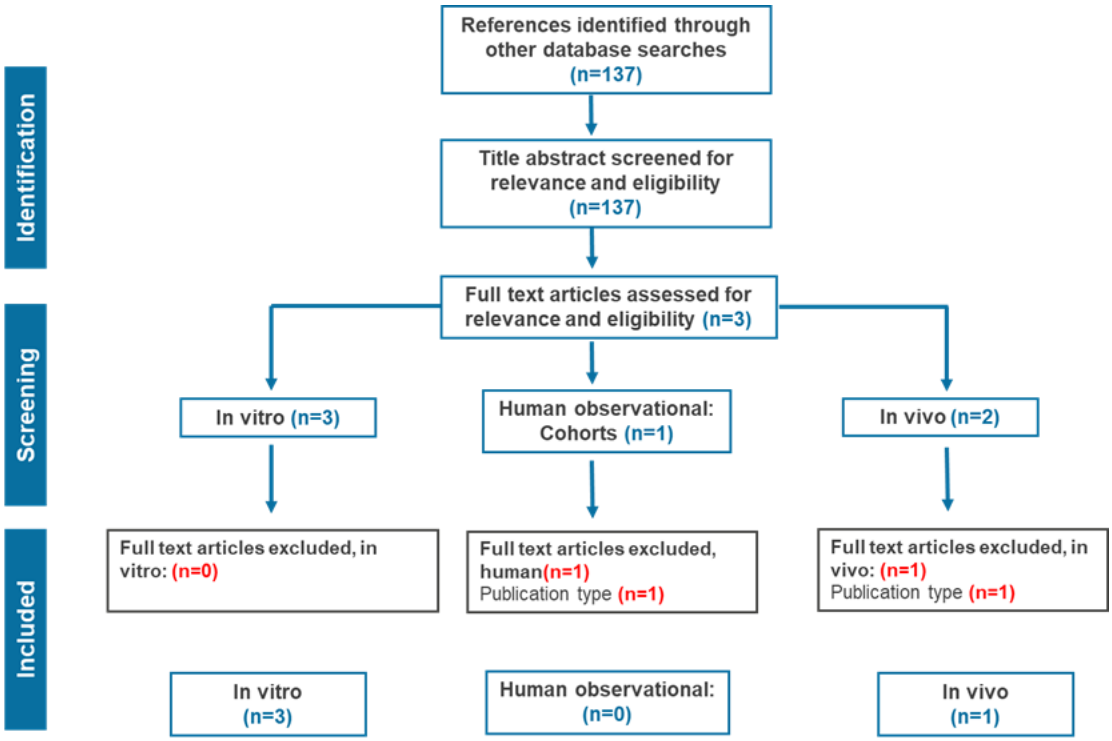
Figure 1. PRISMA chart – systematic literature review result of the screening for relevance for deltamethrin



1.1.4. Flufenacet systematic review results

For flufenacet, two independent reviewers screened the literature identified through the searches; 137 unique references were identified after removing duplicates (see PRISMA Chart, Figure 2). The evidence was clustered as *in vivo* (containing *in vivo* experimental studies), *in vitro* (containing *in vitro* mechanistic studies and behavioural studies with exposure conducted in zebrafish up to 120 hours post fertilisation) or human (containing human observational studies) in the title and abstract screening. The title and abstract screening left five relevant articles that underwent a full-text review, of those, three were classified as *in vitro*, 1 as human evidence and two as *in vivo* (of those one was also classified as *in vitro*). For *in vivo* a one study was included as providing relevant data, for *in vitro* three publications were included. For human no relevant publications were retrieved (see full list of references included and excluded and reasons for exclusion, Appendix B2.2).

Figure 2. PRISMA chart systematic literature review result for flufenacet of the screening for relevance



1.1.5. Lines of evidence

Data related to three lines of evidence have been collected and appraised by the systematic review process and the high-throughput battery (Masjosthusmann et al., 2020):

- Human observational data used to address the causal association between exposure to deltamethrin and health outcomes in humans (scientific literature).
- *In vivo* data on zebrafish and rodents used to address the causal association between exposure to deltamethrin and health outcomes in zebrafish, rats and mice (scientific literature and studies from the applicant submitted as part of the dossier).
- *In vitro* data retrieved from the literature and collected using a battery of tests whose experimental design was defined by EFSA. The battery unit was human and/or rat and/or mouse neuro cells in development. Exposure duration ranged between several hours up to 28 days.

Since there is no commonly accepted thresholds for biologically significant effects, biological relevance was based on expert judgement taking into consideration heterogeneity of the response measures (i.e. EC50, IC50, BMR50) and uncertainties in the experiments.

1.1.6. Risk of bias

Risk of bias (RoB) in eligible studies on rodents (experimental toxicological studies) and humans (observational studies) was appraised by endpoints using tailored versions of the OHAT-NTP RoB tool (OHAT/NTP, 2015). For the *in vitro* studies, a non-validated tool developed by OHAT-NTP for a specific project (PFOA and PFOS Monograph (2016)) was adapted. Critical appraisal tools were defined upfront and are described in the protocol (Appendix A). Studies were classified as being at low (Tier 1), moderate (Tier 2) or high (Tier 3) RoB for each of the endpoints they measured. RoB tiers were derived weighing the appraisal from the individual RoB domains some of which were identified as key. RoB was appraised by endpoint in the cases that it was considered that the different endpoints measured in a study used different methodology and therefore have different RoB.

For deltamethrin the outcome of the RoB appraisal and descriptive forest plots are presented in Appendix B3.1 for *in vivo, in vitro*, human and zebrafish lines of evidence. For flufenacet the outcome of the RoB appraisal and descriptive forest plots are presented in Appendix B3.2 for *in vivo, in vitro* and zebrafish lines of evidence.

1.1.7. Data extraction and analysis

As anticipated in the protocol (Appendix A), the specific endpoints retained for the synthesis have been summarised in a qualitative manner. Due to the sparse and heterogenous nature of the available data, no meta-analysis has been performed. Available evidence have been clustered hierarchically by evidence streams first and then by apical adverse outcomes categories and related subcategories and/or specific endpoints. The studies' results have been graphically displayed together with the main study characteristics as follows:

For HUMAN studies, the following characteristics are displayed in the tables:

- metabolite (specific/unspecific) and LOD
- concentration' of metabolite
- endpoint and method for assessment
- statistical model (Linear regression; Logistic regression; Negative binomial; Bayesian Kernel Machine Regression; Linear mixed model; Generalised model; Reverse scale Cox regression; Difference between exposure OUT?)
- adjustment and for what
- tier
- Q2 (confounding) (key question in appraisal)
- Q4 (confidence in exposure characterisation) (key question in appraisal).

For *IN VIVO* studies:

- species
- sex
- dose
- tier of internal validity
- maternal or system toxicity (Q9 in CAT)
- group sample size
- exposure duration

- exposure stage.

For *IN VITRO* studies:

- species
- test system and origin of the test system
- stage of development of the primary cells
- concentration
- tier of internal validity
- system toxicity (Q9a in CAT)
- number of biological replicates
- exposure duration.

For ZF (behavioural) studies: same as for *IN VITRO*.

The outcome of the data extraction and analysis by line of evidence and endpoint is presented in the Appendix B4.1 containing all the graphs for *in vivo*, *in vitro* and zebrafish for deltamethrin, and the human evidence table in Appendix B4.2. For flufenacet graphs for *in vivo, in vitro* and zebrafish are presented in Appendix B4.3.

1.2. Accounting for the uncertainty: the probabilistic approach

As for *in vivo* and *in vitro* evidence, only studies appraised as RoB Tiers 1 and 2 were retained for further analysis.

To screen the evidence and to identify MIEs/KEs/AOs that can enter the putative AOP, the evidence within each line has been clustered according to hierarchical levels.

For evidence from *in vivo* studies on rodents and zebrafish (ZF) the following grouping has been considered:

- SE: Specific Endpoint (e.g. for ZF 'Locomotor activity-total distance moved'; for rodents: 'Acquisition – Learning and memory – MWM acquisition speed').
- EC: Endpoint Category (e.g. for ZF: 'Locomotor activity'; for rodents: 'Learning'; for humans: 'Expressive Communication (Bayley Scales of Infant Development, third edition (BSDID-III))'.
- AO: Adverse Outcome (e.g. for ZF: 'Behaviour'; for rodents: 'behavioural'; for humans 'behavioural').

For the evidence on humans only the last two levels (i.e. EC and AO) have been used as grouping factors.

Similarly, the evidence on neuro cells in development have been grouped by:

- SE: Specific endpoint (e.g. 'events/burst');
- MIE/KE: Molecular Initiating Event or Key Event (e.g. 'patch clamp – electrical activity').

Questions related to both hazard identification (i.e. identification of the MIE/KE/adverse outcome to be considered in the putative adverse outcome pathway as potentially causally associated with deltamethrin exposure) and characterisation have been addressed for the three lines of evidence on *in vitro*, zebrafish and *in vivo*. Only hazard identification questions could be answered using human observational studies (evidence on humans) due to high RoB affecting all the studies. The list of questions is provided in Annex A. As for the definition of the causal relationship, the classical Bradford–Hill criteria have been considered (Fedak et al., 2015).

1.2.1. Uncertainty analysis for screening the evidence to identify potential MIEs, KEs, AOs

Following EFSA recommendations to address uncertainty in addressing scientific assessment questions possibly using quantitative approaches (EFSA Scientific Committee, 2018a, 2018b), an uncertainty analysis was performed within each line of evidence and hierarchical level to support conclusions on the HI and HC questions and identify MIEs, KEs, AOs to be included in the putative AOP network. A stepwise approach was taken. The assessment of the uncertainty started with the lower hierarchical levels and progressed at the higher levels (e.g. conclusions on the endpoint category for *in vivo* data were based on those achieved for the specific endpoints). Progression of the assessment towards a higher level (e.g. endpoint category – locomotor activity) was carried out only when at the lower level (i.e. specific endpoint) at least one outcome (e.g. locomotor activity – total distance moved) was assessed as possibly (probability greater than 0.66) causally association with deltamethrin exposure.

The uncertainty analysis was performed using predefined lists of factors/domains and related guiding questions tailored by lines of evidence. The factors/domains were assessed in two ways. First potential explanations for the identified heterogeneity in the results (if any) were assessed. If inconsistencies could not be justified by any factor/domain, the unexplained inconsistencies were treated as a source of uncertainty. Secondly the same factors/domains were appraised for adequateness in the body of evidence in relation to the specific endpoint/endpoint category/adverse outcome. Factors/domains considered not adequate were retained as sources of uncertainty. A detailed list of factors/domains by line of evidence is provided in Annexes B (from now on referred to as uncertainty tables). For both steps (assessment of the inconsistencies and of the potential sources of uncertainty), the judgement was achieved answering to domain and line of evidence specific 'guiding questions'. The WG experts were instructed to provide synthetic answers (Yes/No/Not Relevant) and accompany them with a narrative explanation providing the rationale for the assessment. The tables displaying the results of the uncertainty analysis for screening KEs/MIEs/AOs are provided in Appendix B5.1 and B5.2 for deltamethrin and flufenacet, respectively.

