Toxicology and Applied Pharmacology 489 (2024) 116995

Contents lists available at ScienceDirect

Toxicology and Applied Pharmacology

journal homepage: www.elsevier.com/locate/taap

Weight of evidence evaluation of the metabolism disrupting effects of triphenyl phosphate using an expert knowledge elicitation approach

Claire Beausoleil [a,*], Anne Thébault [a], Patrik Andersson [b], Nicolas J. Cabaton [c], Sibylle Ermler [d], Bernard Fromenty [e], Clémentine Garoche [f], Julian L. Griffin [l], Sebastian Hoffmann [g], Jorke H. Kamstra [h], Barbara Kubickova [i], Virissa Lenters [h], Vesna Munic Kos [j], Nathalie Poupin [c], Sylvie Remy [k], Maria Sapounidou [b], Daniel Zalko [c], Juliette Legler [h], Miriam N. Jacobs [i], Christophe Rousselle [a]

[a] French Agency for Food, Environmental and Occupational Health and Safety (Anses), 94701 Maisons-Alfort, France
[b] Chemistry Department, Umeå University, SE-901 87 Umeå, Sweden
[c] INRAE. UMR1331 Toxalim (Research Center in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UT3, 31027 Toulouse, France
[d] Department of Life Sciences, Centre of Genome Engineering and Maintenance, College of Health, Medicine and Life Sciences, Brunel University London, UB8 3PH Uxbridge, United Kingdom
[e] INSERM, Univ Rennes, INRAE, Institut NUMECAN (Nutrition Metabolisms and Cancer) UMR_A 1341, UMR_S 1317, F-35000 Rennes, France
[f] Institut de Recherche en Cancérologie de Montpellier (IRCM), Inserm U1194, Université Montpellier, Institut Régional du Cancer de Montpellier (ICM), Montpellier, France
[g] Seh Consulting + Services, Stembergring 15, 33106 Paderborn, Germany
[h] Institute for Risk Assessment Sciences, Department of Population Health Sciences, Utrecht University, Utrecht, the Netherlands
[i] Radiation, Chemical and Environmental Hazards (RCE), Department of Toxicology, UK Health Security Agency (UKHSA), Harwell Science and Innovation Campus, Chilton OX11 0RQ, Oxon, United Kingdom
[j] Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden
[k] Flemish Institute for Technological Research (VITO), 2400 Mol, Belgium
[l] The Rowett Institute, Foresterhill Health Campus, University of Aberdeen, Aberdeen, UK

ARTICLE INFO

Editor: Lawrence Lash

Keywords:
Metabolism-disrupting chemicals
Triphenyl phosphate (TPP)
Obesity
PPARγ
Weight of evidence
Elicitation

ABSTRACT

Identification of Endocrine-Disrupting Chemicals (EDCs) in a regulatory context requires a high level of evidence. However, lines of evidence (*e.g.* human, *in vivo*, *in vitro* or *in silico*) are heterogeneous and incomplete for quantifying evidence of the adverse effects and mechanisms involved. To date, for the regulatory appraisal of metabolism-disrupting chemicals (MDCs), no harmonised guidance to assess the weight of evidence has been developed at the EU or international level. To explore how to develop this, we applied a formal Expert Knowledge Elicitation (EKE) approach within the European GOLIATH project. EKE captures expert judgment in a quantitative manner and provides an estimate of uncertainty of the final opinion. As a proof of principle, we selected one suspected MDC -triphenyl phosphate (TPP) - based on its related adverse endpoints (obesity/adipogenicity) relevant to metabolic disruption and a putative Molecular Initiating Event (MIE): activation of peroxisome proliferator activated receptor gamma (PPARγ). We conducted a systematic literature review and assessed the quality of the lines of evidence with two independent groups of experts within GOLIATH, with the objective of categorising the metabolic disruption properties of TPP, by applying an EKE approach. Having followed the entire process separately, both groups arrived at the same conclusion, designating TPP as a "suspected MDC" with an overall quantitative agreement exceeding 85%, indicating robust reproducibility. The EKE method provides to be an important way to bring together scientists with diverse expertise and is recommended for future work in this area.

* Corresponding author.
E-mail address: claire.beausoleil@anses.fr (C. Beausoleil).

https://doi.org/10.1016/j.taap.2024.116995
Received 18 March 2024; Received in revised form 5 June 2024; Accepted 7 June 2024
Available online 9 June 2024

1. Introduction

Recent estimates suggest that overweight and obesity cause >1.2 million deaths across the WHO's European Region every year (Afshin et al., 2017; Brock et al., 2020). Emerging evidence indicates that xenobiotic chemicals can have obesogenic effects, referred as "metabolism-disrupting chemicals" (MDCs), or metabolic disruptors that can alter any aspect of metabolism (Heindel et al., 2017). MDCs are generally suspected to contribute to the incidence of obesity and related metabolic disorders such as type II diabetes and non-alcoholic fatty liver disease (Legler et al., 2020).

In 2022, new hazard categories for Endocrine Disrupting Chemicals (EDCs) have been proposed to be included in the Classification and Labelling of Products regulation (CLP Regulation (EC) No 1272/2008 (EC, 2023)), particularly in relation to oestrogen, androgen, thyroid and steroidogenesis (EATS) modalities. As with the current CLP guidance for classification of carcinogenic, mutagenic and reprotoxic (CMR) substances applicable in Europe (EC, 2017), the possibility offered by CLP to classify EDCs in different categories depending on the level of evidence, would enable more effective accounting of uncertainties and facilitate expert judgment in reaching a conclusion. In addition, this categorisation would allow tailored regulatory implementation according to sectorial legislations and data requirements. Within European chemical regulations, criteria to identify EDCs have been proposed that require information on a chemicals' endocrine mode of action (MoA) and related adverse effects relevant for human health (ECHA, EFSA, and JRC, 2018). However, whilst MDCs are suspected to play an important role in the worldwide epidemic of metabolic disorders, to date there are no standardised approaches that can be used for regulatory assessment. It is thus of paramount importance to not only develop standardised test methods but also to derive a weight-of-evidence (WoE) approach to assess and possibly identify MDCs.

Expert group evaluation based on a qualitative ordinal scale (*i.e.* "known", "presumed", "suspected") have severe limitations. Qualitative wording has subjective interpretation, thus terminology understanding may differ between experts and organisations, and this has been analysed utilising and comparing with numerical probabilities (Morgan, 2014). Differences in the ways that hazard and risk assessment organisations evaluate similar bodies of information have also given rise to different classifications/evaluations by organisations, as seen for example in recent years for some pesticides (Tarazona et al., 2017) and also bisphenol A (Zoeller et al., 2023). In an effort to build consensus, by improving mutual understanding and interpretation of data, here we take a multidisciplinary approach and use the WoE methodology as proposed by EFSA (EFSA, 2017) with a systematic review and quality assessment of lines of evidence, which also integrates the evidence using a formal elicitation process. This multidisciplinary approach builds upon social science and participatory approaches, and applies them specifically to the scientific and regulatory community. A formal Expert Knowledge Elicitation (EKE) refers to the drawing out of the opinion of a group of experts in a quantitative way, taking into account the uncertainty directly in the estimates (EFSA, 2014). Furthermore, these quantitative estimates can be translated in a harmonised way in ordinal categories, taking into account quantified uncertainty (Anses, 2021). The term 'elicitation' has many meanings, all of which represent different aspects of the general meaning of 'drawing out some information that is needed'. EKE clearly refers to the drawing out of knowledge from one or more experts (EFSA, 2014).

On the basis of these considerations, the aim of this work, conducted as part of the EU-funded Horizon 2020 GOLIATH project (Legler et al., 2020) (https://beatinggoliath.eu/; https://cordis.europa.eu/project/id/825489), was to assess the weight of evidence using an EKE approach, with the intention of categorising a chemical into one of five distinct EDC categories: known; presumed; suspected; not categorised; and non-MDC: "metabolism disrupting compounds" (MDCs) are natural and anthropogenic chemicals that can promote metabolic changes that can ultimately result in obesity, diabetes, and/or fatty liver in humans".

A first prioritisation step was to select one chemical from the six chemicals scrutinized in the GOLIATH project, namely, Bisphenol A (BPA), Perfluorooctanoic acid (PFOA), p,p Dichlorodiphenyldichloroethylene (p,p DDE), Tributyltin (TBT), Triclosan (TCS) and Triphenyl phosphate (TPP), together with a preliminary assessment of adverse effects and potential MoA with sufficient level of evidence to enter into the elicitation process. For this purpose, the level of evidence for the link between individual chemicals and health effect was assessed using a plausibility database previously developed at ANSES for another EU-funded project (ATHLETE-About - Athlete (athleteproject.eu) (Colzin et al., 2024)) and adapted herein.

Following this approach for the six candidates, we retrieved the conclusions of agency reports or published reviews (between 2015 and 2021) regarding metabolic effects. For each chemicals-outcome pair, conclusions in three streams of evidence (epidemiological, toxicological and mechanistic) were translated into stream-specific Levels of Evidence (LoEs) and then combined into an overall LoE ranging from "very unlikely" to "very likely" (for more detail see Colzin et al., 2024 (Colzin et al., 2024). Based on this preliminary work, TPP was selected and the endpoints related to metabolism and obesity identified were: diabetes in offspring, obesity and adipogenesis in offspring, body weight and cholesterol changes. Among these endpoints, obesity was selected for its relevance for humans and adipogenesis for its relevance in experimental models. Adipocytes originate from mesenchymal stem cells, these are multipotent cells that can differentiate into various cell types, including adipocytes. Adipogenic signalling pathways (*e.g.* Wnt, BMP or Hedgehog signalling) and primary molecular initiating events (MIE) and transcription factors such as PPARγ, are considered as master regulators of genes in adipogenesis (Tontonoz and Spiegelman, 2008; Jakab et al., 2021) and their differential expression/ activation determines the adipocytic phenotype. On this basis, we selected PPARγ as the putative MIE to be evaluated in the elicitation exercise. It is acknowledged that there are other important co-regulators as C/EBPs, sterol regulatory element-binding protein, and the glucocorticoid receptor.

The overall level of evidence for the effects related to metabolic disorders for TPP is described in the Table below. The overall level of evidence for the effects related to metabolic disorders for TPP is described in the Table 1 below.

Overall, even if the LoE of TPP was lower compared to the other chemicals partly due to a less extensive data set compared to compounds such as PFOA or BPA, the data generated on TPP itself on PPARγ mediated mechanism or mimicking the insulin signalling pathway and stimulating glucose uptake were considered as a good candidate for a quantitative WoE approach.

The overall objective of this study was to show how the weight of evidence analysis combined with an elicitation approach can be applied to MDCs, and more generally to EDCs. In order to assess the robustness and reproducibility of the method the process of quality assessment of publications and elicitation was conducted in two separate groups of GOLIATH consortium members, each of whom are experts in different fields.

2. Methods

2.1. Context and problem formulation

The strategy follows the definition of an EDC given in WHO/IPCS, 2002 (WHO/IPCS, 2002)and is in agreement with the EFSA/ECHA/JRC guidance describing hazard identification for endocrine-disrupting properties for Plant Protection Products (PPP) or Biocidal Products (BP)(ECHA, EFSA, and JRC, 2018). So far, ECHA/EFSA/JRC guidance document describes how to gather, evaluate and consider all relevant information for the assessment of endocrine-disrupting properties in order to establish whether the endocrine disruptor (ED) criteria are fulfilled. It should be emphasized that this guidance has been written for

Table 1
Level of Evidence for effects related to metabolic disorders for TPP.

Substance	Institution / Authors	Effects	Overall LoE	Probability of causation
Triphenyl phosphate (TPP) CAS N° 115–86-6	(ANSES (French Agency for Food, Environmental ond Occupationnal Health and Safety, 2018)	Obesity and adipogenesis in offspring	As likely as not	50% [40% - 60%[
		Diabetes in offspring	As likely as not	50% [40% - 60%[
		Body weight (↘)	Unlikely	30% [20% - 40%[
		Birth weight (↗)	Unlikely	30% [20% - 40%[
		Liver function	As likely as not	50% [40% - 60%[
		Cholesterol level	As likely as not	50% [40% - 60%[
	(U.S.EPA, 2020)	Thyroid function	Very unlikely	10% [0% - 20%[
		Body weight (↘)	Likely	70% [60% - 80%[
	(Department of ecology, 2018)	Diabetes and obesity in offspring	As likely as not	50% [40% - 60%[
		Body weight (↘)	As likely as not	50% [40% - 60%[
		Prolactin level	Very unlikely	10% [0% - 20%[
		Liver function	Unlikely	30% [20% - 40%[

data-rich substances. It is important to note, that depending on the data available, a substance can be identified for its potential endocrine-disrupting activity for environment or human health or both. The main vehicle for EDC identification at the EU regulatory level is now foreseen to be the CLP (CLP Regulation (EC) No 1272/2008 (EC, 2023)). This regulation bases its assessment on respective criteria and considers all available relevant information without setting information requirements for classification purposes. The information requirements depend upon the legal frameworks within PPP Regulation (PPPR), BP Regulation (BPR) or Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH).

