Regulatory Toxicology and Pharmacology 90 (2017) 51–71

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

A quantitative weight of evidence assessment of confidence in modes-of-action and their human relevance

Wolfgang Dekant [a, *], James Bridges [b], Anthony R. Scialli [c]

[a] Department of Toxicology, University of Würzburg, 97078 Würzburg, Germany
[b] University of Surrey, Guildford, UK
[c] Reproductive Toxicology Center, George Washington University, USA

ARTICLE INFO

Article history:
Received 17 May 2017
Received in revised form
1 August 2017
Accepted 19 August 2017
Available online 22 August 2017

Keywords:
Weight of evidence
Mode of action
Reproductive toxicity
Systematic analysis

ABSTRACT

A quantitative weight of evidence (QWoE) methodology was developed to assess confidence in postulated mode(s) of action for adverse effects in animal toxicity studies. The QWoE is appropriate for assessing adverse effects as relevant endpoints for classification and labeling purposes. The methodology involves definition of mode of actions and scoring supporting data for all key steps using predefined criteria for quality and relevance/strength of effects. Scores for all key steps are summarized, and the summary score is compared to the maximal achievable score for the mode of action. The ratio of the summary score to the maximal achievable scores gives an indication of confidence in a specific mode of action in animals. The mode of action in animals with highest confidence is then taken forward to assess appropriateness to humans. If one of the key steps cannot occur in humans, the mode of action is not relevant to humans. The methodology developed is applied to four case studies.

1. Introduction

The assessment of potential human health risks from exposure to chemicals requires the evaluation of many datasets providing information of widely differing nature. Toxicity testing in experimental animals remains the most common basis for human health risk characterization. Results from toxicity testing are the major basis for classification and labeling (C&L) of chemicals under the Globally Harmonized System of Classification, Labeling and Packaging (GHS) regarding specific toxicities (Dekant and Bridges, 2016b). Clear adverse effects, as defined by WHO/IPCS, in appropriately performed toxicity studies usually trigger classification for specific hazards with consequences such as restriction in use for chemicals classified as toxic to reproduction or as carcinogenic. However, the EU guidance (EC-Regulation, 2008) states that even in the presence of adverse effects, classification may not be appropriate if the adverse effect is only observed in the presence of marked differences in toxicokinetics and/or toxicodynamics between experimental animals and humans, e.g non-linear toxicokinetics (Saghir, 2015). This provision acknowledges the issues of animal to human extrapolation and the presence of a number of animal-specific modes of action without human relevance (Corton et al., 2014; Swenberg and Lehman-McKeeman, 1999). The EU guidance does not elaborate the approaches to assessing human relevance except to state that expert judgment and "weight of evidence" should be used.

Weight of evidence provides a more transparent communication of scientific judgments that should be less susceptible to bias (Lutter et al., 2015; US-OSHA, 2016). However, applied weight of evidence evaluations vary widely in their scope. Recently, quantitative weight-of-evidence (QWoE) methods have been developed to evaluate inconsistent databases on the toxicity of chemicals with the aim of generating support for decision-making in classification and labeling (Dekant and Bridges, 2016b) and assessing persistent, bioaccumulative and toxic organic pollutant properties (Bridges and Solomon, 2016). This approach relies on scoring aspects of study quality and reported effects, including weighting of effects depending on the level of biological organization that is influenced or relevance of the endpoint evaluated (Bridges and Solomon, 2016; Dekant and Bridges, 2016b; Van Der Kraak et al., 2014).

Assessment of the human relevance of an observed adverse effect in experimental animals requires information on the mode of action in animals that produces the adverse effect (Borgert et al., 2015). In the past, evaluations of the mode of action for a specific

* Corresponding author. Department of Toxicology, University of Würzburg, Versbacherstrasse 9, 97078 Würzburg, Germany.
E-mail address: dekant@toxi.uni-wuerzburg.de (W. Dekant).

http://dx.doi.org/10.1016/j.yrtph.2017.08.012
0273-2300/

chemical and its human relevance relied on narrative descriptions, which may not provide the necessary transparency. Therefore, QWoE, based on predetermined scores for how well the data support a mode of action and the absence/presence of human relevance of the effects in animals, may offer a less bias-prone and more transparent procedure.