Based on the identified unexplained inconsistencies and uncertainties (if any), a judgement was made on:

- whether exposure to deltamethrin is causing the KE/specific endpoint/endpoint category/adverse outcome (hazard identification);
- only for the KE/specific endpoint/endpoint category/adverse outcome for which a causal association was identified, the lowest concentration/dose triggering the KE/specific endpoint/endpoint category/adverse outcome.

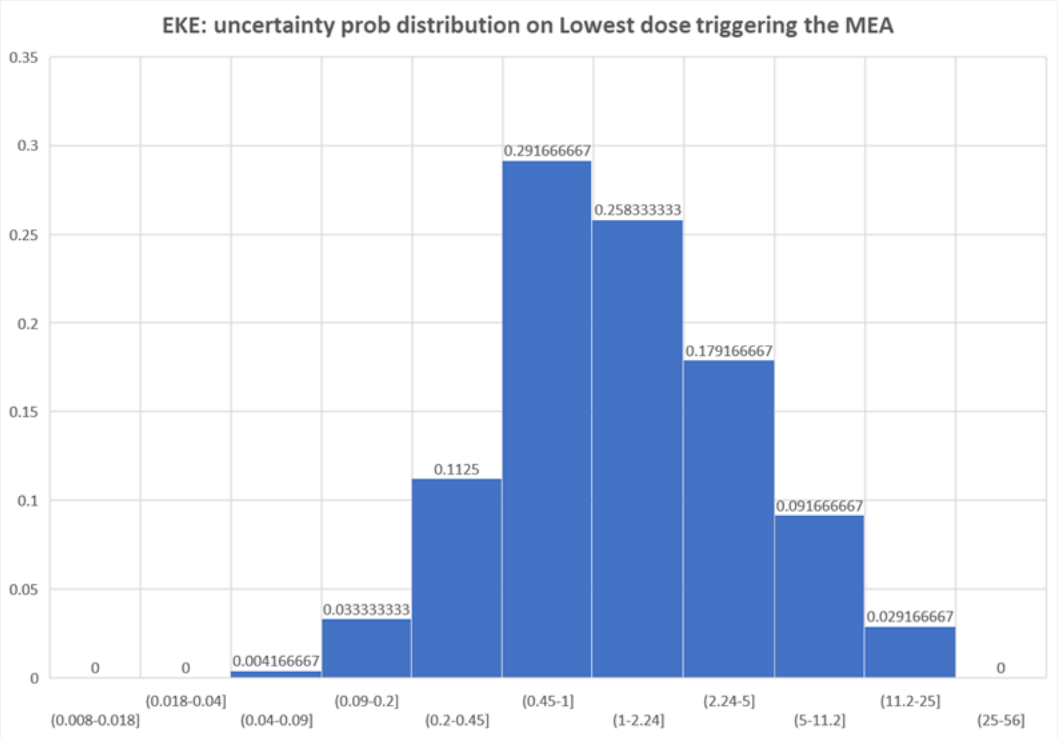
To take account of the identified uncertainties (including the unexplained inconsistencies) in the hazard identification, a probabilistic expression of the results was adopted. The causal association was judged as occurring or not occurring with an overall level of certainty summarised by the WG consensus subjective probability on the causal relationship to occur. The threshold of 0.66 (twice as possible as not) was used as the minimum subjective probability (minimum level of certainty) leading to the conclusion of a causal association. The use of a bounded probability was thought to reflect adequately the difficulties in expressing the level of certainty more precisely and was considered sufficiently precise considering the purpose of this step of the process (screening potential KEs and AOs for the AOP). For the human line of evidence, the working group decided not to address the hazard characterisation due to the overall high RoB affecting the body of evidence. Only for the human line of evidence, the certainty/uncertainty on the hazard identification question was assessed more precisely using the approximate probability scale (%): [0–10), [10–33), [33–50), [50–66), [66–100]1.

1 Squared parenthesis indicates that the extreme is included, rounded parenthesis that it is

The uncertainty affecting the body of evidence used for identifying the lowest concentration/dose triggering the KE/specific endpoint/endpoint category/adverse outcome (hazard characterisation), was expressed using either ranges (under the assumption that all values in the range were equally probable – uniform distribution) or individual probability distributions. The estimates were derived using a semi-formal expert knowledge elicitation (EFSA, 2014; EFSA Scientific Committee, 2018a, 2018b). In the cases were a range was used, the experts were first requested to provide an individual estimate. Then a consensus range was achieved based on the discussion among the experts. For two endpoints (i.e. MEA and behavioural adverse outcome) the uncertainty distribution was derived using the roulette method (EFSA, 2014; O'Hagan, 2019). In the roulette method, first the experts have to agree on an overall plausible range (i.e. range that would include the true value almost certainly) that is partitioned in a number of subsets. The latter would depend on the level of accuracy that is requested to achieve with the uncertainty estimate (i.e. the bigger the number, the larger the accuracy). Then the experts are requested to provide their subjective probabilities that each subset (bin) of the overall plausible range include the true value of the parameter (lowest dose/concentration in the case at hand) by allocating chips to each bin. The chips can be equally distributed among the bins or concentrated in one or more bins depending on the subjective belief of the expert on which ones are more probable to contain the true value. Each chip has a predefined value in terms of probability that depends on the total number of chips (e.g. 0.05 when using 20 chips, 0.10 for 10 chips). The strength of the method is that it is very intuitive and provides an immediate graphical representation of the experts' beliefs. The individual uncertainty distributions are then summarised in a single distribution by mathematical averaging. Based on the collegial discussion the experts had the possibility to revise their individual judgements until a consensus probability distribution was achieved (Figure 3 and Figure 4). The evidence dossier and the detailed report of the expert knowledge elicitation exercise is provided in Appendices C5.1 and C5.2 (deltamethrin and flufenacet respectively).

excluded

Figure 3. Uncertainty distribution on the lowest concentration triggering the MEA

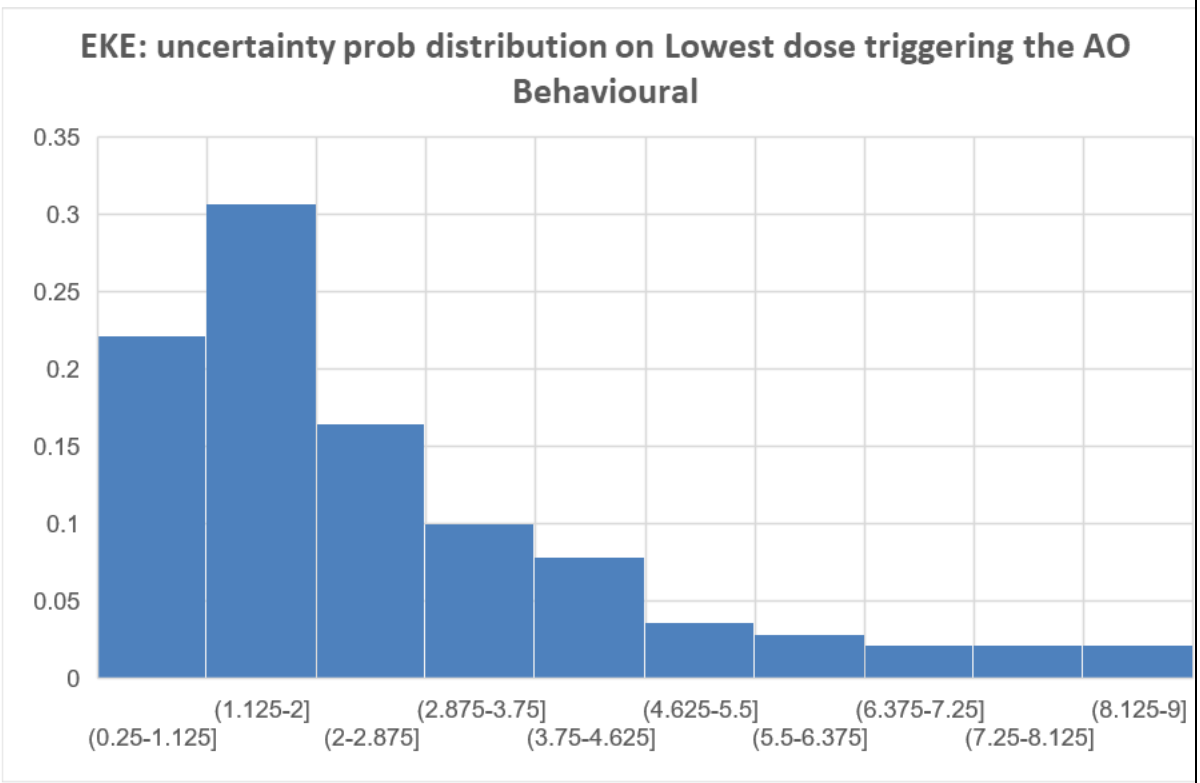


Rationale for the uncertainty distribution: The overall range expected to include the lowest concentration almost certainly was chosen considering: variability of the different tests and systems, the shape of the different dose–response and a pragmatic approach to guarantee sufficient separation among concentrations. A logarithmic scale was considered by the experts more appropriate to split the overall credibility range into subsets. It was highlighted that in the laboratory practice the step of increase of concentrations/doses is set using logarithmic scale. The bounds of the bins are in the range of -3;+2.5 in a log scale of base 5 with step increase of 0.5 (reported above in the original scale). Twelve experts participated to the elicitation. The outcome of the elicitation was based on all the NNF (MEA) studies and provided a credibility range of 0.04–5 µM that is expected to include the true lowest concentration with around 88% probability. The judgement is based on the consideration that lower bounds of concentration for individual studies with different test systems and methods mostly overlap in the range of 0.45–5 µM covering both the distribution for single administration (acute protocol) and multiple administration over developmental period of the network (developmental protocol) (see Appendix C for detailed results of the discussion).

During the second round of elicitation, the initial distribution shifted from the ‘right’ (higher concentrations) to the ‘left’ (lower concentrations) following a discussion among participants to the EKE. The initial thoughts for some of the participants for having a higher concentration in place was mainly based on the following considerations: more weight to the ‘developmental’ protocol as expression of DNT, more weight to the ‘human derived’ cell system or because the area that has most overlap in the confidence intervals was considered more likely to cover the true value. After discussion on these relevant points consensus was reached considering the arguments that are summarised at the beginning.