According to the guidance for the identification of endocrine disruptors (ECHA, EFSA, and JRC, 2018), a substance is considered as having endocrine-disrupting properties if it meets all of the following three criteria:

a) Criteria (1): "It shows an adverse effect in [an intact organism or its progeny]/[non-target organisms], which is a change in the morphology, physiology, growth, development, reproduction or life span of an organism, system or (sub)population that results in an impairment of functional capacity, an impairment of the capacity to compensate for additional stress or an increase in susceptibility to other influences;"
b) Criteria (2): "It has an endocrine mode of action, i.e. it alters the function(s) of the endocrine system;"
c) Criteria (3): "The adverse effect is a consequence of the endocrine mode of action".

For the purpose of this assessment, the approach previously developed by the French Agency for Food, Environmental and Occupational Health and Safety (Anses) to categorise substances of interest as regards to their potential endocrine-disrupting activity (Anses, 2021) was adapted for the GOLIATH project and applied to TPP.

We drafted 4 specific questions to be answered in order to address the 3 EDC identification criteria presented above. The four questions included in Table 2 were adapted from the ANSES 2021 approach to hazard characterization of potential EDCs and relate to determining whether a chemical exposure is an EDC using an EKE approach (Anses, 2021).

The EDC categorisation (known, presumed, suspected, not categorised, not EDC) of the studied substance is then based on the answer to question 4.

To identify the adverse effect of interest related to the metabolic disruption properties of TPP, a systematic review covering several outcomes such as obesity, adipogenesis, metabolic syndrome or lipid metabolism disorder (including dyslipidemia) was performed. Obesity was considered as the more relevant endpoint for human and adipogenesis for the (animal) experimental model. Regarding the potential MoA related to this endpoint, a dedicated search was run (see Supplementary material). On the basis of the collated data, peroxisome proliferator-activated receptor γ (PPARγ) activation was selected as a putative MIE leading to obesity/adipogenesis.

The four questions were then adapted to TPP as given in Fig. 1 below.

Table 2
Specific questions relate to determining whether a chemical exposure is an EDC using an EKE approach.

Question 1	o What is the plausibility that the studied substance has the potential to cause the effect? The adverse effect induced by the substance has to be indicated.
Question 2	o What is the plausibility that the studied substance acts through an endocrine MoA? The pathway has to be indicated, and can concern an endocrine MoA related to oestrogenic (E), androgenic (A), thyroid (T) or steroidogenesis (S) pathways of the substance; but it is not limited to EATS pathways (other endocrine signalling pathways can be considered).
Question 3	o What is the plausibility that the endocrine mode(s) of action induces the adverse effect(s) identified? This question concerns the link (biological plausibility) between the adverse effect and the endocrine MoA, which shall be determined in the light of current scientific knowledge. However, to conclude on the biological plausibility of the link, it may not be necessary to have demonstrated for the substance under evaluation the whole sequence of events leading to the adverse effect. Existing knowledge from endocrinology and/or toxicology may be sufficient to address the link and come to a conclusion biological plausibility between adverse effects and the endocrine activity (ECHA, EFSA, and JRC, 2018).
Question 4	o Knowing the plausibility of QUESTION 1, QUESTION 2 and QUESTION 3, what is the plausibility that the studied substance has the potential to cause the adverse effect through the endocrine MoA? This last question integrates the evidence from QUESTION 1 to QUESTION 3 and relates to both the environment (ENV) and the human health (HH).

2.2. Overview of the whole process

Three stages are described in the EFSA guidance on the WoE approach (EFSA, 2017): assemble the evidence; weigh the evidence; and integrate the evidence. Our overall strategy of WoE is integrating those different stages (Fig. 2). Within these three stages, specific and different steps are described, in the following sections.

Some steps indicated in boxes with solid line (see Fig. 2) were

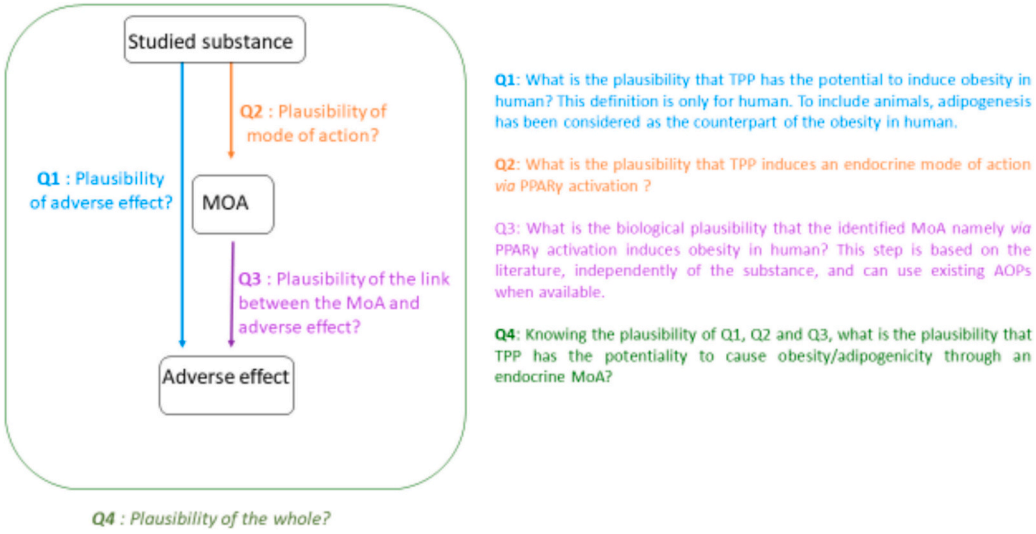

Fig. 1. Integration of the different questions to identify an ED based on the approach developed by Anses (2021).

conducted by a steering committee, which comprised 5 senior toxicologists with expertise in endocrine disruption assessment with 2 specialised in metabolic disorders, and one Facilitator. While for intermediate and latter steps, additional experts were involved (see dotted boxes and italicized text in Fig. 2).

2.3. Collect and assemble the lines of evidence

2.3.1. Systematic review

Based on the preparatory work (see Fig. 2 above), the final review question was formulated as: "What is the evidence available on obesity and adipogenesis of TPP in experimental animal studies or in humans?" In accordance with the EFSA guidance, this review question was described in terms of four key items: population (P), exposure (E), comparator (C) and outcome (O) (PECO). Search terms for these key items were then identified. A combination of search terms for exposure and outcome with the Boolean operator "AND" was used and are described in the supplementary material (see Supplementary material 1). This literature search performed in Scopus and Pubmed followed an iterative approach. Sequential search was run with a first search aimed to identify the scientific papers related to TPP (or TPP's main synonyms) and obesity and a second search aiming to identify scientific papers related to putative TPP metabolites and obesity. These two searches were completed with a third search scrutinizing the literature available on TPP and its putative MIE. While a last search aimed to gather TPP's omics data. For population (P), screening the abstracts and the full-text of the studies allowed us to identify the relevant experimental animal or human studies. For an overview of the scrutinized key words, the reader is invited to refer to the supplementary material (Supplementary material 1).

2.3.2. Allocation to the dedicated questions

The retrieved publications were allocated to the appropriate question.

2.4. Weigh the evidence: assess the relevance and reliability of the evidence

2.4.1. Expert selection

Among the members of the GOLIATH consortium, experts with epidemiological, *in vivo*, *in vitro*, *omics* or *in silico* competencies were recruited on a voluntary basis and assigned by the steering committee in a balanced manner to equally distribute the seniority level and range of experience of the experts to one of the two groups. For the individual elicitation step, each group included 9 experts while for the collective elicitation step, group 1 included 9 experts and group 2 included 8 experts. For each field of expertise, to avoid unbalanced expertise in both groups, two experts were appointed in each group (some experts may have several domains of expertise such as *in silico* and omics), while the Anses team played the role of Facilitator.

2.4.2. Corpus of publications

The completeness of the collected dataset (total number of relevant studies) was further checked by each group during a written consultation phase.

Quality assessment of publications

(1) Evaluation grid: All studies (*i.e. in silico*, *in vitro*, *in vivo*, omics and epidemiological studies) were subjected to a preliminary analysis by an Anses reviewer. Evaluation grids were developed in Excel (see Supplementary material 7) which aimed to gather and extract key elements related to the experimental design and its results, assess its potential limitations and/or whether essentiality according to the ECHA, EFSA, and JRC, 2018 guidance or reversibility of the effects, can be assessed. The evaluation grids were developed in close collaboration with volunteer GOLIATH experts. Particular attention was given to the field of applicability and predictive capacity of *in silico* studies and on the analytical approach (*e.g.*, NMR / LCMS / GCMS / LCMSMS…) in omics studies. When completed, these grids were submitted to one expert from each group for a further critical review of the extracted data.
(2) Evaluation of the reliability and relevance of each scientific paper: Around three to five scientific papers were submitted for the critical review of one expert *per* group. Each critical review was gathered in a dedicated grid specific either to the adversity of the effect or to the implicated MoA.
(3) Elaboration of supportive documents: in a final step, all the individual analyses aiming to assess the reliability and the relevance of each scientific paper were then compiled in two supportive working documents. These documents include specific argumentation and assessment at four qualitative levels of assessment for each criteria (strong, moderate, weak, irrelevant).

The findings derived from the assessments of individual experts were compiled into respective collaborative documents, with one document created for each group. Subsequently, these documents were disseminated to their respective group (Group 1 or Group 2) to serve as

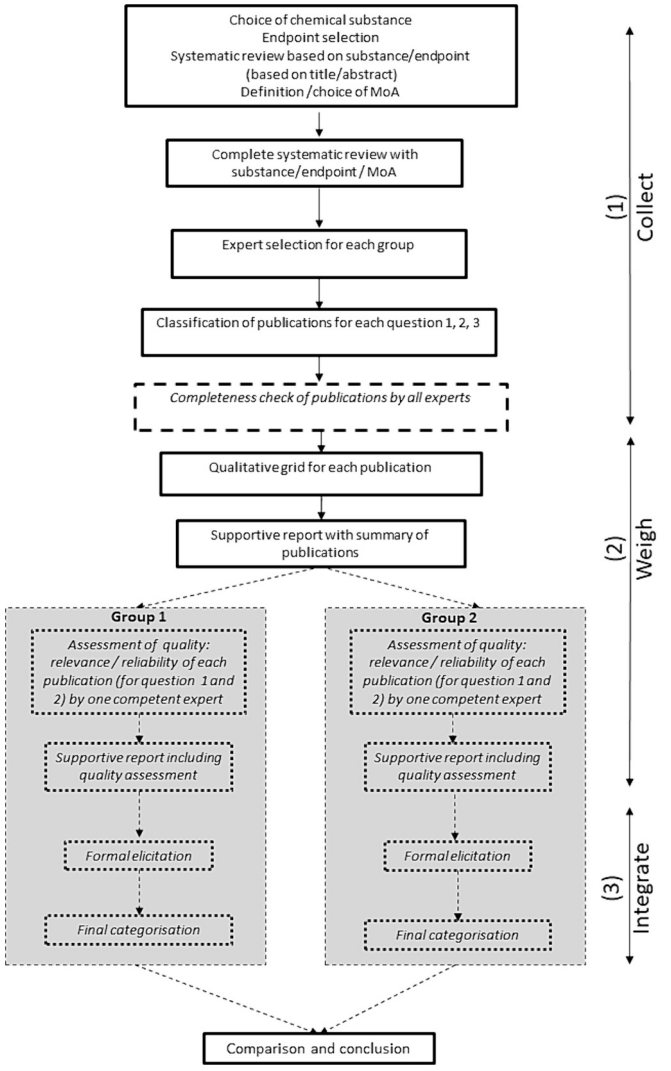

Fig. 2. Flow chart illustrating the integration of the different questions from question 1 to question 4 based on the approach described by Anses (2021).
The boxes with solid line describe the steps where all experts from the dedicated working groups were involved.
The dotted boxes and italicized text describe the steps carried out by the steering committee.

supportive materials during the elicitation phase.