To better structure information on mode of action, the World Health Organization (WHO/IPCS) developed a mode-of-action and human relevance framework based on an understanding of toxicity pathways (termed adverse outcome pathways) leading to disease development (Meek et al., 2013, 2014a, 2014b; OECD, 2016). The interaction of a chemical with biological macromolecules, the molecular initiating event, is a fundamental concept (Ankley et al., 2010). Within a mode of action, this molecular initiating event is followed by one or more key events that are connected by a key event relationship that describes the toxicodynamic relationships between individual key events (Becker et al., 2015; Hill, 1965; Meek et al., 2013, 2014a, 2014b; OECD, 2016; Patlewicz et al., 2013). Mode of action needs to be biologically plausible and a number of modes of action for adverse effects have found their way into textbook knowledge (Klaassen, 2013). In many cases, mode of action has been defined based on case studies where plausibility and empirical support for key effect relationships and essentiality of key effects for a disease process have been developed based on test results from a range of chemicals (OECD, 2016). Mode of action may also be based on chemical-specific information.

However, in practice, several different modes of action with potential widely differing relevance to humans may result in the observed adverse effect in experimental animals, and information to support a specific mode of action may be highly variable. Therefore, a comparative evaluation of biological plausibility and experimental support for a specific mode of action and its human relevance is required. While concepts for a systematic comparison of different possible adverse outcome pathways have been developed (Becker et al., 2017), systematic analysis of the confidence in a specific pathway (as compared to other possible pathways), applying quantitative approaches is lacking. Since basic approaches in the application of mode of action analysis have already been incorporated into guidance for regulatory approaches to chemical safety (EC-Regulation, 2008; US-EPA, 2005), a quantitative evaluation of the level of evidence is desirable and should be included in regulatory guidance for risk characterization. The approach presented here may also strengthen conclusions made in systematic reviews.

This manuscript describes a QWoE methodology to compare support for different modes of actions that may cause an adverse effect and their human relevance. The highest score supports the most reliable dataset and supports decisions on human relevance (Fig. 1). To illustrate the utility of the approach, four case studies are evaluated by these criteria. These case studies include a well-recognized male rat-specific mode of action (α_{2u}-globulin nephropathy) as case 1. The 2nd and 3rd cases involve octamethylcyclotetrasiloxane (D4) for which a narrative human relevance assessment of the animal toxicity data both on reproductive toxicity and tumorigenicity was recently published (Dekant et al., 2017). Case 4 is the developmental toxicity of diethylhexyl phthalate in male rats where a mode of action has been developed to demonstrate human relevance.