Note: A round or squared parenthesis indicates that the extreme is excluded or included, respectively

Figure 4. Uncertainty distribution on the lowest dose triggering the AO behavioural



Rationale for the uncertainty distribution: The initial range expected to include almost certainly the lowest dose was split in 10 equally sized subranges reflecting the level of precision the experts felt being able to achieve in their judgement. The experts discussed the distribution by considering the expected corresponding brain concentration in the pup. They agreed that the different route and period of administration is associated with some inconsistencies and uncertainties (listed in Appendix C) making the use of external dose complicated when assessing evidence coming from three different studies. The experts finally concluded that the range of external doses between 0.25 and 2 mg/kg/bw/day (based on the study likely to yield the highest concentration in the brain) had a 53% probability to include the true lowest dose that would trigger the AO. This distribution mainly reflects the occurrence of the adverse outcome in what is expected to be the most sensitive population, i.e. the pups, and the adverse outcome at the lowest dose is considered based on effect on sensory motor and cognitive behaviours. Overall, the credible range 0.25–7.25 mg/kg/bw/day is expected to include the true lowest dose with 96% probability (see Appendix B1 for details in the EKE process and Appendix C for detailed results of the discussion).

Note: A round or squared parenthesis indicates that the extreme is excluded or included, respectively

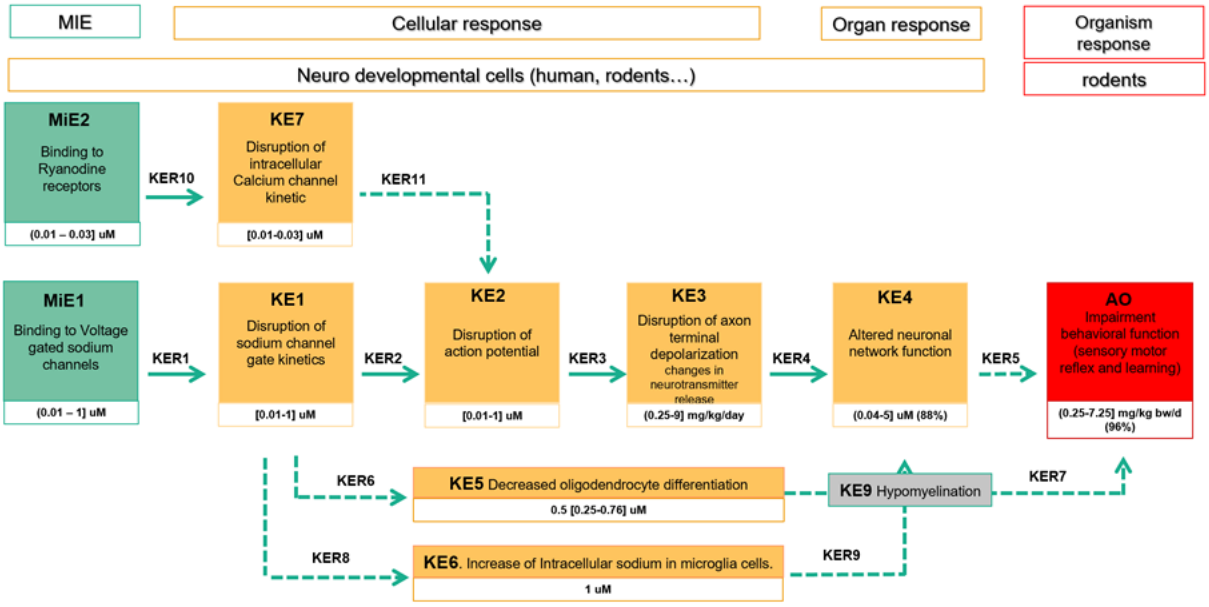
Appendix C provides a summary of the assessment of the causal association of MIEs/KEs/AOs with deltamethrin exposure and the associated estimate of the lowest concentration/dose (when appropriate) for the various lines of evidence. For the events/Adverse outcomes for which a full probability distribution was elicited, the 88% and 96% credibility interval (respectively for MEA and AO) is derived keeping out 12% and 4% of the distribution from the upper tail.

The assessment of the causal association with deltamethrin exposure in light of the identified uncertainties allowed the MIEs/KEs/AOs to be considered for a putative adverse outcome pathway for developmental neurotoxicity.

1.3. Putative stressor-based AO Network (deltamethrin): structure of the AOP (MIEs, KEs, AO and KERs)

The putative AOP structure for DNT as defined on the basis of the discussion among WG experts is displayed in Figure 5. Below each MIE/KE/AO a range or a 96%-97% credibility interval are given expressing the uncertainty in the estimate of the lowest concentration/dose triggering the event.

Figure 5. Putative AOP network structure.



1.4. A quantitative approach to the AOP: the Bayesian network

Lack of approaches to quantitatively model AOPs and AOP networks including the associated uncertainty have been recently identified as the main obstacles to successfully implement the AOP framework in the regulatory context (LaLone et al., 2017). The Bayesian network (BN) framework represents a promising approach here. As any quantitative AOP (qAOP) a BN can serve as a computational tool for translating or extrapolating from mechanistic measurements of an upstream KE to a predicted severity/status of the AO (Fenton and Neil, 2012, Muller et al., 2015, Perkins et al., 2019).

Three types of probabilities are associated to the BN structure and can be used to infer conclusions on the KEs and the triggering stressor:

Conditional probability distributions: the conditional probability is the probability of each of the possible statuses of a downstream event given each possible statuses (or combination of statuses) of the connected upstream event(s) (i.e. the conditioning event). For each combination conditioning upstream KE(s)status/status of the target KE, the conditional probability can be higher or lower than the marginal probability depending on the strength of the KER (i.e. the stronger the association the more unequal the conditional and marginal distribution will be).

Marginal probabilities the marginal probability distribution describes the probabilities associated to each possible state of a KE/variable (e.g. activated/not activated, occurrence/not occurrence) irrespective of the state of all the others. This probability distribution can be used to infer what is the most probable status of a KE/AO (e.g. altered behavioural function) assuming exposure to a stressor (e.g. deltamethrin) and therefore to classify the stressor as able to trigger or not the KE/AO taking into consideration the uncertainty in all the KERs in the AOP network. The marginal probability (as the conditional and joint probabilities) can be updated once new 'evidence' (e.g. Disruption of action potential generation activated with certainty) becomes available.

When dealing with categorical variables, the marginal probabilities can be derived from the conditional probabilities using an exact formula (given here for a binary variable taking values 0 and 1) given below for a pair and a triplet.

For a pair:

$$Prob(X_i = k) = \sum_{j = \{0,1\}} \left[Prob(X_i = k/P_{1,X_i} = j) * Prob(P_{1,X_i} = j)\right] \quad [1.1]$$

For a triplet

$$Prob(X_i = k) = \sum_{j = \{0,1\}} \sum_{l = \{0,1\}} \left[Prob(X_i = k/P_{1,X_i} = j, P_{2,X_i} = l) * Prob(P_{1,X_i} = j) * Prob(P_{2,X_i} = l)\right] \quad [1.2]$$

With $k = \{0,1\}$

Joint probability: the joint probability distribution of a set/network of KEs describes the probability of all the possible combinations of the status of the KEs in the network. A natural choice for a joint distribution representing a set of dichotomous variables is a multinomial distribution assigning a probability to each combination of states of the KEs in the AOP. Since the number of combinations dramatically increases when the number of KEs raises, so does the number of distribution parameters. As illustrative examples for a network with 10 KEs each of which entailing only two possible status (active/not active), the number of possible combinations is 1,023. As a consequence, the probability attached to each combination is rarely very high unless the evidence is supporting it with high certainty.

1.4.1. Bayesian networks: general description of the approach

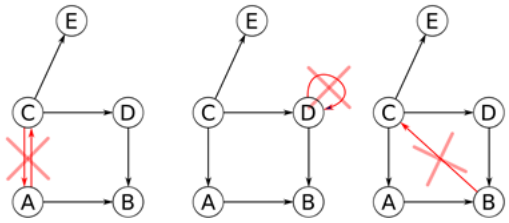
The Bayesian Networks or Bayesian Belief Networks are graphical models that allows representation of the probabilistic structure of multivariate data using a graphical display (Scutari and Denis 2015). The BN approach has recently started to be applied successfully in the context of AOP (Moe et al 2020, Jeong et al. 2019, Jaworska et al 2013, Pirone et al 2014). They entail: 1) a set of random variables X = {X1, ….,Xk} with the associated joint probability distribution (named global probability distribution); 2) a graphical representation, namely a Directed Acyclic Graph (DAG), describing the dependencies/independencies within the set of variables X;

The probability that an upstream node(s) will activate or inactivate a downstream node is defined experimentally or by expert judgement and summarised in probability tables associated with each node. A quantitative AOP-BN uses information on the activation of each node across the network to model the potential for a chemical to cause the adverse outcome.

In a DAG, the random variables in the set X are represented as nodes (MIEs, KEs and AOs in the context of an AOP) and the links between variables (KE relationships in an AOP) as directed arcs. The links in a DAG have to satisfy the requirements of having a direction and being acyclic (i.e. not involving

any cycles or loops, Figure 6). Nodes in the network play a different role depending on the direction of the arcs from which they are connected to other nodes (Figure 6). A node is defined as parent when an arc departs from this node towards another/other node(s) e.g. node 'A' is a parent of 'B', 'C' is parent of 'E'. A child node is one to which an arc arrives from one or more nodes e.g. node 'B' is a chid of 'A' and 'D'. Nodes sharing a child are named spouses. Nodes of this type are 'A' and 'D' since they are both parents of 'B'. The role of the nodes is crucial in the definition of the conditional dependency/independency in the network. In fact each node in the DAG is conditionally independent of any other non-descendent node given its parents. This property of the DAG combined with the recursive application of the Bayes Theorem allows factorisation of the joint probability distribution of the set of random variables X (KEs) using the conditional probability distributions (named local probability distributions). The latter quantify the strength of the dependencies between couple or triplets of variables conditioning the distribution of each node to the (combined) state of its parent(s). The factorisation property makes the inference on the BN computationally lighter since the set of local distributions has overall fewer parameters compared with the global distribution. The dimensional reduction is a central property of the BN that allows its application also to high dimensional problems.

Figure 6. Illustrative example of a DAG with parents, children and spouse nodes



Note: red crosses highlight cycles and loops not allowed in a DAG.

The BN allows the 'belief updating' or 'belief propagation' that consists in updating the conditional and marginal probability distribution of each variable in the set X, as encoded by the BN, in the face of new evidence.

The BN-AOP can have different applications. It can be used to perform prognostic inferences i.e. to prospectively predict the probability of the occurrence and/or the severity of an adverse outcome based on different scenarios (i.e. different combinations of the possible status of the upstream variables/events in the networks). BN-AOP can be also used to derive diagnostic inference, running the model from the adverse outcome backward to identify KEs that are the main determinants of the AO. Furthermore the model can be run from any intermediate MIE or KE backwards and forwards.

Associations entailed by the BN can also be interpreted as causal relationships if the following additional assumptions are met (Scutari and Denis, 2015):

- Each variable is conditionally independent of its non-effects, both direct and indirect, given its direct causes.
- There must be a DAG which maps not only all conditional independence but also dependences in the set of vars X.
- There must be no latent variables (unobserved variables influencing the variables in the network) acting as confounding factors. Such variables may induce spurious correlations between the observed variables, therefore introducing bias in the causal network.

The assumption of the absence of latent variables is particularly difficult to meet. Frequently a latent KE being the pathway between an upstream and a downstream KE (i.e. a confounding factor) exists but evidence is lacking on it.