2.5. Integrate the evidence: expert knowledge elicitation (EKE) method

Considering the different pieces of evidence and their experience on the subject, each expert provided values for the 25, 50 and 75% quartiles based on their level of confidence. Elicitation is not only considering studies numbers but it relies on the overall expert's judgment about the available level of information to answer "yes" to the question. Whenever an expert considers that its own competency does not cover all the pieces of evidence, he or she, should rely on the opinion expressed by the other experts in the supportive report. Most importantly at collective stage, they can ask explanations from other experts and explanations provided collectively should help to agree on a final opinion.

Based on previous work (Anses, 2021), the Sheffield method was selected as a formal elicitation method (EFSA, 2014). The Sheffield method is described in different documents such as the EFSA (2014), O'Hagan et al. (2006), as well as in several published studies (Butler et al., 2015; Pietrocatelli, 2008) and online material (http://www.tonyohagan.co.uk/shelf/ecourse.html founded by the United States Office of Naval Research).

The principle of the Sheffield method is based on 6 structured stages:

1. Steering committee constitution,
2. Problem formulation (see questions 1 to 4).
3. Selection of experts: based on complementary expertise, representative of the different aspects of the question,
4. Expert training,
5. Individual elicitation stage (weigh and integrate, see also below),
6. Collective elicitation stage (weigh and integrate see also below).

The selection of the experts and the problem formulation was considered in the preliminary Steps 2 and 3. Based on previous expertise in the process of elicitation, the Anses team played the role of facilitator and guided the experts towards optimally expressing their knowledge *e.g.* through structured forms and quantitative online tools The collective phase of elicitation allows a controlled interaction between experts ("behavioural aggregation") to obtain a consensus quantitative judgment, and exchange of arguments. The main qualities required to achieve successful elicitation are based on the clarity of the questions raised, on a common and agreed view on the definitions used, and on transparency on the process followed and anonymous reporting. Lastly, in order to appraise the reproducibility of the applied weight of evidence approach, two independent groups of experts were set up and their elicitations were carried out independently and in parallel.

2.5.1. Quantitative aspects of expert opinion

The objective of 'formal elicitation' is to capture expert knowledge/judgment on uncertainties and to quantify these in the form of a probability distribution. Formalized methods allow to correctly and quantitatively describe the uncertainty around a desired value (EFSA, 2014), such that the process can be considered reproducible. The parameter of interest is the probability that the answer to the question is "yes". Because the probability is not assessed by data but by expert opinion, the probability is called "a subjective probability". However, this value should be justified by an *ad-hoc* argumentation based on evidence.

In order to simplify and for consistency to ensure methodological reproducibility, it was agreed that:

- The elicited values being a probability, the minimum and maximum limits are bounded between 0 and 1.
- Uncertainty about a probability is classically described by a distribution of values according to a Beta distribution (Vose, 2000). This is the default, and we therefore selected this distribution to characterise the subjective probability.
- In the Sheffield method, we chose the quartile method (EFSA, 2014), where the expert provides the 25, 50 and 75% quantiles.
- Likewise for reasons of simplicity, and ease of understanding to facilitate expert engagement, the quartile summaries were provided for individual and collective elicitation. From the values of these quartiles, a distribution following a Beta distribution is fitted (by the maximum likelihood method). The adjusted distribution obtained makes it possible to describe other characteristics of the uncertainty distribution, such as its credibility interval at 95%, 99%, or the average. This information allows the experts to express their feedback and to validate, or not, their elicitation regarding the distribution obtained. For each of the four questions, each expert was asked to provide their own quartiles for the quantitative aspects, 25%, 50% (median), and 75%, and to check their feedback on the quality of fitting with appropriate tools (from Anses: https://shiny-public.anses.fr/elicittools/ and from Tony O'Hagan-SHELF: the Sheffield elicitation framework: https://jeremy-oakley.shinyapps.io/SHELF-single/
- With respect to representation and feedback with experts, to represent a probability distribution, there are two possibilities: the Probability Density Function (PDF) and the Cumulative Distribution

Function (CDF). The interpretation of the axes of a PDF is as follows: the abscissa axis corresponds to the elicited value, the ordinate axis to the frequencies corrected by a normalisation constant. A CDF describes the probability or quantiles (y-axis) that the value sought is less than or equal to a certain value (x-axis). With the CDF, the quality of fitting can be assessed graphically, quartiles provided by experts should be aligned with the fitted Beta distribution. Graphically if the CDF is predominantly close to 0, the probability of the chemical being an ED is low, and to the contrary, if the CDF is close to 1, the probability of the chemical being an ED is high. In Fig. 3, an example of quantitative opinion is provided where CDF and PDF are shown on the left and right hand side respectively.

2.5.2. Training

Ahead of the official launch of the individual elicitation and due to the Covid-19 pandemic, remote training sessions were organised. These were conducted over several hours, and provided a full specification of the role of the elicitation and/or training on the concepts of statistics, probability, and uncertainty and visualisation interpretation of these as well as a digest of the background and context information resources. These training sessions were followed by a dedicated workshop with a face-to-face meeting in Elche, Spain, May 2022.

At this Elche workshop, participants were again reminded of the aims of this task, as well as the need not to communicate with the members of the parallel group on the subject matter, until guided to do so by the Facilitators.

2.5.3. Individual elicitation opinion

For each of the four questions, experts were asked to provide their own individual quantitative evaluation within evaluation grid to the Facilitators. Each expert in the respective group was provided with the same supportive materials for judgment assessment. Altogether, the publications, the evaluation grid, training material, and a supportive report with qualitative assessment for each publication provided by a competent expert of the group, supported the expert in their response to the four questions, even if some pieces of evidence were not within their field of expertise. A questionnaire and access to the online webtool were provided for each question, so that each expert could document their respective quantified opinion. A reasonable amount of time (several weeks) was given to allow experts to read carefully the different documents and to give their quantitative values and argumentation. For each of the four questions, each expert was asked to provide their response independently from each other to avoid collective biases (meaning they did not share their view or discuss with any other experts at this stage). For each group separately, at the end of the individual elicitation phase, all brought arguments, were organised and ranked for preparing collective elicitation using an interactive visual internet platform using the tool in Klaxoon (Klaxoon, 2024).

2.5.4. Collective elicitation

For each expert group, the Facilitator organised a day-long collective elicitation process (14th of September 2022 and 28th of September 2022 for Group 1 and Group 2, respectively). For each session, the Facilitators moderated and hosted each session.

Each session was organised as following:

a) Short introduction and summary of the previous steps and reminder about the objectives of the meeting,
b) First round table: individual elicitation debriefing:
   - o For each question, each expert provided their argumentation and quantitative results to the group. Unless any clarification was needed, no direct exchange with the other participants was allowed;
   - o All arguments were compiled using the Klaxoon tool;
   - o When the collection of views was completed, duplicates/redundancies were removed;
   - o Discussion on positions/views were initiated new views could be introduced and included in the Klaxoon tables if needed;
c) At the last step, an on-line voting phase allows a ranking of the main arguments.
d) Second round table: collective elicitation phase
   - o This time, the experts were asked to consider what an intelligent and impartial observer might now reasonably believe, having assimilated the experts'different opinions and arguments (EFSA, 2014).
   - o Arguments for the various categories were subsequently ranked with participants utilising a 'like' vote in an interactive visual internet platform, such that the group view was reflected;

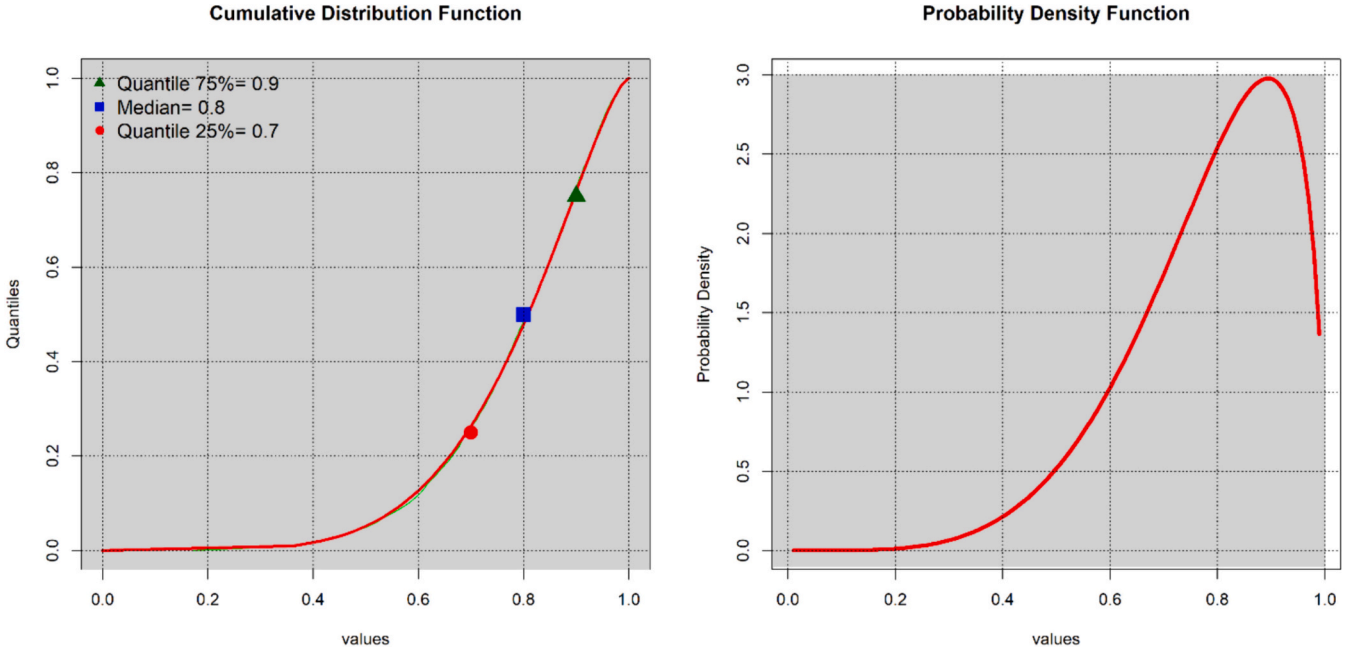

Fig. 3. a) Cumulative distribution function (CDF) and b) the related probability density function (PDF) of a Beta distribution with quartiles 0.7 (in red); 0.8 (in blue) and 0.9 (in green). The red line is the fitting with a Beta distribution.

- o Finally, one expert proposed quantified values for the group. Each expert was invited to approve the proposed values or instead proposed new ones. The process continued until consensus was reached.

e) The (consensus) opinion derived at the collective phase for each group was the final result of the elicitation exercise, applied for the purpose of categorisation of the substance in question.

2.6. Qualitative categorisation in ED categories

According to the method developed previously (Anses, 2021), the correspondence between summary characteristics of the opinion from the group of elicitation experts (median, 5 and 95% percentile of the distribution) for the final integrated question (question 4): "*Knowing the plausibility of* question *1,* question *2 and* question *3, what is the plausibility that the studied substance has the potential to cause the adverse effect through the endocrine MoA?",* and the qualitative category of opinion (known, presumed, suspected, not categorised, not an EDC) was assessed collegially by each group.

The elicitation process establishes Q25, Q50 and Q75 quartiles for the final question. The quartiles are fitted with a Beta distribution, and give 5 and 95 percentiles. The level of evidence given to answer question 4 is then converted to an EDC category following the decision tree below (Fig. 4).

The lowest quantile of the opinion (5% = Q5) gives the first node between non-negligible probability of the chemical not being an ED (below 0.05 or 5% in the categorisation tree) and the other categories (Fig. 4). For the upper categories: known, presumed, suspected, the median of the opinion leads to the final categorisation (known, presumed, and suspected). For the lower categories (not an EDC or not categorised), we differentiate between two situations depending on the quantile 95% (Q95): if the probability of being an ED is also not negligible (Q95 $\geq$ 5), there is too much uncertainty to conclude (category Not Categorised). If this upper bound is strictly below 5%, we consider that the substance is not likely to be an ED (not an ED). The opinion (being/not being an EDC) is not completely symmetric because it can be difficult to raise "perfect" lines of evidence for a group of scientists and also to protect human safety according to the precautionary principle, which is by nature conservative, erring on the side of caution. Examples of applications of this rule in relation with different quantitative opinions are given in appendix of the Anses report (Anses, 2021: https://www.anses.fr/fr/system/files/REACH2019SA0179Ra.pdf).