The structure and the parameters of a BN (i.e. the parameters of the conditional probability distributions) can be either learned directly from data using mathematical algorithms or derived using expert judgement informed by available data.

Resorting to expert judgement easily allows mirroring the suspected causal relationships in the modelled system but it can be more prone to biases in the knowledge. Conversely, it turns advantageous when available data are scarce and affected by missingness. In both cases, whether using an algorithm or expert judgement, the BN structure would require a validation process by an external set of data to confirm the strength of the relationships and the robustness of the identified dependencies/independencies.

The BN approach has recently started to be applied successfully in the context of AOP (Moe et al., 2020; Jeong et al., 2019; Jaworska et al., 2013; Pirone et al., 2014). Advantages of its implementation in this context include fitting naturally the concept of the AOP network, allowing the integration of data coming from different lines of evidence such as *in vitro* high-throughput assays and *in vivo* toxicological data and measuring the impact of introducing a new information in the system. Last but not least, being a probabilistic model, it allows accounting and propagating uncertainty also in complex assessment systems

1.4.2. A stressor-based AOP-BN for deltamethrin

A binary AOP-BN has been used to describe the AOP for deltamethrin (DM) with random variables (MIEs, KEs and AO from now on referred to as KEs) assuming only two categories (active, inactive). Although more sophisticated models could have been developed using quantitative dose–response and response–response to model the KE relationship, this option would have required more ample data to be implemented. Those data were not available.

The putative AOP was provided by the experts based on the available data (literature based and invitro battery data) and their knowledge of the mechanistic process leading from the exposure to deltamethrin to altered behavioural functions. The structure of the hypothesised AOP-BN is described in Figure 7.

The AOP-BN for DM is composed of 10 nodes each taking two possible levels/states (active/inactive) and 11 KERs:

> X = {MIE1,MIE2, KE1, KE2, KE3, KE4, KE5, KE6, KE7, AO1}
> KER = {KER1: KER between MIE1 and KE1; KER2: KER between KE1 and KE2; KER3: KER between KE2 and KE3; KER4: KER between KE3 and KE4; KER5: KER between KE4 and AO; KER6: KER between KE1 and KE5; KER7: KER between KE5 and AO; KER8: KER between KE1 and KE6; KER9: KER between KE6 and KE4; KER10: KER between MIE2 and KE7; KER11: KER between KE7 and KE2}

Considering the binary nature of the variables (KEs) a natural choice for the joint probability distribution (i.e. the global distribution) and the conditional probability distributions (local distributions) is represented by the multinomial distribution.

The conditional probability distributions (CPD) associated with pairs (e.g. KE1 and its parent MIE1) or triples of nodes (e.g. KE2 with its parents KE1 and KE7) describe the probability that a child/descendent node (downstream KE) is active or inactive given the (combined) state of its parent node(s) (upstream MIEs/KEs). They quantify the strength of the KE relationships and allow the propagation of the uncertainty throughout the pathway.

Dealing with the joint probability distribution is not convenient since it entails a large number of parameters to be estimated. Resorting to the property of the conditional independence implied by the DAG (i.e. each variable/node is conditionally independent by all other nodes except its children given

its parents), it is possible to factorise the joint probability distribution of the set of events X into the product of the local CPDs – formula [2]:

$P(X) = \prod_{i=1}^{n} P(X_i/\Pi_{X_i};\ \Theta_{X_i})$ Joint probability distribution [2]
where $\Pi_{X_i} = \{parents\ of\ X_i\}$
and $\Theta_{X_i} = parameters\ of\ the\ conditional\ distribution\ of\ X_i\ given\ the\ parents\ \Pi_{X_i}$

Therefore, for the set of variables/KEs in the AOP-BN for deltamethrin, the joint probability distribution is given by:

$$P(X) = P(\mathrm{MIE1}, \mathrm{MIE2}, \mathrm{KE1}, \mathrm{KE2}, \mathrm{KE3}, \mathrm{KE4}, \mathrm{KE5}, \mathrm{KE6}, \mathrm{KE7}, \mathrm{AO1}) =$$
$$P(MIE1) * P(MIE2) * P(KE1/MIE1) * P(KE2/KE1, KE7) * P(KE3/KE2) * P(KE4/KE3, KE6) * P(KE5/KE1) * P(KE6/KE1) * P(KE7/MIE2) * P(AO/KE4, KE5)$$

Interpreting the network as a causal one rests to the assumption that there are no latent variables. Such an assumption is difficult to meet in the AOP-BN for deltamethrin since there are key events that are expected to be in the pathway, based on expert judgement, but due to the lack of evidence they are not appearing in the putative AOP-BN (except for hypomyelination displayed in grey in Figure 7). Therefore the relationships embedded in the AOP-BN cannot be characterised as causal.

1.4.3. The conditional probability tables: approach used for the estimate

An expert judgement approach has been adopted to estimate the parameters of the CPDs. This choice was made necessary by the lack of large sets of data covering the whole pathway.

For each pair of nodes in the BN the following question was addressed: 'What is the probability that the downstream KEx is activated given the upstream (parent) KE is activated/not activated?' For each triplet the question to answer was: 'What is the probability that the KEx is activated given all possible combinations of the state of the upstream (parents) KEs (i.e. active/active, active/non-active, non-active/active, non-active/non-active)?'

To answer these questions the following steps were taken:

1) The experts were requested to assess individually first and then collectively, for each of the (combinations of) instances of the conditioning event(s) and for the conditioned event instance 'activated', three criteria: the biological plausibility (BP from now on), essentiality (E from now on) and empirical evidence (EE from now on) in support to the KER. Those criteria are indicated in the AOP handbook (OECD, 2018) as the three critical Bradford–Hill considerations for AOPs items to characterise the KEs, KER and their uncertainty when demonstrating causality. The criteria definition and the recommendations on how to assess them being either low, moderate or high were the ones suggested in the AOP handbook (OECD, 2018).
2) The consensus judgements on the three criteria (two for the instance 'not active' in the pairs, 'not active/not active' in the triplets) were combined according to predefined rules for each of the instances of the conditioning events to derive ranges for the conditional probabilities (Tables 1–6).
3) By a collegial discussion the experts were requested to find an agreement on a single probability value within the range derived at the previous step.
4) The conditional probabilities for the various instances of the conditioning event(s) and for the conditioned event instance 'non-activated' were derived based on the property that the sum of the probabilities/density over the various classes/values of a variable distribution must be equal to 1.

Note that in the tables a round or squared parenthesis indicates that the extreme is excluded or included, respectively.

The CPD are provided in Annex C.

Table 1. Rules for combining the criteria assessment – pairs combination 'active (KE downstream)/active (KE upstream)'

Criteria assessments	3 high	2 high & 1 moderate	1 high & 2 moderate or 2 high & 1 low (only E)	3 moderate or 1 high & 1moderate & 1 low (only E)	Any other combination
Probability range of KEx to be activated given the upstream (parent) KE is activated	(0.95–1.0]	(0.85–0.95]	(0.75–0.85]	(0.66–0.75]	[0–0.66]

Table 2. Rules for combining the criteria assessment – pairs combination 'active (KE downstream)/not active (KE upstream)'

Criteria assessments (only biological plausibility and essentiality) Note: essentiality is interpreted in the reverse direction (the higher the essentiality the lower the probability range)	BP high & E low	BP high & E moderate or BP moderate & E low	BP moderate & E moderate	Any other combination
Probability range of KEx to be activated given the upstream (parent) KE is not activated	(0.85–1.0]	(0.75–0.85]	(0.66–0.75]	[0–0.66]

Table 3. Rules for combining the criteria assessment – triplet combination 'active (KE downstream)/active (KEs upstream)'

Criteria assessments	3 high	2 high & 1 moderate	1 high & 2 moderate or 2 high & 1 low (≠BP)	3 moderate or 1 high & 1moderate & 1 low(≠BP)	Any other combination
Probability range of KEx to be activated given the upstream (parents) KEs are activated	(0.95–1.0]	(0.85–0.95]	(0.75–0.85]	(0.66–0.75]	[0–0.66]

Table 4. Rules for combining the criteria assessment – triplet combination 'active (KE downstream)/active (first KE upstream)/not active (second KE upstream)'

Criteria assessments	3 high	2 high & 1 moderate	1 high & 2 moderate or 2 high & 1 low (?BP)	3 moderate or 1 high & 1moderate & 1 low (?BP)	Any other combination
Probability range of KEx to be activated given the first upstream (parents) KE is activated whereas the second is not activated	(0.95–1.0]	(0.85–0.95]	(0.75–0.85]	(0.66–0.75]	[0–0.66]

Table 5. Rules for combining the criteria assessment – triplet combination 'active (KE downstream)/not active (first KE upstream)/active (second KE upstream)'

Criteria assessments	3 high	2 high & 1 moderate	1 high & 2 moderate or 2 high & 1 low	3 moderate or 1 high & 1moderate & 1 low	Any other combination

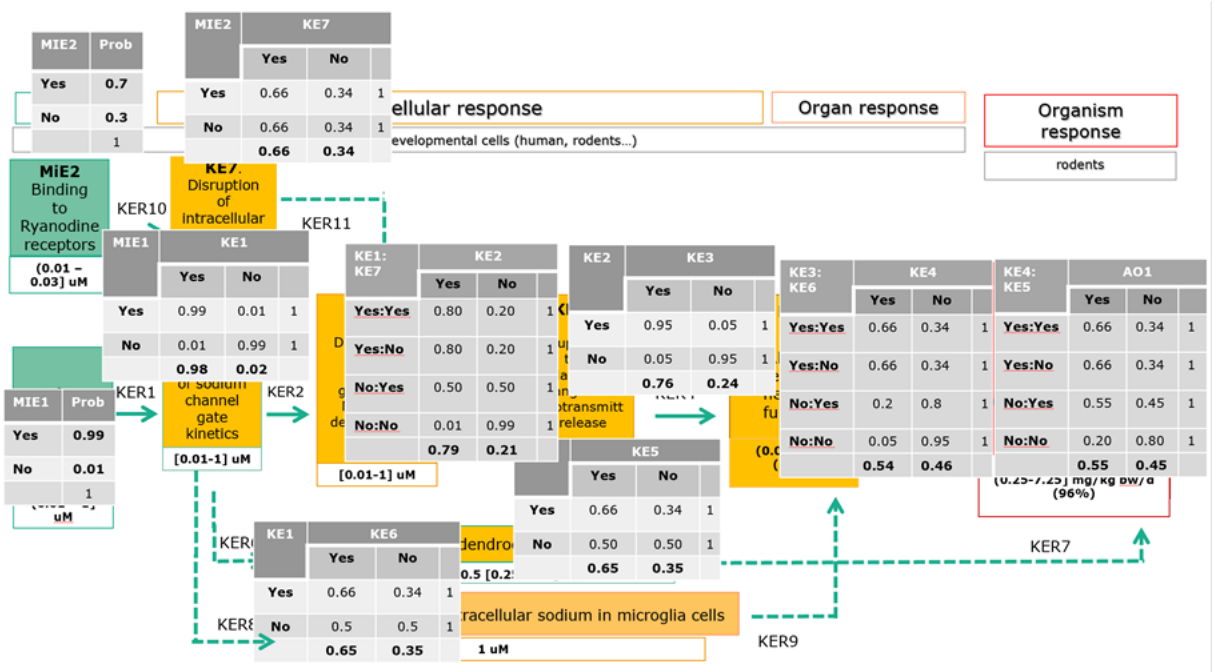
			(≠BP)	(≠BP)	
Probability range of KEx to be activated given the first upstream (parents) KE is not activated whereas the second is activated	(0.95–1.0]	(0.85–0.95]	(0.75–0.85]	(0.66–0.75]	[0–0.66]

Table 6. Rules for combining the criteria assessment – triplet combination 'active (KE downstream)/not active (both KEs upstream)'

Criteria assessments (only biological plausibility and essentiality) Note: essentiality is interpreted in the reverse direction (the higher the essentiality the lower the probability range)	BP high & E low	BP high & E moderate or BP moderate & E low	BP moderate & E moderate	Any other combination
Probability range of KEx to be activated given both the upstream (parents) KEs are not activated	(0.85–1.0]	(0.75–0.85]	(0.66–0.75]	[0–0.66]

The results of the expert elicitation process are reported in Annex C. The related conditional probability tables (CPT) are also displayed in Figure 7. The rationale for the assessment of the three criteria for each of the combination of statuses of the conditioning and conditioned events are provided in Table 1 of the Appendix C.

Figure 7. Conditional probability tables for AOP-BN KERs



The conditional probability tables describe the subjective belief of the experts on how probable they consider a KE downstream to be activated/not activated after knowing that the KE upstream has been either activated or not activated. The number in bold indicate the marginal probability distributions associated to each KE (also reported in Table 7).

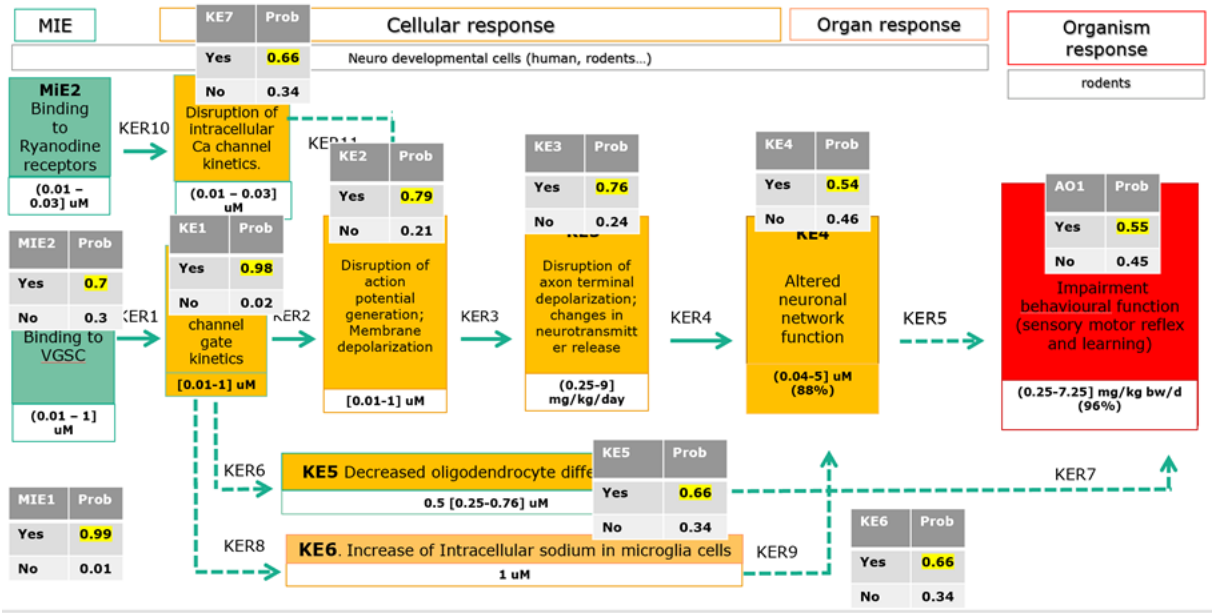
1.4.4. Results (marginal probs, overall certainty, influence analysis)

As discussed above the marginal probability distributions represent a useful tool to predict the most likely state of a KE assuming exposure to the stressor. They are reported in Table 7and displayed in Figure 8 suggesting that, for all the KEs, it is more probable than not to be 'activated/occurring'.

Table 7. Marginal probability distributions for MIE/KE/AO

		Probability	
MIE/KE/AO		To be activated	To be not activated
MIE1	Binding to VGSC	0.99	0.01
MIE2	Binding to Ryanodine receptors	0.7	0.3
KE1	Disruption of sodium channel gate kinetics	0.98	0.02
KE2	Disruption of action potential generation; membrane depolarisation	0.79	0.21
KE3	Disruption of axon terminal depolarisation; changes in neurotransmitter release	0.76	0.24
KE4	altered neuronal network function	0.54	0.46
KE5	decreased oligodendrocyte differentiation	0.66	0.34
KE6	Increase of intracellular sodium in microglia cells	0.66	0.34
KE7	Disruption of intracellular Ca channel kinetics	0.66	0.34
AO	Impairment behavioural function	0.55	0.45

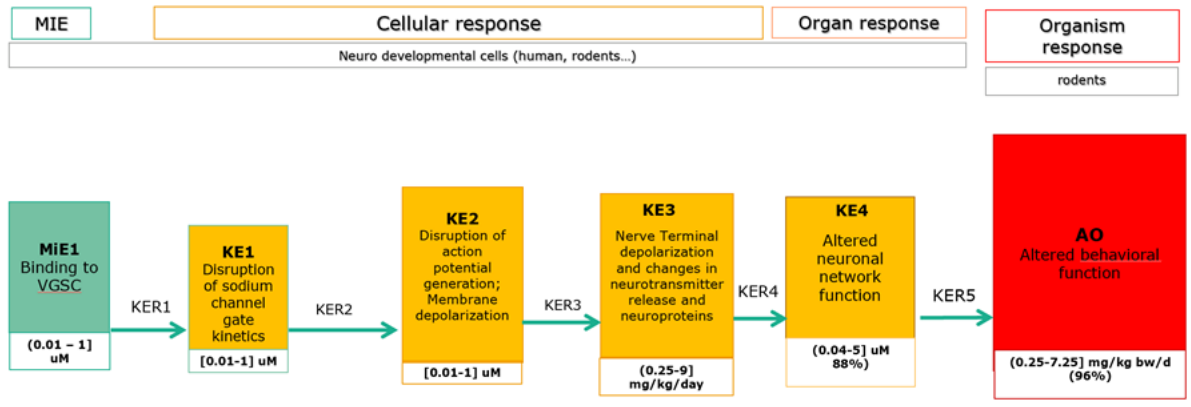
Figure 8. Marginal probability distribution



The marginal probabilities also illustrate how the uncertainty propagates across the pathway. In fact KEs closer to the root of the network (MIEs as triggered by exposure to the stressor – deltamethrin) generally have, with few exceptions, high probability to be activated whereas for KEs related to higher biological complexity (i.e. organ and organism response) this probability tends to approach the maximum uncertainty (i.e. 0.5).

Four single string AOP-BNs could be derived from the AOP network. They are displayed in Figure 9-Figure 12 together with the associated conditional probability tables (marginal probabilities in the bottom line).

Figure 9. Liner string AOP (AOP1)



MIE1	Prob
Yes	0.99
No	0.01
	1

MIE1	KE1		
	Yes	No	
Yes	0.99	0.01	1
No	0.01	0.99	1
	0.98	0.02	

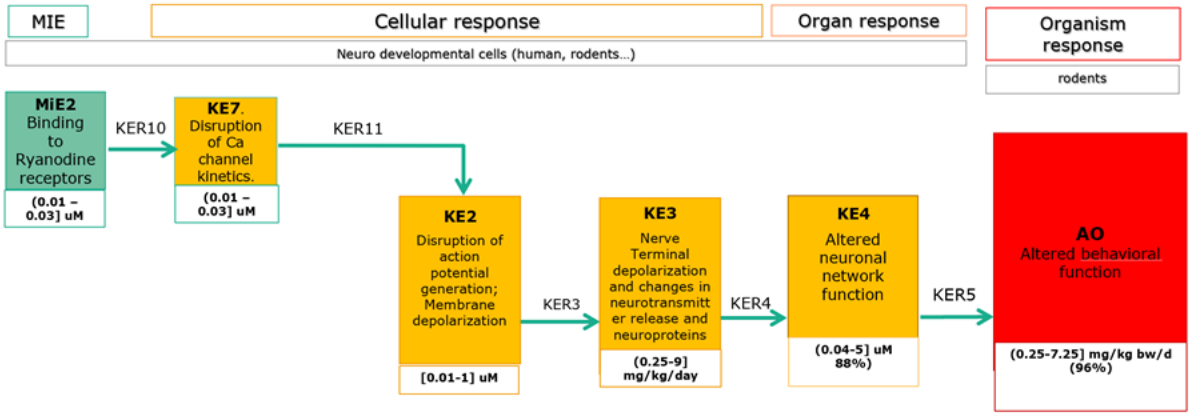
KE1	KE2		
	Yes	No	
Yes	0.80	0.20	1
No	0.33	0.67	1
	0.79	0.21	

KE2	KE3		
	Yes	No	
Yes	0.95	0.05	1
No	0.05	0.95	1
	0.76	0.24	

KE3	KE4		
	Yes	No	
Yes	0.66	0.34	1
No	0.15	0.85	1
	0.54	0.46	

KE4	AO1		
	Yes	No	
Yes	0.66	0.34	1
No	0.43	0.57	1
	0.55	0.45	

Figure 10. Liner string AOP (AOP2)