In summary, the final categorisation is led by the median and the uncertainty (5–95% quantiles) of the quantitative opinion of the group.

2.7. Statistical analysis of the homogeneity of opinion inside group and agreement between groups

For individual elicitation synthesis, the results at the group level were estimated by the medians of quartile 25%, 50% (median), and 75% of all individual elicitations within each group. In order to assess the homogeneity of medians, we use the inter-quartile range of the group level.

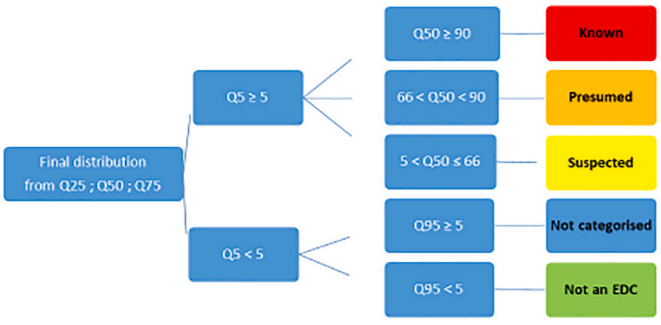

Fig. 4. Decision tree making the link between the level of evidence to be an ED and the final ED categorisation.

Comparing the distribution of opinion of the two groups of experts can be described by comparing summary statistics (medians, quantiles, mean), and for uncertainty analysis, by inter-quartile agreement.

However, detailed comparison of the summary statistics does not take into account the whole distribution of opinion. For that purpose, we utilised an indicator based on CDF. The Kolmogorov-Smirnov statistic quantifies the distance between an empirical and cumulative distribution functions. The CDF of a Beta distribution can be described inside a square of area 1. The difference between the areas under the curve of two cumulative of Beta distribution function is comprised between 0 and 1; by example of a numerical application, between Beta (alpha = 1, beta = 10,000) (mean probability around 10^{-4}) and Beta (alpha = 100,000, beta = 1) (mean probability around 0.9999), the difference of CDF is close to 1; for Beta distribution with equal parameters the difference is 0. Consequently, 1 minus the difference is the level of agreement between the 2 distributions, which is a parameter varying between 0 and 1.

To interpret this indicator, we applied another criterion of agreement between 0 and 1; the Cohen's kappa coefficient of agreement. Cohen suggested the Kappa result can be interpreted as follows: values $\leq$0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement (Cohen, 1960).

3. Results

3.1. Selection of the corpus of publications

The systematic review allowed us to identify 37 publications (see Fig. 5) that are specifically linked to TPP and metabolic disruption either in human or animals, or to TPP and PPARγ activation. The initial data set of 17 publications retrieved from the first search on Scopus and PubMed was augmented after full reading with 20 publications that were cited in these articles. This shows the importance to not only rely on references retrieved only by key word search but to also include targeted reviews by experts in the field, to identify and supplement with more relevant references.

Among the retrieved publications, 9 were allocated to questions 1 and 19 to question 2 (see supplementary material 6). Two reviewers were involved, and disagreements, if any, were resolved after discussion involving a third reviewer.

3.2. Quality assessment of publications

According to the procedure presented above, supportive working documents were built based on the quality assessment of the publications, the completion and careful review of the evaluation grid, the reliability and the relevance check of each scientific paper, and the evaluation of each critical review.

3.3. Results of individual elicitation

The median of individual elicitation for each group is given to the four questions presented in Table 3.

3.3.1. Group 1

Fig. 6 describes results for group 1. The CDF of the group is based on the medians of each of the quartiles for the group. Each different data point (dot) is the result of the quartile for each expert. The dispersion around the median of the group can be seen to be quite homogenous, whatever the question is.

Fig. 6 shows that in some situations, the median of some individual experts was outside the median interquartile range for the group. In particular, one expert (n°8) has a low level of confidence to answer "yes"

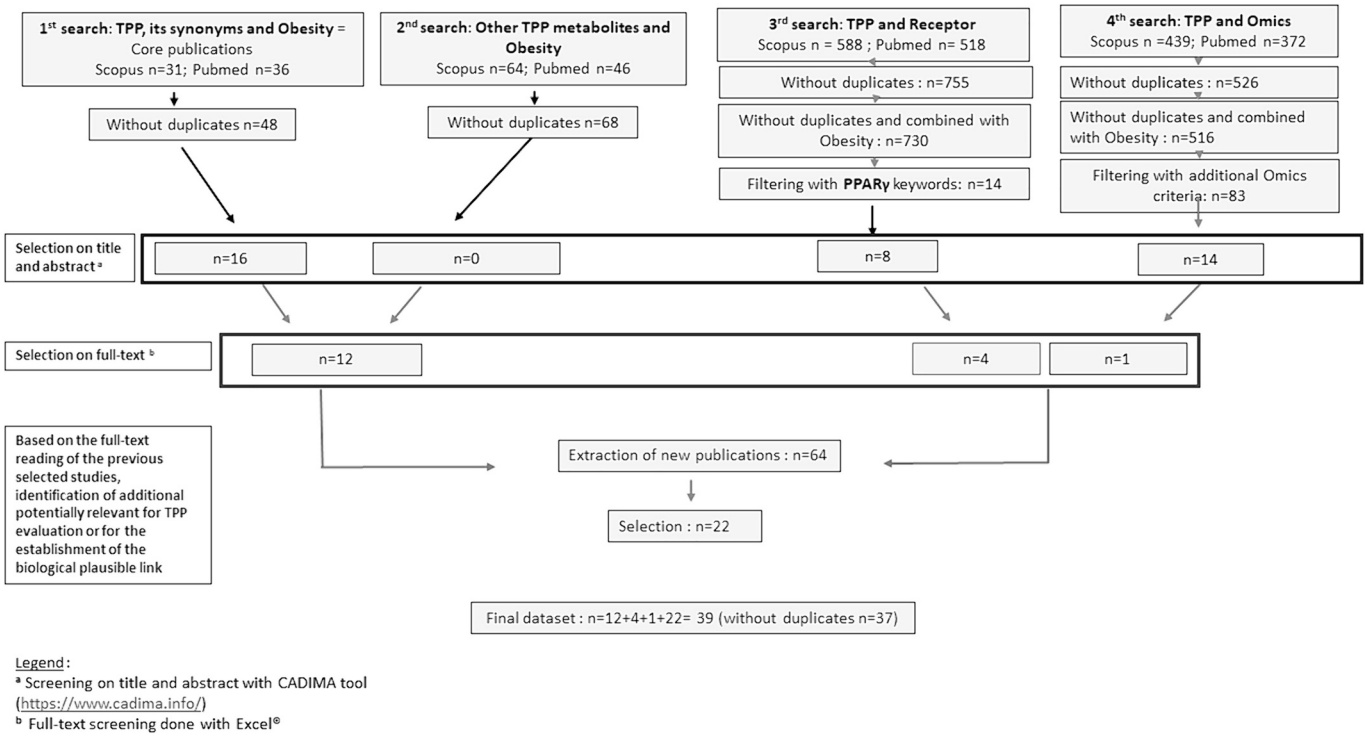

Fig. 5. Prisma diagram.

Table 3
Medians of individual elicitations for each quartiles for group 1 and group 2.

	Group 1			Group2		
Question	Q25	Q50	Q75	Q25	Q50	Q75
Question 1	0.25	0.5	0.65	0.4	0.5	0.7
Question 2	0.5	0.6	0.75	0.6	0.7	0.8
Question 3	0.3	0.6	0.75	0.5	0.7	0.85
Question 4	0.3	0.5	0.6	0.5	0.6	0.7

Legend: Question 1: What is the plausibility that TPP has the potential to induce obesity in human? This definition is only for human. To include animals, adipogenesis has been considered as the counterpart of the obesity in human; Question 2: What is the plausibility that TPP induces an endocrine mode of action *via* PPARγ activation?; Question 3: Q3: What is the biological plausibility that the identified MoA namely *via* PPARγ activation induces obesity in human? This step is based on the literature, independently of the substance, and can use existing AOPs when available; Question 4: Knowing the plausibility of Q1, Q2 and Q3, what is the plausibility that TPP has the potentiality to cause obesity/adipogenicity through an endocrine MoA?

to the questions 1, 3 and 4 or two experts (N°1 and N°6) for question 3. However, no clear trend dividing the group in subgroups of different opinion can be observed.

3.3.2. Group 2

The results are given in Fig. 6. The dispersion of the opinion is higher than for the Group 1. In particular, the experts 4 and 7 give more extreme values than others in this group. As shown in this figure, the medians generated for some of the experts are not in the interquartile range of the group. Experts 1, 4 and 7 give lower medians than the rest of the group.

3.4. Results of collective elicitation

3.4.1. Qualitative assessment

During the final steps of the elicitation exercise, on the collective elicitation day for each group, we progressed through each question (question 1–4, see section 2.1), asking experts to provide supportive and explanatory arguments to justify their quantitative figures. We then discussed altogether what could be the final values at the group level.

In the next sections, the justifications for question 1 to question 4 are presented for each group. For a full overview of the arguments brought by each group and for each question, the reader is invited to refer to Supplementary material 8 to Supplementary material 15) where they are extensively reported and classified according to the number of "likes" with the Klaxoon tool.

Regarding question 1: "*What is the plausibility that TPP has the potential to induce obesity in humans?*" as the obesity definition relates to humans only, increased adiposity in animals has been considered as the counterpart of the obesity in humans. Thus, the following source of evidence was considered: human epidemiological studies but also experimental studies performed in intact animals (*in vivo*) to capture adipogenic properties, and if it may reinforce the level of evidence studies performed in environmental organisms (*e.g.*, fish). Relevance and reliability of studies were also considered. Lastly as this evaluation is focused on hazard assessment only, exposure levels were not considered within this work.

For Group 1, the main uncertainties and responses given to question 1 were that there is only moderate evidence based on animal data (3 *in vivo* studies available) showing that TPP promotes body weight increase and adipogenicity. Two relevant epidemiological studies were identified but some reliability issues were raised. The evidence based on epidemiological data was overall considered weak or not relevant. Lastly, the evidence based on studies focussing on omics (*e.g.*, transcriptomics, metabolomics) was also considered weak.

For Group 2, the main uncertainties and responses given to question 1 were again that there is a low level of evidence based on epidemiological studies, as human studies were not considered directly comparable since different endpoints were assessed (adiposity in Luo et al., 2020a, 2020b and birth weight in Boyle et al., 2019). In addition, evidence based on animal data that TPP increases body weight or body fat mass was considered moderate.