MIE2	Prob
Yes	0.70
No	0.30
	1

MIE2	KE7		
	Yes	No	
Yes	0.66	0.34	1
No	0.66	0.34	1
	0.66	0.34	

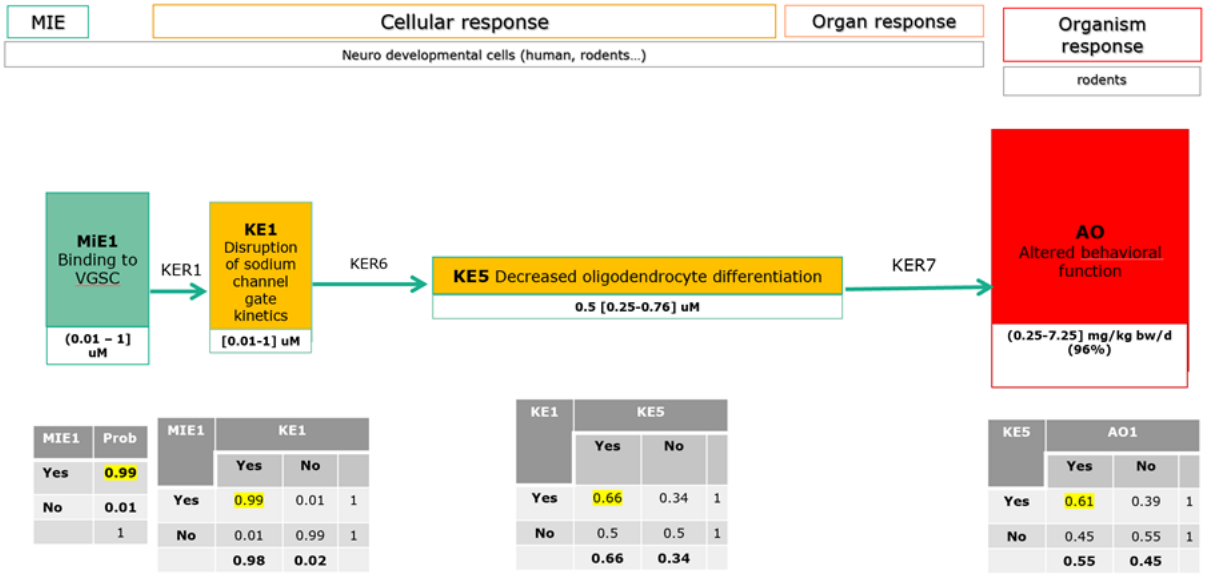
KE7	KE2		
	Yes	No	
Yes	0.79	0.21	1
No	0.78	0.22	1
	0.79	0.21	

KE2	KE3		
	Yes	No	
Yes	0.95	0.05	1
No	0.05	0.95	1
	0.76	0.24	

KE3	KE4		
	Yes	No	
Yes	0.66	0.34	1
No	0.15	0.85	1
	0.54	0.46	

KE4	AO1		
	Yes	No	
Yes	0.66	0.34	1
No	0.43	0.57	1
	0.55	0.45	

Figure 11. Liner string AOP (AOP3)



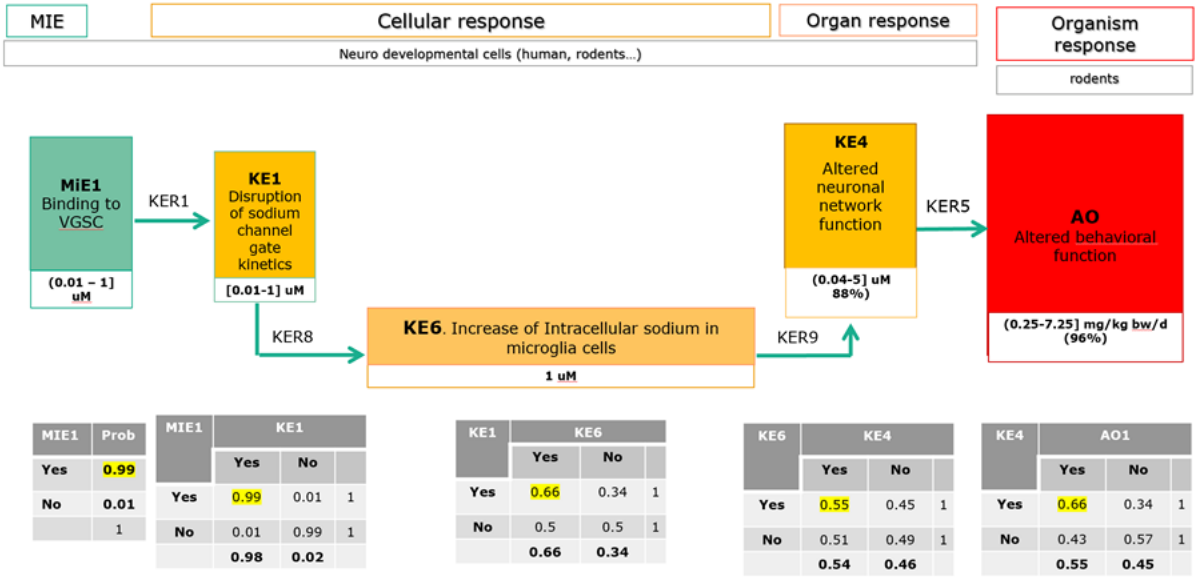
MIE1	Prob
Yes	0.99
No	0.01
	1

MIE1	KE1		
	Yes	No	
Yes	0.99	0.01	1
No	0.01	0.99	1
	0.98	0.02	

KE1	KE5		
	Yes	No	
Yes	0.66	0.34	1
No	0.5	0.5	1
	0.66	0.34	

KE5	AO1		
	Yes	No	
Yes	0.61	0.39	1
No	0.45	0.55	1
	0.55	0.45	

Figure 12. Liner string AOP (AOP4)



MIE1	Prob
Yes	0.99
No	0.01
	1

MIE1	KE1		
	Yes	No	
Yes	0.99	0.01	1
No	0.01	0.99	1
	0.98	0.02	

KE1	KE6		
	Yes	No	
Yes	0.66	0.34	1
No	0.5	0.5	1
	0.66	0.34	

KE6	KE4		
	Yes	No	
Yes	0.55	0.45	1
No	0.51	0.49	1
	0.54	0.46	

KE4	AO1		
	Yes	No	
Yes	0.66	0.34	1
No	0.43	0.57	1
	0.55	0.45	

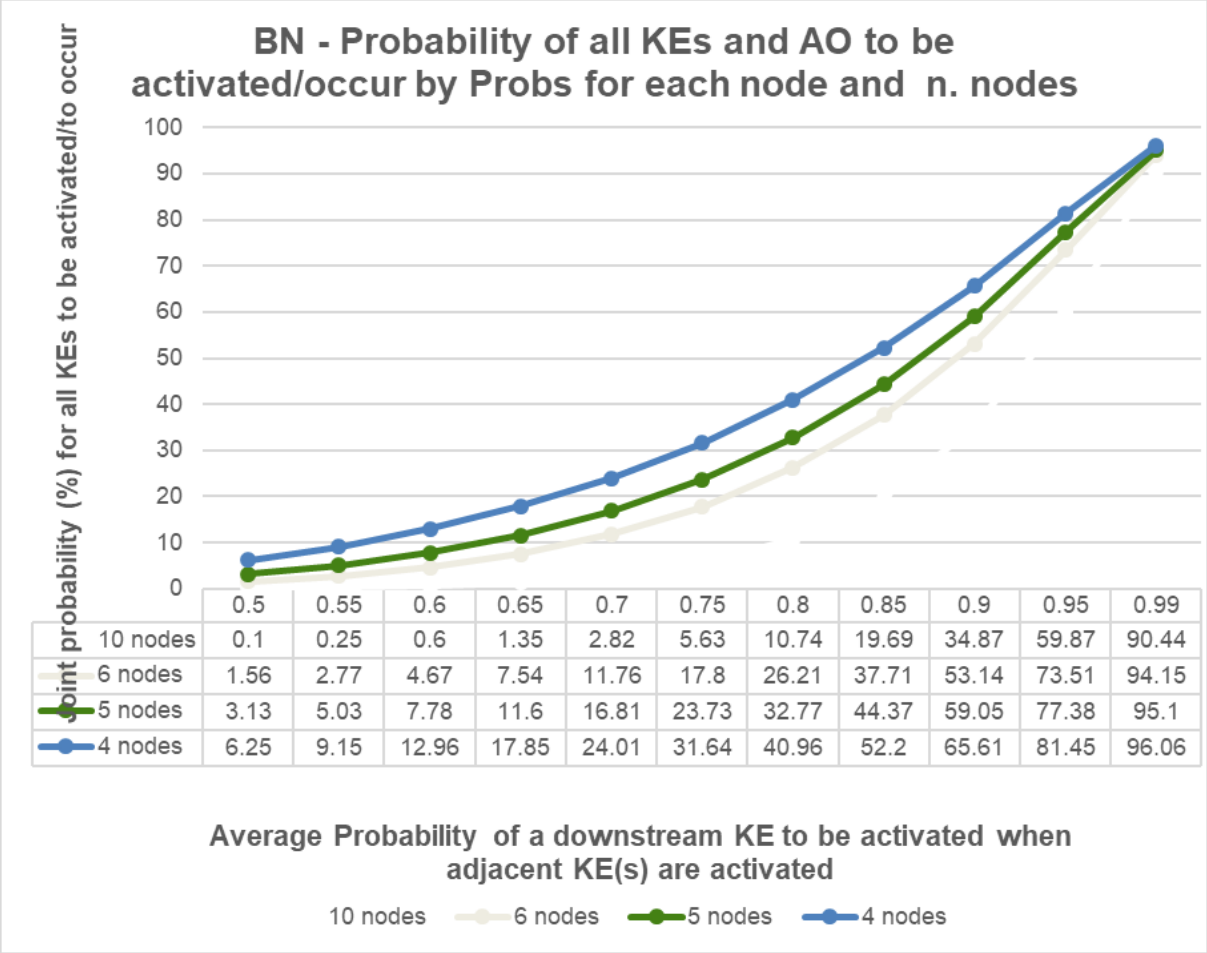
For these liner string AOPs, CPT for pairs have been derived from the CPT for triplets when appropriate.

1.4.5. A measure of the overall certainty of the AOP network and each individual AOP

Resorting onto the Bayesian theorem and property of the nodes in a BN to be conditionally independent from all non-descendent nodes given the parents (conditioning nodes) , the conditional probabilities for the individual KEs to occur under the condition that the upstream KEs were activated, were used to estimate the joint probability that all events (MIEs, KEs, AO) in the network are activated/occur. This probability is 6.5% for the full AOP network.

It is noted that the joint probability for all KEs and the AO to be activated within a network depends also on the number of nodes within the network as shown in Figure 13 (e.g. considering an average probability of 0.95, would allow a joint probability of all KEs to occur of 60% with a network of 10 nodes, 74% with six nodes, 77% with five nodes, 81% with four nodes). Therefore, an interpretation of this probability of 6.5% is neither meaningful in absolute terms, nor by comparison with absolute probabilities for other AOPs or AOP networks. Coverslyt is meaningful to consider the probabilities for each KE being activated, averaged over all KEs in the network. The 6.5% probability for all KEs and the AO in the 10-nodes network being activated corresponds to a situation, when the average probability for the individual KEs being activated under condition that upstream KEs are activated is about 0.76. This is a moderate probability, considering that it represents a situation between random, where KEs would have an average probability of 0.5 and a situation of very high certainty, where KEs would have an average probability of 0.95. In comparison, the probability for all KEs and the AO being activated is 32.4% for the AOP string including only the best documented KERs for deltamethrin (MIE1–KEs1-2-3-4-AO). This corresponds to a situation where KEs would have an average probability of 0.84, which is slightly higher compared with the average probability of 0.76 for the KEs in the complete network (Table 8).