Taken together, both groups agree about the lack of evidence

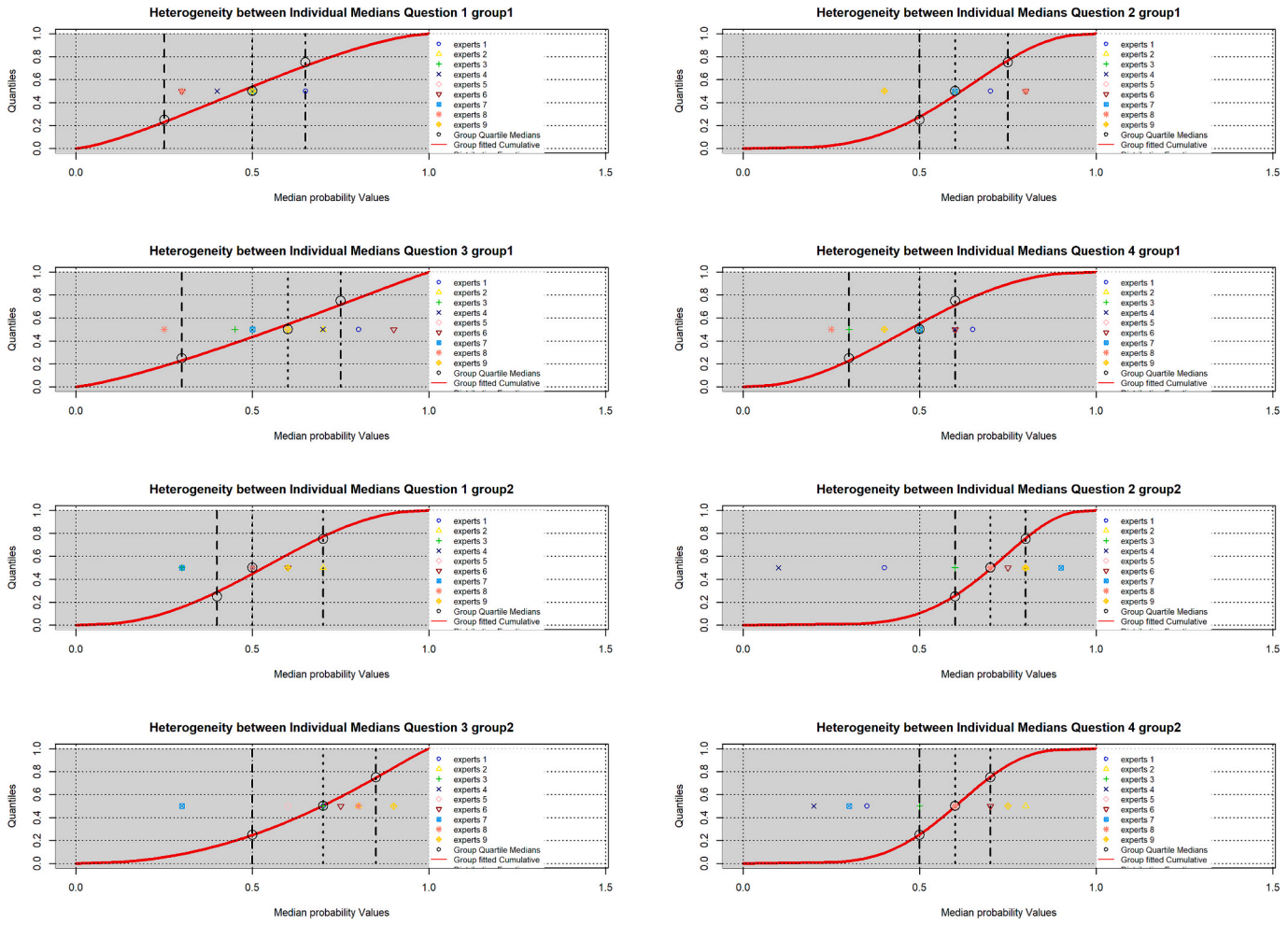

Fig. 6. Heterogeneity of quartiles for each individual expert for each question group 1 and group 2.
Legend: Median of interquartile ranges: vertical long dashed lines; median of the medians: vertical short dashed line; red line: beta fitting of the median of quantiles given by each expert.

coming from human data and the limitations of the animal studies to adequately answer the question whether TPP has the potential to induce obesity in humans.

Regarding question 2: "*What is the plausibility that TPP induces an endocrine mode of action via PPARγ activation*? *in vivo* or *in vitro* mechanistic experimental studies were considered as well as any supportive data on the hypothesised MoA.

For group 1, the main uncertainties and responses given to question 2 were that there is (strong) evidence mainly based on binding, (*in silico*) docking and *in vitro* data showing that TPP binds to and activates PPARγ1 and consistently demonstrates effects in different cell lines, with varied levels of confidence. There are no *in vivo* studies that link TPP to PPARγ activation and obesogenic effects in the corpus of publications. It is quite uncertain whether PPARγ activation is the sole mechanism involved and if this activation shown at very high concentration will be sufficient to activate the entire adverse outcome pathway. Lastly, not all studies have proper controls or adequate description of methods.

For group 2, the main responses given to question 2 indicated were that the studies are overall in agreement regarding the effects observed with TPP. Two studies show PPARγ interaction in *in silico* molecular docking studies and a large amount of convincing data demonstrates a clear interaction with PPARγ in several species. In the three *in vitro* studies that studied adipogenesis, the mechanisms were not examined in sufficient detail. *In vivo* data is considered poor. Omics studies were not considered relevant and did not provide data about the expression of lipogenic genes/proteins associated with an activation of the PPARγ signalling pathway.

Overall, both groups agreed about the interaction of TPP with PPARγ mainly based on *in silico*, binding or *in vitro* studies. But *in vivo* studies were considered limited to answer whether TPP induces an endocrine MoA *via* PPARγ activation.

Regarding question 3: "*What is the biological plausibility that the identified MoA namely via PPARγ activation induces obesity in human?*" scientific literature independent of TPP itself was considered as well as any existing or under investigations Adverse Outcome Pathways (AOPs) if available, as well as levels of evidence coming from other related compounds in a « grouping approach ». The biological plausibility was weighted as follows:

- o Strong: if there is extensive understanding of the Key Event (KE) and Key Event Relationship (KER) based on extensive previous documentation and broad acceptance,
- o Moderate: if the KEs or KERs are plausible based on analogy with accepted biological relationships, but scientific understanding is not completely established,
- o Weak: the structural or functional relationship between the KEs is not understood.

For group 1, the main arguments raised to answer question 3 originates from the lack of investigations to demonstrate if the MoA of PPARγ activation alone is sufficient for obesity induction in humans. It is possible, based on data generated with rosiglitazone (ROSI) and

troglitazone that PPARγ activation induces obesity in humans but epidemiological data demonstrating this link are missing. PPARγ activation may be involved in obesity development, but this is not the only mechanism involved.

For group 2, the main arguments used to answer question 3 were that based on animal data, it is known that PPARγ agonists may elicit energy metabolism (identified with metabolomic approach), change body weight and key obesity markers. However, activation of PPARγ *per se* is insufficient to classify a chemical as an obesogen, as it is only a MIE, and does not sufficiently inform on the occurrence/manifestation of downstream KEs. How relevant animal data is to humans and how well the results *in vivo*, *in vitro* and *in silico* can really be extended to meaningful weight gain in humans is not known. In total, the biological plausibility was considered moderate. It is likely that other factors (*e.g.*, hormonal regulation of satiety and appetite) play a role in the manifestation of obesity.

Both groups consider that even if PPARγ activation could be implicated in the induction of obesity in humans, there is limited evidence that this involves only an endocrine MoA and besides PPARγ activation other mechanisms, either ED or non-ED-related may also be involved.

Regarding question 4 "*Knowing the plausibility of* question 1, question 2 and question 3 *, what is the plausibility that TPP has the potentiality to cause obesity/adipogenicity through an endocrine MoA?"*

For group 1, the main responses given to question 4 were that the PPAR signalling pathway is involved in adipogenesis and obesity. However, epidemiological evidence is limited. The essentiality of the effects depending on the dose and period of exposure is not clearly evaluated in the studies and there is a lack of relevant data that could be used for humans.

For group 2, the main responses given to question 4 where that there is moderate/high consistency between rodent and human *in vitro* system results. There is a lack of clear evidence from epidemiological studies mostly due to lack of data. The specificity is considered weak /moderate. TPP may exert its effects *via* direct PPARγ activation, but other mechanisms may be involved for instance, metabolic adaptation or increased food intake and/or reduced energy expenditure *via* alteration of hypothalamic peptidergic circuits.

The PPARγ signalling pathway is well known to beinvolved in the endpoints of adipogenesis and obesity. However, both groups did consider that epidemiological evidence was limited in this particular case. TPP may exert its obesogenic/adipogenic effects *via* direct PPARγ activation, but other mechanisms may also be involved.

3.4.2. Quantitative assessment

3.4.2.1. Comparison between individual and collective elicitation results in each group. Results are described in Fig. 7. For group 1, when comparing the CDF at the individual and collective level, for each question, the level of confidence slightly increased after the collective discussion (the green curve moved to the right) and the range of uncertainty slightly narrower (the slope of the curve is a little bit sharper). Except for question 3, the medians before and after collective phase did not change noticeably.

This tendency was less obvious for group 2. For example, the answer to question 1 was a bit less certain at the collective stage compared to the individual one; for question 2, the uncertainty increased and for question 3 and question 4, it was pretty much the same (see Fig. 7). For group 2 when comparing the CDF at the individual and collective level, for each question, the medians before and after collective phase did not change noticeably.

3.4.2.2. Comparison between group 1 and 2 elicitation results. The results of the collective elicitation for the two groups are collated in Table 4. Overall, the medians of the two groups were found to be close for question 2 and question 3, and identical for question 4. For question 1, the median was lower for group 2 (0.45) compared to group 1 (0.6). We identified no particular systematic trend between groups. The interquartile range was found to be higher for question 1 in group 1, and higher for question 2 in group 2. Whenever we translated numeric values with a corresponding qualitative category, all the answers in both groups, for each question, felt into the same category, except for question 1. The interquartile range was [0.2 to 0.6].

Notably, for both groups, the categorisation of TPP in answering to question 4, fell into the "suspected ED" category.

For both groups the final ED category concluded for TPP considering its potential to induce obesity in human *via* an endocrine MoA was **"suspected".** For both groups, answers to the four questions at the collective level were similar, as shown in the graphs below (see Fig. 8). However, there were some minor differences in the analyses between both groups for question 1 for which group 1 seems to be a little bit more convinced that TPP may induce an adverse effect linked to obesity/adipogenicity, but with more uncertainty than group 2 (as shown by a less sharp curve for group 1 compared to group 2). The lack of human data supporting such a relationship is the likely explanation as both groups considered this to be a key uncertainty.

The overall quantitative agreement, based on Beta fitting of the quantiles given by the group between the two distributions is given in the bullet list below. The overall agreement between the CDF is >86% for all questions, showing the relative robustness of the method between the two groups for categorising TPP with four questions:

- Question 1: 86.5% agreement between both groups
- Question 2: 89.9% agreement between both groups
- Question 3: 95.8% agreement between both groups
- Question 4: 95.2% agreement between both groups

4. Discussion

Given the absence of standardised hazard assessment approaches for identifying MDCs at the EU level and globally as well and recognizing that the GOLIATH project brings together numerous European experts in the field of metabolic disruption, this study investigates a transparent means of addressing this gap. Specifically, we explore the development and analysis of a comprehensive WoE assessment approach. This entailed the creation of a steering group to support the development of the iterative steps and encourage expert participation, followed by a systematic review of the scientific literature, a mapping of the evidence, an evaluation of the evidence, and finally a comparative weighing of the evidence by an elicitation process. As a collective elicitation approach was quite recently applied in the ED field (Anses, 2021), we considered that it would be a useful and unifying collective exercise for most of the GOLIATH partners to better understand the WoE approach that regulators generally need to follow when assessing chemical hazards. By having a sufficient number of experts, it was also possible to conduct the exercise with TPP, *a priori*tized chemical within the GOLIATH project, in parallel by two independent groups.

As a case study, this WoE approach was conducted for TPP, in relation to obesity with PPARγ activation being the MIE. A systematic review of the scientific literature was performed and 37 relevant papers were retrieved. These papers were used as a basis to constitute the data set that was made available to both groups of experts participating in the elicitation exercise.

Two groups of 8 and 9 experts, with a varied range of training and competencies, were elicited individually and independently, next in dedicated groups, and lastly in a collective phase, with the aim to assess the reproducibility of the overall expert elicitation process. A formal elicitation process adapted from the Sheffield method was applied in particular to address remote meeting needs, during the Covid-19 pandemic. As the judgment is expressed in terms of probability, statistical simplifications were made, the range of values was fixed *a priori*

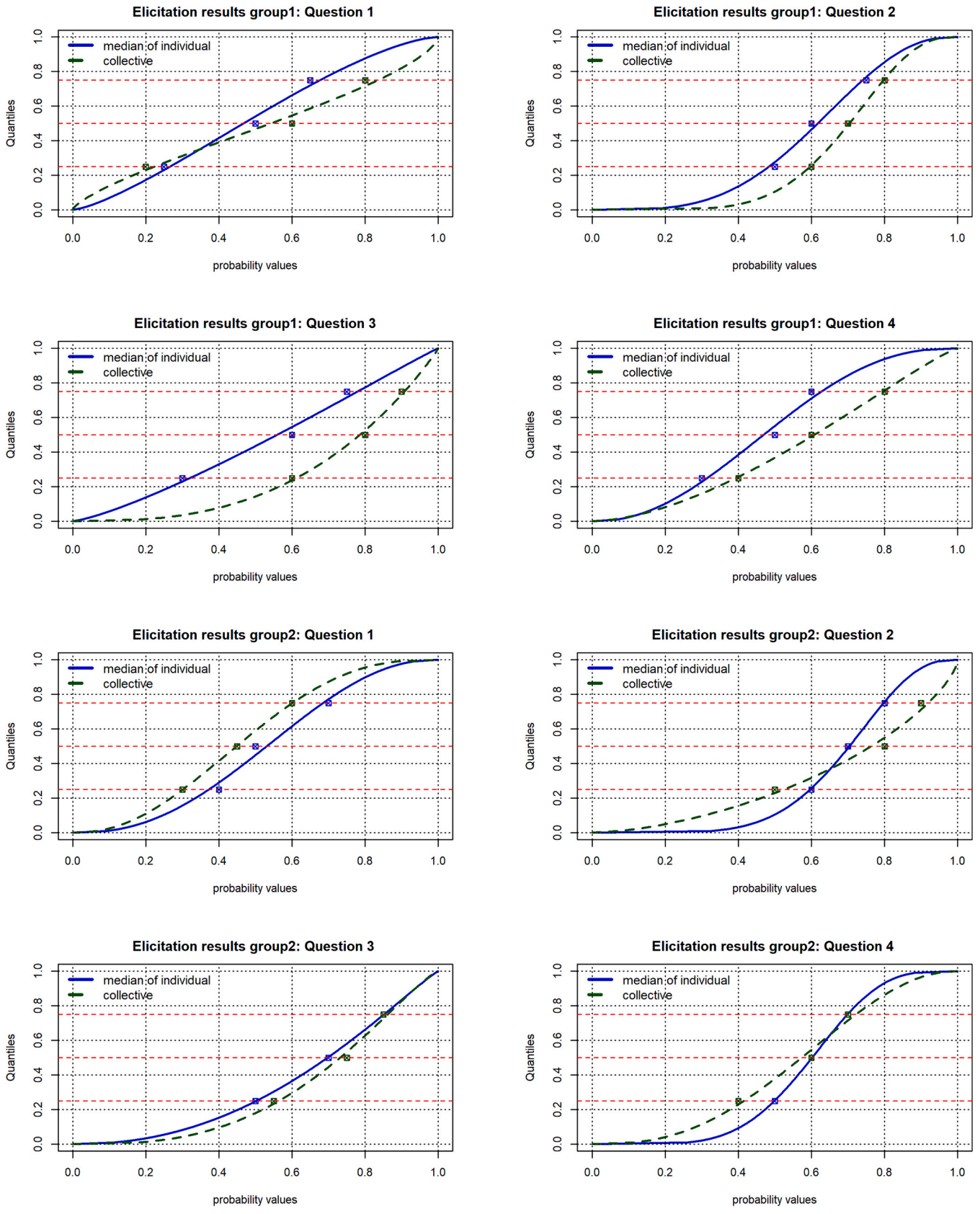

Fig. 7. Cumulative Distribution Function (CDF) for each question for group 1 and for group 2 before (median of individual elicitations) and after the collective elicitation.

Table 4
Results of the collective elicitation for groups 1 and 2.

Question	Median		Q25		Q75		P5		P95		Q25-Q75 (range)		Qualitative category	
Group	G1	G2	G1	G2	G1	G2	G1	G2	G1	G2	G1	G2	G1	G2
Question 1	0.6	0.45	0.2	0.3	0.8	0.6	0.02	0.14	0.99	0.79	0.6	0.3	NA	NA
Question 2	0.7	0.8	0.6	0.5	0.8	0.9	0.43	0.2	0.90	0.99	0.2	0.4	NA	NA
Question 3	0.8	0.75	0.6	0.55	0.9	0.85	0.34	0.32	0.98	0.96	0.3	0.3	NA	NA
Question 4	0.6	0.6	0.4	0.4	0.8	0.7	0.15	0.22	0.95	0.88	0.4	0.3	S	S

Legend: P stands for presumed, S for suspected and NC for not categorised, G1: group1, G2: group2, NA: not applicable.

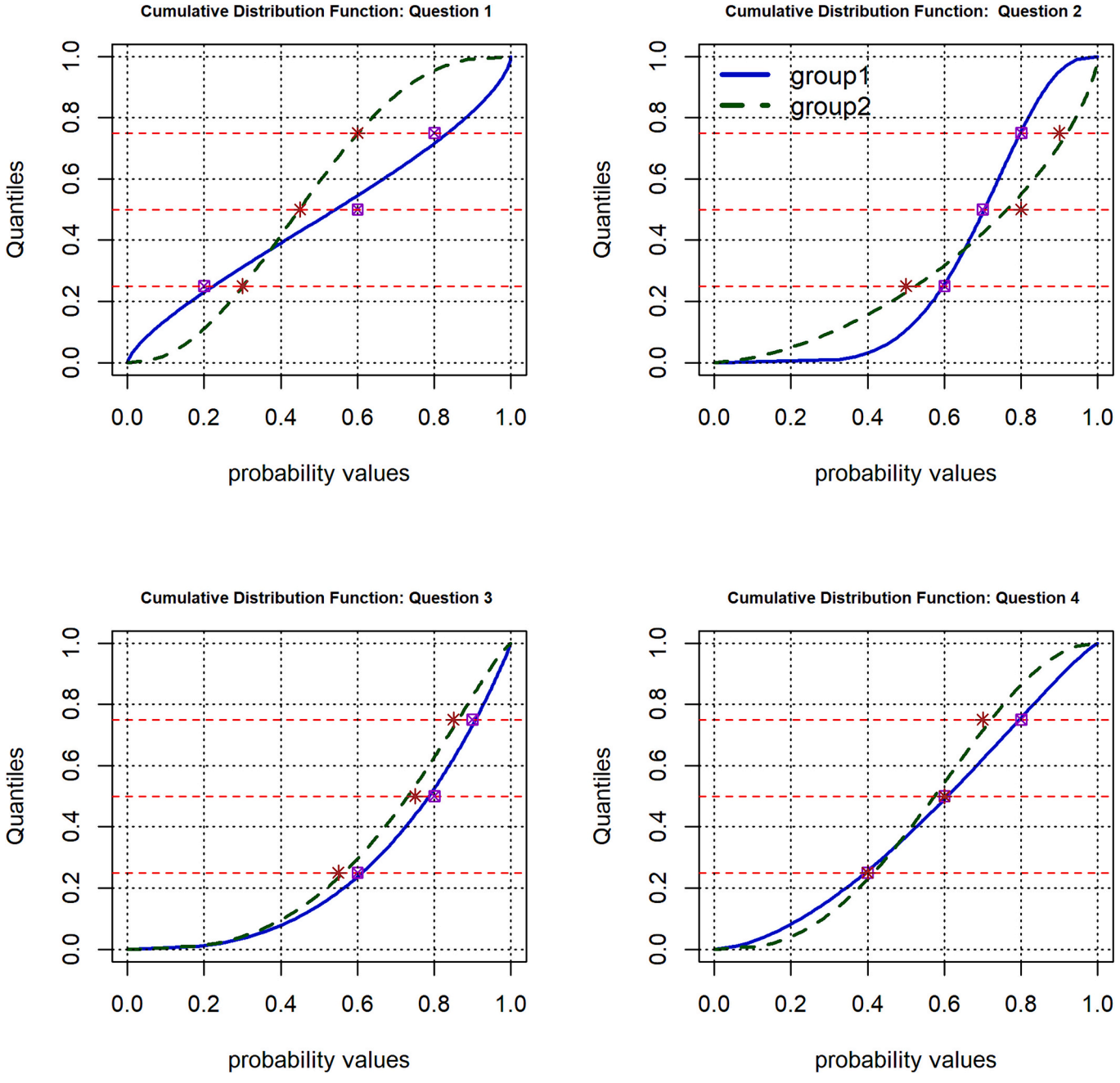

Fig. 8. Comparison of the results between groups after the collective elicitation for question 1 to question 4.
Legend: The X axis represents the cumulative probabilities describing the opinion and the Y axis the quantile of the CDF (0.5 is the median). Each symbol are values of the quartiles given by the respective group.

between 0 and 1, and the default probability distribution considered was Beta distribution. To reach a more rapid consensus, experts were asked to give a value for three quartiles. The two groups came to the same final conclusion and considered TPP to be a "suspected ED". Importantly, the level of agreement quantitatively assessed by formal elicitation, with a subjective probability distribution, and considering opinions and related

uncertainties, showed a very high level of agreement, above 86% and reaching >95% for the concluding question (question 4), which synthesised the whole opinion. Altogether, this work supports the reproducibility of the process.

The Sheffield method is an interesting tool to reach a consensus for complex questions, with heterogeneous lines of evidence weighted to obtain an answer. The collective phase was a structured opportunity for listening to each other first, and then exchanging views and argumentation, often arising out of different competencies. Such a structured approach gives predefined space for experts to listen and understand each other's opinion, and this helps in reaching a final consensus. This final consensus is analysed using numeric values, so that we are confident that the understanding of the results is harmonised between the experts. Considering the uncertainty in a quantitative manner, at the same time with the opinion itself, yields an answer about the relative weight of uncertainty in the final conclusion. All these reasons support the use of formal elicitation with quantitative estimates, in comparison with a qualitative approach.

Other adaptations of the method were made: supportive documents were carefully prepared to support the WoE approach, the time given to each expert to reply to the individual elicitation was several weeks, so that experts could consider all the information available. For the preparation of the collective stage, a mapping of the individual elicitation arguments using defined templates within online support facilitated the derivation of summaries and organisation of seemingly complex arguments. An on-line voting phase and ranking of the main arguments before the final quantitative process was also an adaptation from the initial Sheffield protocol.

The application of this method required an extensive amount of preparatory work. This included the systematic review, then the mapping of the evidence, quality assessment of the data available, allocation to respective expertise groups, selection and training, compiling information in a supportive report with summaries of publications related to each question, building and completing evaluation grids and subsequent analysis.

All these steps require several months of work. The number of experts recruited is also intended to allow deep exchange and discussion. For each expert, the assessment of 3 to 5 publications, time to reply to the individual and collective elicitations, requires a strong motivation and engagement. Fortunately, in this particular case, the total number of key publications was manageable. A higher number than that would need a process of preliminary selection of publications to be put in place, in order to retain a limited corpus of key studies. In this particular case, frequency of citation was used as a pragmatic approach to select the potential key MIE, namely PPARγ activation. That does not mean it is the sole MIE, but it is a MIE that is clearly flagged in the literature Other approaches. Based on artificial intelligence could be used to systematically explore available toxicological data that can be parsed in the scientific literature. As an example, a new tool called AOP-helpFinder was developed to identify associations between stressors and KEs supporting thus documentation of AOPs (Jornod et al., 2021). Ultimately the selection of the MoA and adverse effect is limited by the number and quality of publications, and the AOP understanding. This selection is an important part of the problem formulation as the questions will be framed to it. The collection and analysis of the data set will also be tailored by it. In a general manner, the rationale behind this choice should be well documented and balanced with other likely MIEs, end points or MoA. Again, this selection strongly depends on the problem formulation and the question to be answered. In some specific cases, the end point of interest and/or the MoA can be predetermined and then, the elicitation is organised to address them. In other situations, as it was the case for this study, the initial raised question was to illustrate the process by which a hazard assessor can categorise a chemical of interest for its endocrine-disrupting properties and in this particular work for its metabolic disruption properties. It was also a useful learning exercise to assist many of the academics within the GOLIATH project to better understand the WoE approach that regulators generally need to take when assessing chemical hazards.

As this exercise progressed, the following methodological question was also raised: Would it be possible to combine results of question 1, question 2 and question 3 to deduce the answer for the question 4 by a mathematical relationship without asking the 'elicited' experts to answer question 4? However, the answer is not obvious. Most of mathematical approaches for WoE available nowadays are Bayesian approaches (Buist et al., 2013; Vermeire et al., 2013; J.P. Gosling et al., 2013; J. P. Gosling, 2019), but these methods are developed for homogenous data aggregated in a Bayesian framework. Other methods such as Multicriteria Decision Analysis and Dempster-Shaeffer theory or utility mathematical functions require the establishment of relative weight (for Multicriteria decision analysis in relation to obesity policy options see for example (Mohebati et al., 2007)). There are many relevant elicitation/ data analysis approaches that are of utility, depending upon the framing of the question that one wants to find a robust answer to, and how to integrate different expertise and stakeholder views. However, while the synthesis may appear to be challenging, each question can be seen as parts of a puzzle, and combining parts of a puzzle can sometimes create interesting surprises.