Figure 13. Joint probability of all KEs and AO to be activated by number of nodes in the network and average conditional probability to occur when parents occur by node



	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	0.99
10 nodes	0.1	0.25	0.6	1.35	2.82	5.63	10.74	19.69	34.87	59.87	90.44
6 nodes	1.56	2.77	4.67	7.54	11.76	17.8	26.21	37.71	53.14	73.51	94.15
5 nodes	3.13	5.03	7.78	11.6	16.81	23.73	32.77	44.37	59.05	77.38	95.1
4 nodes	6.25	9.15	12.96	17.85	24.01	31.64	40.96	52.2	65.61	81.45	96.06

Table 8. Joint probability of all KEs to occur, number of nodes and average probability per node for AOP network and single AOPs

	AOP_net	AOP1	AOP2	AOP3	AOP4
N. nodes	10	6	6	4	5
Joint Prob of all KEs to occur (%)	6.53%	32.45%	15.18%	39.41%	23.5%
Average prob per node	0.76	0.83	0.73	0.79	0.75

1.4.6. Influence analysis

The BN approach allows also to perform scenario analyses assessing the impact of hard evidence such as an individual MIE/KE occurring/not occurring with certainty (probability of occurring equals to 1 or probability of not occurring equal to 1 irrespective of the upstream KEs status) on the status of the MIEs/KEs/AO in the network. Computing the difference between the marginal AO probability to occur assuming that each KE is activated and not activated with certainty allows the assessment of the influence of each KE on the final outcome. Then MIEs/KEs can be ranked for their impact and this ranking could be used for recommending further research. For the putative AOP network outlined here, the strongest impact is from the most downstream KEs, i.e. KE4 (altered neuronal network function), followed by KE5 (decreased oligodendrocyte differentiation) – Table 9.

Table 9. Impact of uncertainty in MIEs/KERs on certainty in AO to occur within the putative AOP

Marginal Probability of AO to occur =' ' 0.55					
MIEs/KEs		MIEs/KEs not active with certainty, i.e. Prob(KEx = 0) = 1	MIEs/KEs active with certainty, i.e. Prob(KEx = 1) = 1		
		Probability for AO to occur	Probability for AO to occur	Difference in the Prob for AO to occur	Rank for influence
MIE1	Binding to VGSC	0.47	0.55	0.08	6
MIE2	Binding to Ryanodine receptors	0.55	0.55	0	8
KE1	Disruption of sodium channel gate kinetics	0.46	0.56	0.10	5
KE2	Disruption of action potential generation; membrane depolarisation	0.47	0.58	0.11	4
KE3	Disruption of axon terminal depolarisation; changes in neurotransmitter release	0.46	0.58	0.12	3
KE4	altered neuronal network function	0.43	0.66	0.23	1
KE5	decreased oligodendrocyte differentiation	0.45	0.61	0.16	2
KE6	Increase of intracellular sodium in microglia cells	0.55	0.56	0.01	7
KE7	Disruption of intracellular Ca channel kinetics	0.55	0.55	0	8

2. Remaining sources of uncertainty

Data were lacking to test the robustness of the network structure and the strength of the relationships using external datasets (validation of the AOP network structure). The possibility to perform this exercise rests onto the availability of adequate evidence in the future.

Although demonstrating causality for each of the KERs embedded in the AO-BN was implicitly the purpose of the assessment, this was not possible due the likely presence of latent variables in the pathways.

It is important to acknowledge the subjectivity of all the expert judgements. They are all based on a structured and transparent process that is described in this report to make it repeatable. However it is reasonable to expect that, applying the same process, a different group of experts would end up with different conclusions in terms of uncertainty assessment and numerical estimates for the various probability measures.

3. Software

Statistical analyses were carried out using R version R-3.6.0 (R Core Team, 2013) and Rstudio version 1.2.1335. The BN modelling was performed using the package ‘bnlearn’, ‘Rgraphviz’ and ‘gRain’ (Scutari and Denis, 2015).

References

Chandler J, Cumpston M, Thomas J, Higgins JPT, Deeks JJ and Clarke MJ, 2020. Chapter I: Introduction. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ and Welch VA (eds). Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane, 2020. Available from www.training.cochrane.org/handbook.

EFSA (European Food Safety Authority), 2014. Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA Journal 2014;12(6):3734. [278 pp.] doi:10.2903/j.efsa.2014.3734

EFSA (European Food Safety Authority), Martino L, Aiassa E, Halldórsson TI, Koutsoumanis PK; Naegeli H, Baert K, Baldinelli F, Devos Y, Lodi F, Lostia A, Manini P, Merten C, Messens W, Rizzi V, Tarazona J, Titz A and Vos S, 2020. Draft framework for protocol development for EFSA's scientific assessments. EFSA supporting publication 2020:EN-1843. 46 pp. doi:10.2903/sp.efsa.2020.EN-1843

EFSA (European Food Safety Authority) 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 2018a;16(1):5123, 39 pp. https://doi.org/10.2903/j.efsa.2018.5123ISS

EFSA (European Food Safety Authority) 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, Germini A, Martino L, Merten C, Mosbach-Schulz O, Smith A and Hardy A, 2018. Scientific Opinion on the principles and methods behind EFSA's guidance on uncertainty analysis in scientific assessment. EFSA Journal 2018b;16(1):5122, 235 pp. https://doi.org/10.2903/j.efsa.2018.5122ISS

Fedak KM, Bernal A, Capshaw ZA and Gross S (2015): Applying the Bradford–Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. Emerging Themes in Epidemiology, 12, 14 DOI 10.1186/s12982-015-0037-4

Fenton N and Niel M, 2012. Risk assessment and decision analysis with Bayesian networks. Chapman and Hall/CRC Press: NW Boca Raton, FL · United States

Jaworska J, Dancik Y, Kern P, Gerberick F, and Natsch A, 2013. Bayesian integrated testing strategy to assess skin sensitization potency: from theory to practice. Journal of Applied Toxicology 33, 1353–1364.

Jeong J, Song T, Chatterjee N, Choi I, Kyung Cha Y and Choi J, 2019. Developing adverse outcome pathways on silver nanoparticle-induced reproductive toxicity via oxidative stress in the nematode Caenorhabditis elegans using a Bayesian network model. Nanotoxicology, 12(10), 1182–1197. doi: 10.1080/17435390.2018.1529835. Epub 2019 Jan 21. PMID: 30663905.

LaLone CA, Ankley GT, Belanger SE, Embry MR, Hodges G, Knapen D, Munn S, Perkins EJ, Rudd MA, Villeneuve DL, et al., 2017. Advancing the adverse outcome pathway framework—an international horizon scanning approach. Environmental Toxicology and Chemistry, 36(6), 1411–

1421.

Masjosthusmann S, Blum J, Bartmann K, Dolde X, Holzer A-K, Stürzl L-C, Hagen Keßel E, Förster N, Dönmez A, Klose J, Pahl M, Waldmann T, Bendt F, Kisitu J, Suciu I, Hübenthal U, Mosig A, Leist M and Fritsche E, 2020. Establishment of an a priori protocol for the implementation and interpretation of an in-vitro testing battery for the assessment of developmental neurotoxicity. EFSA supporting publication 2020:EN-1938. 152 pp. doi:10.2903/sp.efsa.2020.EN-1938

Moe SJ, Wolf R, Xie L, Landis WG, Kotamäki N and Tollefsen KE, 2021. Quantification of an adverse outcome pathway network by Bayesian regression and Bayesian network modelling. Integrated Environmental Assessment and Management, 17(1), 147-164

Muller EB, Lin S and Nisbet RM (2015): quantitative adverse outcome pathway analysis of hatching in zebrafish with CuO nanoparticles. Environmental Science and Technology, 49(19) 11817-24. doi: 10.1021/acs.est.5b01837

OECD, 2018. Users' Handbook Supplement to the guidance document for developing and assessing AOPs. Series on Testing and Assessment No. 233, Series on Adverse Outcome Pathways No. 1.

O'Hagan A, 2019. Expert Knowledge Elicitation: subjective but scientific. The American Statistician, 73(Suppl 1), 69–81; DOI: 10.1080/00031305.2018.1518265

Perkins EI, Gayen K, Shoemaker JE, Antczak P, Burgoon L, Falciani F, Gutsell S, Hodges G, Kienzler A, Knapen D, McBride M, Willett C, Doyle III FJ and Garcia-Reyero N, 2019. Chemical hazard prediction and hypothesis testing using quantitative adverse outcome pathways. ALTEX, 36(1), 91–102. doi:10.14573/altex.1808241

Pirone JR, Smith M, Kleinstreuer NC, Burns TA, Strickland J, Dancik Y, Morris R, Rinckel LA, Casey W and Jaworska JS, 2014. Open source software implementation of an integrated testing strategy for skin sensitization potency based on a Bayesian network. ALTEX, 31, 336–340.

Scutari M and Denis JB, 2015. Bayesian networks with examples in R. Taylor & Francis/CRC Press: NW Boca Raton.

Annex A: Questions by lines of evidence

Line of evidence	Species/subject age	Hazard Identification	Expression of uncertainty for HI	Hazard characterisation	Expression of uncertainty for HC
Human	Children	What is the probability that a causal association between individual exposure to deltamethrin in uterus (mothers might have been exposed through dietary and non-dietary sources) and the specific endpoint/adverse outcome occurs	Approximate probability scale (%):[0–10), [10–33), [33–50), [50–66), [66–100]		
In vivo	Rats and mice	Does DM exposure affect this specific endpoint/endpoint category/adverse outcome in a dose–response relationship in rodents exposed during pregnancy and/or post-natal until weaning?	Bounded probability No: Prob < 0.66 Yes: Prob > = 0.66	What is the lowest dose at which DM affects the endpoint in a dose–response relationship in rodents exposed during pregnancy and/or post-natal until weaning ?	Range (uniform distribution) or uncertainty probability distribution for the dose
Zebrafish	Zebrafish	Does DM exposure affect this specific endpoint/endpoint category/adverse outcome in a concentration-response relationship in ZF exposed any time and duration up to 120hours Post Fertilisation	Bounded probability No: Prob(causal association) < 0.66 Yes: Prob(causal association) > = 0.66	What is the lowest concentration at which DM affects the endpoint in ZF exposed any time and duration up to 120hours Post Fertilisation?	Range (assumption of uniform distribution) or uncertainty probability distribution for the concentration
In vitro	Human and rodent cells	Does exposure to DM triggers the specific endpoint/KE as measured in acute and developmental protocol (wash-out yes/no) (assuming a monotonic concentration-response relationship) in human and/or rat and/or mouse neuro cells in development. Exposure might having a duration ranging between several hours up to 28 days. There is no commonly accepted thresholds for biologically significant effects. Biological response is based on expert judgements using EC50/IC50/BMR50 taking into consideration also the uncertainties in the experiments.	Bounded probability No: Prob(causal association) < 0.66 Yes: Prob(causal association) > = 0.66	What is the lowest concentration at which the exposure to DM triggers the KE as measured in acute and developmental protocol (wash-out yes/no) (assuming a monotonic concentration-response relationship) in human and/or rat and/or mouse neuro cells in development. Exposure might having a duration ranging between several hours up to 28 days. There is no commonly accepted thresholds for biologically significant effects. Biological response is based on expert judgements using EC50/IC50/BMR50 taking into consideration also the uncertainties in the experiments.	Range (assumption of uniform distribution) or uncertainty probability distribution for the concentration