In summary, the weight given to the level of evidence for the TPP's adverse outcome selected, namely obesity and adipogenicity, was much lower compared to the biological plausibility. The formal elicitation process shows that both groups agree on the moderate evidence based on animal data (*in vivo* studies) that TPP promotes weight increase, adipogenicity or increased body fat mass. This concurs with a preliminary LoE analysis (see the introduction section). Epidemiological data on TPP was considered weak or not relevant, which certainly contributed to a lower CDF allocated to question 1 compared to those allocated to question 2 or question 2. Overall, for question 2, mainly based on *in silico*, binding or *in vitro* studies, both groups agree regarding the interaction of TPP in the PPARγ ligand binding domain, which is thus translated to a higher CDF. Nevertheless, each group noticed some limitations and data gaps with the *in vivo* data: no study links TPP to PPARγ activation and obesogenic effects, its activation was shown at very high tested concentrations, is PPARγ activation the sole MIE? Is the activation of the MIE sufficient to activate the entire pathway? Regarding question 3, both groups considered that PPARγ activation could be implicated in the induction of obesity in humans but that PPARγ activation may not be the sole MIE. In addition, how well the data generated using rodent, *in silico,* and *in vitro* models can really be extended to explain meaningful weight gain in humans was questioned. Finally, although both groups agree that TPP may exert its effects *via* direct PPARγ activation, which is an acknowledged pathway that can lead to adipogenesis and obesity, other mechanisms were recognised as potentially also being involved. For example, in the 3 T3-L1 cell line, pre-adipocyte proliferation and subsequent adipogenic differentiation in 3 T3-L1 cells was enhanced with TPP treatment, coinciding with elevated CEBP and PPARγ pathway transcription. TPP exposure in mature adipocytes increased the basal- and insulin stimulated- uptake of the glucose analog 2-NBDG. Inhibition of PI3K, a member of the insulin signalling pathway ablated this effect (Cano-Sancho et al., 2017). Kim et al. (2021) compared a strong PPARγ therapeutic agonist that also was shown to modify PPARγ phosphorylation (*i.e.*, ROSI, a chemical that was shown to modify only PPARγ phosphorylation (*i.e.*, roscovitine), a weak PPARγ agonist and endogenous molecule (*i.e.*, 15dPGJ2), and two known environmental PPARγ ligands [*i.e.*, tetrabromobisphénol A and TPP]. Important genes were identified for predicting PPARγ ligand/ modification status, specifically the down-regulation of *Rpl13* and the upregulation of *Cidec* (Ozcagli et al., 2024 paper under review). There are also other factors that need to be considered when determining whether TPP has endocrine-disrupting properties as the role of metabolites of the parent chemical (including those generated in livestock), and timing of exposure. Finally, epidemiological evidence remains limited and not causal. Thus, the final CDF leads us to categorise TPP as a

"suspected metabolic ED". Epidemiological studies are difficult and expensive to conduct, and on the whole are seldomly available for man-made chemicals (*e.g.* regulated within EU REACH). Finally, we may suspect that the relative weight given to the first question (question 1) was more important in the final conclusion (question 4) than the answer given to question 3.

To our knowledge, this is the first time that a complete elicitation exercise has been performed in the context of ED categorisation, in a similar way with two different groups, run independently. Conducted in this way, we have been able to illustrate the reproducibility of the approach, and the exercise derived great benefit from the opportunity to have quite a large pool of experts in the field, within the GOLIATH project. Under such good conditions, we also successfully managed the whole process to avoid interactions between both groups as the study progressed, and while most of the participants were also working remotely, during the Covid-19 pandemic. The results show that both groups were similar in their answers and uncertainty for each question, which suggests that on the whole, the GOLIATH partners, with different areas of expertise, have a lot in common. Comparison with other external expert stakeholders from industry, different non-governmental organisational sectors and regulators would be a useful next step in the application of this approach. The elicitation of scientific and technical judgments from experts, in the form of subjective probability distributions, can be a valuable addition to other forms of evidence in support of public policy decision making (Morgan, 2014). Collective elicitation is used whenever direct and quantitative data are not available or in case of contradictory opinion or information and/or whenever an expert judgment is necessary (Morgan, 2014). In addition, formal elicitation is needed for quantifying the uncertainty about parameter estimate (EFSA, 2014). It allows to transparently obtain an expert's carefully considered judgment based on a systematic consideration of all relevant evidence. The formal elicitation is structured to avoid biases such as individual or collective biases (EFSA, 2014; O'Hagan et al., 2006; Tversky and Kahneman, 1974).

Based on this experience, we emphasise that a formal elicitation, as its name indicates, has to be well structured and should follow several pre-defined steps:

a) Questions need to be clearly defined at the very beginning of the procedure,
b) Experts should be well selected (without any conflict of interest, and with complementarity of expertise)
c) Sufficient training needs to be organised,
d) At the individual elicitation step, each expert opinion (both quantitative/qualitative) needs to be provided independently (with no exchange with other experts) with a clear supportive argumentation,
e) The group level opinion should be expressed at the collective elicitation step in 2 rounds:
   - o **Round 1**: debriefing individual elicitation results: each expert explains his opinion and argumentation to the group; no direct exchange except for clarifying (equality/ listening & understanding each other).
   - o **Round 2**: collective phase: each expert needs to be as objective as possible and consider the opinion/view/perspective of an impartial external group observer.

Within this elicitation process, we pursued two objectives: the first was to test the feasibility of the collective elicitation to identify EDC/MDC and the second objective was to achieve a consensus regarding the categorisation of TPP as a (suspected) metabolic disruptor. As the Sheffield method allows group interaction of experts, this was preferred, even if it was anticipated to be more time and resource consuming, as compared to the Delphi or Cook methods. We have at the end recognised the benefits of this approach, as the collective elicitation that took place during a whole day for each group, was very informative and lively discussions helped to identify for each questions the most convincing arguments that were used by the experts to support their quantitative evaluation.

Comparing to the initial plans for the whole process, we have identified only some minor deviations such as:

- Few experts who were initially included could not contribute to the whole process of elicitation, mainly due to time constraints. Nevertheless, we obtained the participation of a sufficient number of experts for each group to deliver reliable results.
- As far as reasonably possible we balanced both groups in terms of numbers, expertise, seniority of the experts, ensuring distribution of the experts within a partner organisation, across the groups, for each cluster of studies (human, *in vivo* or *in vitro*, and also on omics data). However due to the limited amount of expertise in some fields, we may have some minor discrepancies between both groups. This did not appear to have any major detrimental consequences with respect to how the studies were interpreted (results for both groups were very similar for the four questions).

It is important to recognize that the whole process was very time consuming. The first step dealing with literature search, data extraction and preparatory work (*e.g.* grid of evaluation, contribution to the supportive working document) took several months and was resources intensive for the steering group. However, on our view this step is tremendously important and allows the experts to get access to the same data set and share within a same group a common understanding of the full data set. Each expert has also spent in total around 1 week including training sessions. Moreover, for some experts the quantitative assessment methodology was challenging to understand. However, we can anticipate that if we repeat a similar elicitation process with the same experts, they will do it in a shorter time as they are already familiar and now trained with each step, and indeed this approach could be a useful training approach for early career regulators. The process also needs strong support and time commitment from the steering team.

One crucial choice before starting the whole process is the selection of the paired ED adverse effect – MoA – in our case obesity/adipogenicity and PPARγ activation. This decision is very important as the questions will be adapted to the pair and the collection and analysis of the data set will also be tailored to it. The rationale behind this choice should then be well documented and balanced across other possible end points or MoA. It could be the case at the end that the selected pair was not the one for which most evidence are available and therefore the whole process should be repeated for another pair, which will then increase the time and resources.

During the collective elicitation phase, we also realised that, even if we spent quite significant time to discuss during the previous steps (either during the training sessions or at the individual stage) the meaning of the four questions, there were still some differences in the interpretation of the questions between experts. In particular, for question 3, it was not easy for the experts to know what kind of data they had to mobilise to answer it and background/expertise of each expert has also great implication. We indicated that to answer this question they should rely either on their own expertise concerning the link between PPARγ activation and induction of obesity/ adipogenicity or on some published reviews. It was not clear for some experts if activation of PPARγ as such could be considered an endocrine MoA or if other evidence should be provided to justify this assertion.

Question 4 was also for some experts not easy to answer, as for them, the answer could simply be a direct integration of the 3 previous answers instead of being a correlated question. We agree that in a way all questions are linked and in particular as question 4 is the last one, the answer to it might be influenced by the 3 previous questions.

For both groups, the final category for TPP as a metabolic ED was "suspected". This conclusion was mainly supported by some specific data gaps such as the lack of human data, the lack of a clear demonstration of a causal link between activation of PPARγ and obesity/

adipogenicity, and the consideration that other MIEs could also trigger this effect.

The research which is now being performed within the GOLIATH project and sister metabolic disruption projects within the EURION cluster, will be delivering more relevant investigative data that will better document answers to this question. In addition, a more comprehensive review of the plausible ED modes of action leading to obesity currently underway within the EU-funded Horizon 2020 GOLIATH may in the future help us to identify additional specific or sensitive pathways. Once results of this new research are published, the conduct of a further elicitation process would be useful to charter the progress of expert views, update the evaluation and see if such results may further reduce the uncertainties, and be sufficient to modify the final categorisation for TPP.

On the basis of the elicitation study conducted here, we have demonstrated that a collective group of experts is particularly useful to aggregate and provide an opinion on a heterogeneous data set. At last, it would be of interest to investigate how the work currently done within the GOLIATH project or the EURION cluster may contribute to better document this question. Once these new findings are published, a new elicitation process could be envisaged which may ultimately help to refine or to upgrade or downgrade the current TPP's categorisation as a suspected metabolic ED. This elicitation exercise has also identified gaps in the evidence which can inform future research needs.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 825489 ("GOLIATH").

CRediT authorship contribution statement

Claire Beausoleil: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Anne Thébault:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Patrik Andersson:** Writing – review & editing, Investigation. **Nicolas J. Cabaton:** Writing – review & editing, Investigation. **Sibylle Ermler:** Writing – review & editing, Investigation. **Bernard Fromenty:** Writing – review & editing, Investigation. **Clémentine Garoche:** Writing – review & editing, Investigation. **Julian L. Griffin:** Writing – review & editing, Investigation. **Sebastian Hoffmann:** Writing – review & editing, Investigation. **Jorke H. Kamstra:** Writing – review & editing, Investigation. **Barbara Kubickova:** Writing – review & editing, Investigation. **Virissa Lenters:** Writing – review & editing, Investigation. **Vesna Munic Kos:** Writing – review & editing, Investigation. **Nathalie Poupin:** Writing – review & editing, Investigation. **Sylvie Remy:** Writing – review & editing, Investigation. **Maria Sapounidou:** Writing – review & editing, Investigation. **Daniel Zalko:** Writing – review & editing, Investigation. **Juliette Legler:** Writing – review & editing, Validation, Supervision, Investigation. **Miriam N. Jacobs:** Writing – review & editing, Validation, Supervision, Investigation. **Christophe Rousselle:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data availability

Data will be made available on request.

Acknowledgments

All partners acknowledge the contribution of their institutes for additional financial support.

We would also like to acknowledge the scientific contributions from François Pouzaud, Sakina Mhaouty-Kodja and René Habert for the development of the EKE methodology within the frame of the Anses ED Expert group. Contributions from additional members of the GOLIATH consortium: Pierre-Etienne Toulemonde and Romane Multon from Anses are also acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.taap.2024.116995.

References

Afshin, A., Forouzanfar, M.H., Reitsma, M.B., Sur, P., Estep, K., Lee, A., Marczak, L., Mokdad, A.H., Moradi-Lakeh, M., Naghavi, M., Salama, J.S., Vos, T., Abate, K.H., Abbafati, C., Ahmed, M.B., Al-Aly, Z., Alkerwi, A., Al-Raddadi, R., Amare, A.T., Amberbir, A., Amegah, A.K., Amini, E., Amrock, S.M., Anjana, R.M., Ärnlöv, J., Asayesh, H., Banerjee, A., Barac, A., Baye, E., Bennett, D.A., Beyene, A.S., Biadgilign, S., Biryukov, S., Bjertness, E., Boneya, D.J., Campos-Nonato, I., Carrero, J.J., Cecilio, P., Cercy, K., Ciobanu, L.G., Cornaby, L., Damtew, S.A., Dandona, L., Dandona, R., Dharmaratne, S.D., Duncan, B.B., Eshrati, B., Esteghamati, A., Feigin, V.L., Fernandes, J.C., Fürst, T., Gebrehiwot, T.T., Gold, A., Gona, P.N., Goto, A., Habtewold, T.D., Hadush, K.T., Hafezi-Nejad, N., Hay, S.I., Horino, M., Islami, F., Kamal, R., Kasaeian, A., Katikireddi, S.V., Kengne, A.P., Kesavachandran, C.N., Khader, Y.S., Khang, Y.H., Khubchandani, J., Kim, D., Kim, Y. J., Kinfu, Y., Kosen, S., Ku, T., Defo, B.K., Kumar, G.A., Larson, H.J., Leinsalu, M., Liang, X., Lim, S.S., Liu, P., Lopez, A.D., Lozano, R., Majeed, A., Malekzadeh, R., Malta, D.C., Mazidi, M., McAlinden, C., McGarvey, S.T., Mengistu, D.T., Mensah, G. A., Mensink, G.B.M., Mezgebe, H.B., Mirrakhimov, E.M., Mueller, U.O., Noubiap, J. J., Obermeyer, C.M., Ogbo, F.A., Owolabi, M.O., Patton, G.C., Pourmalek, F., Qorbani, M., Rafay, A., Rai, R.K., Ranabhat, C.L., Reinig, N., Safiri, S., Salomon, J.A., Sanabria, J.R., Santos, I.S., Sartorius, B., Sawhney, M., Schmidhuber, J., Schutte, A. E., Schmidt, M.I., Sepanlou, S.G., Shamsizadeh, M., Sheikhbahaei, S., Shin, M.J., Shiri, R., Shiue, I., Roba, H.S., Silva, D.A.S., Silverberg, J.I., Singh, J.A., Stranges, S., Swaminathan, S., Tabarés-Seisdedos, R., Tadese, F., Tedla, B.A., Tegegne, B.S., Terkawi, A.S., Thakur, J.S., Tonelli, M., Topor-Madry, R., Tyrovolas, S., Ukwaja, K. N., Uthman, O.A., Vaezghasemi, M., Vasankari, T., Vlassov, V.V., Vollset, S.E., Weiderpass, E., Werdecker, A., Wesana, J., Westerman, R., Yano, Y., Yonemoto, N., Yonga, G., Zaidi, Z., Zenebe, Z.M., Zipkin, B., Murray, C.J.L., 2017. Health effects of overweight and obesity in 195 countries over 25 years. N. Engl. J. Med. 377 (1), 13–27. https://doi.org/10.1056/NEJMoa1614362.

Anses, 2021. Elaboration of a Method to Categorize Substances of Interest as Regards to their Potential Endocrine Disrupting Activity: Assessment and Categorization of Prioritized Substances. Anses (Maisons-Alfort). https://www.anses.fr/fr/system/files/REACH2019SA0179Ra.pdf, 60 p.

ANSES (French Agency for Food, Environmental ond Occupationnal Health and Safety, 2018. Analysis of the most appropriate risk management option (RMOA), Substance Name: Triphenyl phosphate (TPP), EC Number: 204–112-2, CAS Number: 115–86-6.

Boyle, M., Buckley, J.P., Quirós-Alcalá, L., 2019. Associations between urinary organophosphate ester metabolites and measures of adiposity among U.S. children and adults: NHANES 2013-2014. Environ. Int. 127, 754–763. https://doi.org/10.1016/j.envint.2019.03.055.

Brock, J.M., Billeter, A., Müller-Stich, B.P., Herth, F., 2020. Obesity and the lung: what we know today. Respiration 99 (10), 856–866. https://doi.org/10.1159/000509735.

Buist, Harrie, Aldenberg, Tom, Batke, Monika, Escher, Sylvia, Entink, Rinke Klein, Kühne, Ralph, Marquart, Hans, Pauné, Eduard, Rorije, Emiel, Schüürmann, Gerrit, Kroese, Dinant, 2013. The OSIRIS weight of evidence approach: ITS mutagenicity and ITS carcinogenicity. Regul. Toxicol. Pharmacol. 67 (2), 170–181. https://doi.org/10.1016/j.yrtph.2013.01.002.

Butler, A.J., Thomas, M.K., Pintar, K.D., 2015. Systematic review of expert elicitation methods as a tool for source attribution of enteric illness. Foodborne Pathog. Dis. 12 (5), 367–382. https://doi.org/10.1089/fpd.2014.1844.

Cano-Sancho, G., Smith, A., La Merrill, M.A., 2017. Triphenyl phosphate enhances adipogenic differentiation, glucose uptake and lipolysis via endocrine and noradrenergic mechanisms. Toxicol. in Vitro 40, 280–288. https://doi.org/10.1016/j.tiv.2017.01.021.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. https://doi.org/10.1177/001316446002000104.

Colzin, S., Crépet, A., Wies, B., Rocabois, A., Sanchez, M., Perreau, S., Jean, J., Redaelli, M., Kortenkamp, A., Rousselle, C., Vrijheid, M., Nieuwenhuijsen, M., Slama, R., Angeli, K., 2024. A plausibility database summarizing the level of evidence regarding the hazards induced by the exposome on children health. Int. J. Hyg. Environ. Health 256, 114311. https://doi.org/10.1016/j.ijheh.2023.114311.

Department of Ecology, State of Washington, 2018. Children's Safe Products Reporting Rule Rationale for Reporting List of Chemicals of High Concern to Children 2011-2017. Publication 18-04-025, pp. 40–44.

EC, 2017. In: E.C. European Commission (Ed.), Guidance on the Application of the CLP Criteria Guidance to Regulation (EC) No 1272/2008 on Classification, Labelling and Packaging (CLP) of Substances and Mixtures Draft Version 5.0 April 2017.

EC, 2023. In: E.C. European Commission (Ed.), Commission Delegated Regulation (EU) …/... Of 19.12.2022 Amending Regulation (EC) no 1272/2008 as Regards Hazard Classes and Criteria for the Classification, Labelling and Packaging of Substances and Mixtures.

ECHA, EFSA, and JRC, 2018. Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009, 16 (6), 135. https://doi.org/10.2903/j.efsa.2018.5311.

EFSA, 2014. Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA J. 12 (6), 278. https://doi.org/10.2903/j.efsa.2014.3734.

EFSA, 2017. Guidance on the Use of the Weight of Evidence Approach in Scientific Assessments, p. 69. https://doi.org/10.2903/j.efsa.2017.4971.

Gosling, J.P., 2019. The importance of mathematical modelling in chemical risk assessment and the associated quantification of uncertainty. Computat. Toxicol. 10, 44–50. https://doi.org/10.1016/j.comtox.2018.12.004.

Gosling, J.P., Hart, A., Owen, H., Davies, M., Li, J., MacKay, C., 2013. A Bayes linear approach to weight-of-evidence risk assessment for skin allergy. Bayesian Anal. 8 (1), 169-186, 18.

Heindel, J.J., Blumberg, B., Cave, M., Machtinger, R., Mantovani, A., Mendez, M.A., Nadal, A., Palanza, P., Panzica, G., Sargis, R., Vandenberg, L.N., Vom Saal, F., 2017. Metabolism disrupting chemicals and metabolic disorders. Reprod. Toxicol. 68, 3–33. https://doi.org/10.1016/j.reprotox.2016.10.001.

Jakab, J., Miškić, B., Mikšić, Š., Juranić, B., Ćosić, V., Schwarz, D., Včev, A., 2021. Adipogenesis as a potential anti-obesity target: a review of pharmacological treatment and natural products. Diabet. Metab. Syndr. Obes. 14, 67–83. https://doi.org/10.2147/dmso.S281186.

Jornod, F., Jaylet, T., Blaha, L., Sarigiannis, D., Tamisier, L., Audouze, K., 2021. AOP-helpFinder webserver: a tool for comprehensive analysis of the literature to support adverse outcome pathways development. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab750.

Kim, S., Reed, E., Monti, S., Schlezinger, J.J., 2021. A data-driven transcriptional taxonomy of Adipogenic chemicals to identify white and Brite Adipogens. Environ. Health Perspect. 129 (7), 77006. https://doi.org/10.1289/ehp6886.

Klaxoon, 2024. Klaxoon, Collaborative Tool. https://klaxoon.com/fr.

Legler, J., Zalko, D., Jourdan, F., Jacobs, M., Fromenty, B., Balaguer, P., Bourguet, W., Munic Kos, V., Nadal, A., Beausoleil, C., Cristobal, S., Remy, S., Ermler, S., Margiotta-Casaluci, L., Griffin, J.L., Blumberg, B., Chesné, C., Hoffmann, S., Andersson, P.L., Kamstra, J.H., 2020. The GOLIATH project: towards an internationally harmonised approach for testing metabolism disrupting compounds. Int. J. Mol. Sci. 21 (10) https://doi.org/10.3390/ijms21103480.

Luo, D., Liu, W., Tao, Y., Wang, L., Yu, M., Hu, L., Zhou, A., Covaci, A., Xia, W., Li, Y., Xu, S., Mei, S., 2020a. Prenatal exposure to organophosphate flame retardants and the risk of low birth weight: a nested case-control study in China. Environ. Sci. Technol. 54 (6), 3375–3385. https://doi.org/10.1021/acs.est.9b06026.

Luo, K., Zhang, R., Aimuzi, R., Wang, Y., Nian, M., Zhang, J., 2020b. Exposure to organophosphate esters and metabolic syndrome in adults. Environ. Int. 143 https://doi.org/10.1016/j.envint.2020.105941.

Mohebati, L., Lobstein, T., Millstone, E., Jacobs, M., 2007. Policy options for responding to the growing challenge from obesity in the United Kingdom. Obes. Rev. 8 (Suppl. 2), 109–115. https://doi.org/10.1111/j.1467-789X.2007.00364.x.

Morgan, M.G., 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. Proc. Natl. Acad. Sci. USA 111 (20), 7176–7184. https://doi.org/10.1073/pnas.1319946111.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T., 2006. Areas for research. In: Senn, Marian Scott Stephen, Bloomfield, Peter (Eds.), Uncertain Judgements: Eliciting Experts' Probabilities. John Wiley and Sons, LTd., pp. 223–226

Pietrocatelli, S., 2008. Analyse bayesienne et élicitation d'opinions d'experts en analyse de risques et particulièrement dans le cas d'amiante chrysotile. Université de Montréal.

Tarazona, J.V., Court-Marques, D., Tiramani, M., Reich, H., Pfeil, R., Istace, F., Crivellente, F., 2017. Glyphosate toxicity and carcinogenicity: a review of the scientific basis of the European Union assessment and its differences with IARC. Arch. Toxicol. 91 (8), 2723–2743. https://doi.org/10.1007/s00204-017-1962-5.

Tontonoz, P., Spiegelman, B.M., 2008. Fat and beyond: the diverse biology of PPARgamma. Annu. Rev. Biochem. 77, 289–312. https://doi.org/10.1146/annurev.biochem.77.061307.091829.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185 (4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124.

U.S.EPA, 2020. Draft Scope of the Risk Evaluation for Triphenyl Phosphate (EPA Document# EPA-740-D-20-010), p. 94.

Vermeire, T., Aldenberg, T., Buist, H., Escher, S., Mangelsdorf, I., Pauné, E., Rorije, E., Kroese, D., 2013. OSIRIS, a quest for proof of principle for integrated testing strategies of chemicals for four human health endpoints. Regul. Toxicol. Pharmacol. 67 (2), 136–145. https://doi.org/10.1016/j.yrtph.2013.01.007.

Vose, D., 2000. Risk analysis: a quantitative guide, 3rd edition. John Wiley and sons, West Sussex, England.

WHO/IPCS, 2002. IPCS Global Assessment of the State-of-the-Science of Endocrine Disruptors. https://www.who.int/ipcs/publications/new_issues/endocrine_disruptors/en/.

Zoeller, R.T., Birnbaum, L.S., Collins, T.J., Heindel, J., Hunt, P.A., Iguchi, T., Kortenkamp, A., Myers, J.P., Vom Saal, F.S., Sonnenschein, C., Soto, A.M., 2023. European medicines agency conflicts with the European food safety authority (EFSA) on bisphenol a regulation. J. Endocr. Soc. 7 (9), bvad107. https://doi.org/10.1210/jendso/bvad107.