: Questions by lines of evidence

Annex B: Uncertainty Tables

Uncertainty table for human line of evidence

Heterogeneity in the results and possible explanation	Can the inconsistencies in the results be justified by heterogeneity across kids age at time of outcome assessment?	Can the inconsistencies in the results be justified by heterogeneity across timing of urine collection during gestation for biomonitoring?	Can the inconsistencies in the results be justified by heterogeneity across exposure characterisation? (i.e. the metabolite type and no. of times it was measured)	Can the inconsistencies in the results be justified by heterogeneity in the way exposure is expressed (e.g. classes of exposure, exposure expressed as continuous variable; LOD)?	Can the inconsistencies in the results be justified by heterogeneity across methods for outcome assessment (e.g. different scales for assessing the same endpoint)?	Can the inconsistencies in the results be justified by heterogeneity across RoB tier?	Can the inconsistencies in the results be justified by heterogeneity in the confounding factors for which the models were adjusted across studies??	Can the inconsistencies in the results be justified by heterogeneity in the statistical analysis and related method for expressing the effect (e.g odd ratios, correlation coefficients, p-values etc.)?		Overall assessment of inconsistencies and uncertainties	Hazard Identification (Yes/No)
Uncertainty in the results	Is sensitivity of the kids age at outcome assessment adequate for the SE/AO in the BoE?	Is timing of urine collection for biomonitoring in the BoE adequate for the SE/AO	Is the metabolite and no. times is measured appropriate?	Is the way the exposure is expressed appropriate?	Is method for outcome assessment adequate for the SE/AO?	Is RoB affecting overall the BoE for the SE/AO?	Are confoundings appropriately considered in the BoE for the SE/AO?	Is statistical analysis appropriate for the BoE for the SE/AO	What is the probability of at least one false positive (familywise alpha error due to		

									multiplicity issue) and the probability of false negative		
Heterogeneity in the results and possible explanation	Can the inconsistencies in the results be justified by heterogeneity across kids age at time of outcome assessment?	Can the inconsistencies in the results be justified by heterogeneity across timing of urine collection during gestation for biomonitoring?	Can the inconsistencies in the results be justified by heterogeneity across exposure characterisation? (i.e. the metabolite type and no. of times it was measured)	Can the inconsistencies in the results be justified by heterogeneity in the way exposure is expressed (e.g. classes of exposure, exposure expressed as continuous variable; LOD)?	Can the inconsistencies in the results be justified by heterogeneity across methods for outcome assessment (e.g. different scales for assessing the same endpoint)?	Can the inconsistencies in the results be justified by heterogeneity across RoB tier?	Can the inconsistencies in the results be justified by heterogeneity in the confounding factors for which the models were adjusted across studies??	Can the inconsistencies in the results be justified by heterogeneity in the statistical analysis and related method for expressing the effect (e.g odd ratios, correlation coefficients, p-values etc.)?		Overall assessment of inconsistencies and uncertainties	Hazard Identification (Yes/No)
Uncertainty in the results	Is sensitivity of the kids age at outcome assessment adequate for the SE/AO in the BoE?	Is timing of urine collection for biomonitoring in the BoE adequate for the SE/AO	Is the metabolite and no. times is measured appropriate?	Is the way the exposure is expressed appropriate?	Is method for outcome assessment adequate for the SE/AO?	Is RoB affecting overall the BoE for the SE/AO?	Are confoundings appropriately considered in the BoE for the SE/AO?	Is statistical analysis appropriate for the BoE for the SE/AO	What is the probability of at least one false positive (familywise alpha error due to multiplicity issue) and the probability of false negative		

BoE: Body of Evidence; SE: Specific Endpoint; AO: Adverse Outcome; RoB: Risk of Bias.

Uncertainty table for in vivo and zebrafish line of evidence

Heterogeneity in the results and possible explanation	Can the inconsistencies in the results be justified by heterogeneity across species/strain	Can the inconsistencies in the results be justified by heterogeneity across sexes	Can the inconsistencies in the results be justified by heterogeneity across exposure duration	Can the inconsistencies in the results be justified by heterogeneity across exposure stage	Can the inconsistencies in the results be justified by some studies indicating maternal or systemic toxicity?	Can the inconsistencies in the results be justified by heterogeneity across RoB (all Qs but Q9)	Can the inconsistencies in the results be justified by heterogeneity in the route of administration		Overall assessment of inconsistencies and uncertainties	Hazard Identification (Yes/No)	Hazard Characterisation (only if HI is yes)
Uncertainty in the results	Is sensitivity of the species/strain in the BoE adequate for the SE/EC/AO?	Is sensitivity of the sex (if only one) in the BoE adequate for the SE/EC/AO?	Is exposure duration in the BoE adequate for the SE/EC/AO?	Is exposure stage in the BoE adequate for the SE/AO?	Is maternal or systemic toxicity affecting studies in the BoE?	Is RoB affecting overall the BoE for the SE/EC/AO?	Is route of administration an issue overall the BoE for the SE/AO?	Is large imprecision (i.e. large SD/SE or wide CI) affecting the BoE?			

ZF: zebrafish; BoE: Body of Evidence; SE: Specific Endpoint; EC: Endpoint Category; AO: Adverse Outcome; RoB: Risk of Bias.
Possible answer for questions on heterogeneity (Yes/No) for the questions on uncertainty (Yes/No/Not Relevant). For both, narrative explanation to describe the motivation for the assessment.

Uncertainty table for in vitro line of evidence

Heterogeneity in the results and possible explanation	Can the inconsistencies in the results be justified by heterogeneity across exposure conditions	Can the inconsistencies in the results be justified by heterogeneity across test systems	Can the inconsistencies in the results be justified by heterogeneity in RoB	Can the inconsistencies in the results be justified by heterogeneity across effect measurements. Note Effect can be expressed in different ways across studies (e.g IC50, BMR30. Control might be baseline value for the same group or a different group e.g. of embryos)	Can the inconsistencies in the results be justified by other items not covered before?		Overall assessment of inconsistencies and uncertainties	Hazard Identification (Yes/No)	Hazard Characterisation (only if HI is yes)
Uncertainty in the results	Is sensitivity of the exposure conditions appropriate for the SE/AO in the BoE ?	Is sensitivity of the test system adequate for the SE/AO?	Is RoB in the BoE adequate for the SE/AO?	Is effect measurement adequate for the SE/AO?	Is sensitivity of the methods beyond exposure conditions adequate for the SE/AO in the BoE?	Is large imprecision (i.e. large SD/SE or wide CI) affecting the BoE?			

ZF: zebrafish; BoE: Body of Evidence; SE: Specific Endpoint; AO: Adverse Outcome; RoB: Risk of Bias.
Possible answer for questions on heterogeneity (Yes/No) for the questions on uncertainty (Yes/No/Not Relevant)

Annex C: Assessment of the conditional probability distributions

Conditioning events	Conditioned event	Criteria	Consensus assessment	Probability range	Consensus probability value P(conditioned event =' 1'/conditioning event(s))	Probability value P(conditioned event =' 0'/conditioning event(s))
K0 = 1	MIE1	Biological plausibility	High	(0.95–1]	0.99	0.01
		Essentiality	High			
K0 = 1	MIE2	Biological plausibility	Moderate	(0.66–0.85]	0.7	0.3
		Essentiality	Moderate			
MIE1 = 1	KE1	Biological plausibility	High	(0.95–1]	0.99	0.01
		Essentiality	High			
		Dose/temporal concordance	High			
MIE1 = 0		Biological plausibility	Low	[0–0.66]	0.01	0.99
		Essentiality	High			
KE1 = 1	KE5	Biological plausibility	Moderate	[0–0.66]	0.66	0.34
		Essentiality	Low			
		Dose/temporal concordance	Low			
KE1 = 0		Biological plausibility	Low	[0–0.66]	0.5	0.5
		Essentiality	Low			

KE1 = 1	KE6	Biological plausibility	Moderate	[0–0.66]	0.66	0.34
		Essentiality	Moderate			
		Dose/temporal concordance	Low			
KE1 = 0		Biological plausibility	Low	[0–0.66]	0.5	0.5
		Essentiality	Low			
MIE2 = 1	KE7	Biological plausibility	Moderate	(0.66–0.75]	0.66	0.34
		Essentiality	Moderate			
		Dose/temporal concordance	Moderate			
MIE2 = 0		Biological plausibility	Moderate	(0.66–0.75]	0.66	0.34
		Essentiality	Moderate			
KE1 =' 1' & KE7 = 1	KE2	Biological plausibility Essentiality Dose/temporal concordance	High High Low	(0.75–0.85]	0.8	0.2
KE1 =' 1' & KE7 = 0		Biological plausibility Essentiality Dose/temporal concordance	High High Low	(0.75–0.85]	0.8	0.2
KE1 =' 0' & KE7 = 1		Biological plausibility Essentiality Dose/temporal concordance	Moderate Low Low	[0–0.66]	0.5	0.5
KE1 =' 0' & KE7 = 0		Biological plausibility	Low	[0–0.66]	0.01	0.99
		Essentiality	High			
KE2 = 1	KE3	Biological plausibility	High	(0.95–1]	0.95	0.05
		Essentiality	High			
		Dose/temporal concordance	High			
KE2 = 0		Biological plausibility	Low	[0–0.66]	0.05	0.95
		Essentiality	High			
KE3 =' 1' & KE6 = 1	KE4	Biological plausibility	High	[0–0.66]	0.66	0.34
		Essentiality	Low			

		Dose/temporal concordance	Low			
KE3 =' 1' & KE6 = 0		Biological plausibility	High	[0–0.66]	0.66	0.34
		Essentiality	Low			
		Dose/temporal concordance	Low			
KE3 =' 0' & KE6 = 1		Biological plausibility	Low	[0–0.66]	0.2	0.8
		Essentiality	Low			
		Dose/temporal concordance	Low			
KE3 =' 0' & KE6 = 0		Biological plausibility	Low	[0–0.66]	0.05	0.95
		Essentiality	Moderate			
KE4 =' 1' & KE5 = 1	AO	Biological plausibility	Moderate	[0–0.66]	0.66	0.34
		Essentiality	Low			
		Dose/temporal concordance	Moderate			
KE4 =' 1' & KE5 = 0		Biological plausibility	Moderate	[0–0.66]	0.66	0.34
		Essentiality	Low			
		Dose/temporal concordance	Moderate			
KE4 =' 0' & KE5 = 1		Biological plausibility	Moderate	[0–0.66]	0.55	0.45
		Essentiality	Low			
		Dose/temporal concordance	Low			
KE4 =' 0' & KE5 = 0		Biological plausibility	Low	[0–0.66]	0.2	0.8
		Essentiality	Low			