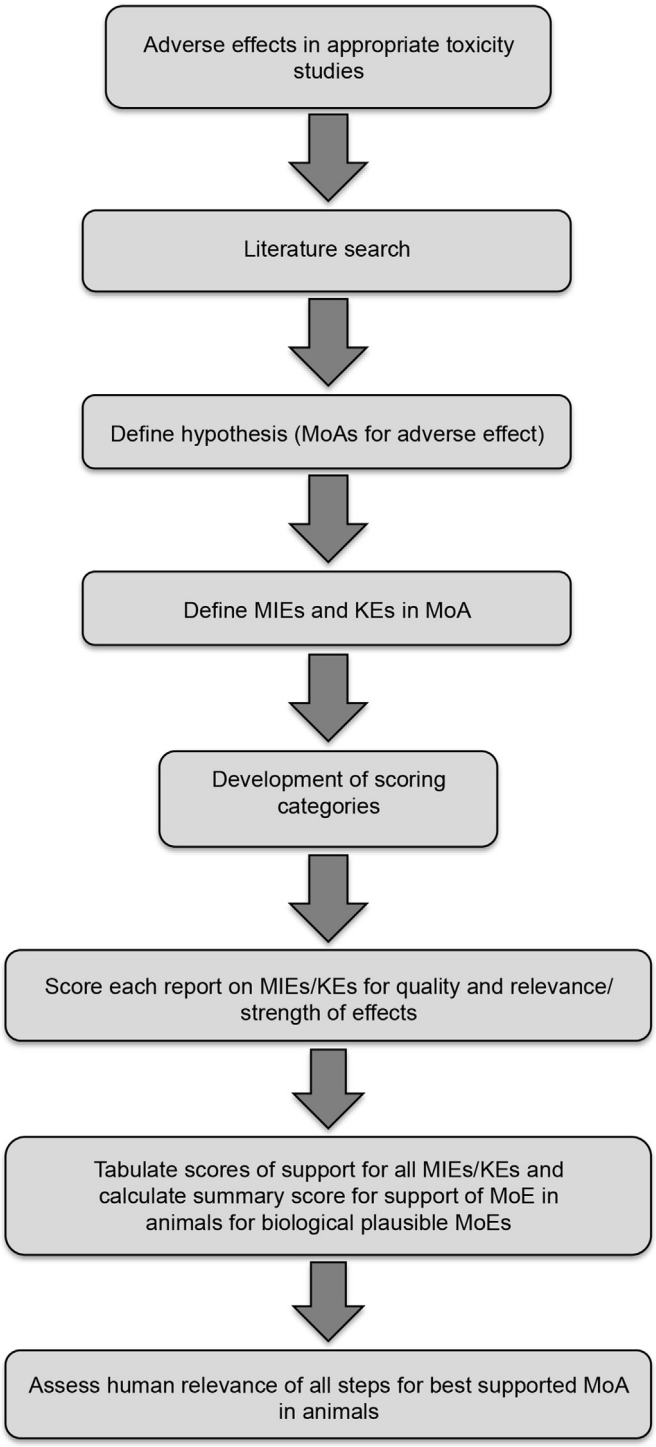

Fig. 1. Basic process of QWoE developed to support hypothesized mode of action for induction of adverse effects by a chemical. MoA = Mode of action, MIE = molecular initiating event, KE = key event. For details of study evaluation, see text.

2. Methodological approach

2.1. Development of QWoE

A weight of evidence analysis includes definition of the causal question (termed problem formulation by the US EPA), development and application of criteria for review, evaluation and integration of evidence, and conclusions based on inference (Rhomberg et al., 2013). In this context, the causal questions are i) to what extent is the mode of action for the adverse effects biologically plausible, ii) what is the confidence that a specific mode of action causes the toxic effects, and iii) how relevant to humans is the plausible mode of action considering species differences in

anatomy, physiology, toxicokinetics, and biochemistry. Responses to these questions require development of a quantitative scoring process and criteria for the application of the scores to the question in a stepwise approach (Fig. 1).

2.2. Development of mode of action

The process starts after identification of an adverse effect as defined by WHO/IPCS. The first step is a search for information applying strategies as outlined previously (Dekant and Bridges, 2016b) followed by the development of hypothesized mode of actions including identification of key events and key event relationships. The hypothesized mode of action needs to provide an explanation for the adverse effect regarding target-organ specificity, type of functional change, and induced pathology. Amenable modes of action have well defined sequential key events that are either experimentally measurable or result in measurable biomarkers, because only such molecular initiating/key events can be scored regarding data quality and relevance/strength of effects. While all possible individual steps fulfilling these criteria should be integrated in a mode of action, in practice, most modes of actions consist of four to six individual steps (Becker et al., 2015). Each component of the mode of action needs to be justified by biological plausibility and, if available, analogy of the proposed mode of action with established examples.

Several modes of actions may result in a common adverse effect. For example, tumors may be induced by a genotoxic mode of action involving DNA-damage as the molecular initiating event but may also involve a receptor interaction as the molecular initiating event or cytotoxicity-induced cell proliferation as a key event. Therefore, alternative mode of action need to be developed and comparatively evaluated to assess weight of evidence in support of a specific mode of action.

2.3. Scoring individual studies for quality

QWoE assessments must consider the study design, experimental systems, endpoints addressed, and changes reported (Bridges and Solomon, 2016; Dekant and Bridges, 2016a; Van Der Kraak et al., 2014). Scoring sheets to best cover all quality aspects of the different experimental approaches need to be developed. Quality should be based on best practice for the different endpoints. OECD/EU/US EPA toxicity testing guidelines are useful in identifying best practice for toxicity studies in experimental animals or the assessment of the potential genotoxicity. However, many of the study types required to investigate key steps in a mode of action rely on best practices derived from the scientific literature. These are, at best, only partially covered by basic principles used for guideline development.

Data to support a mode of action may be generated by experiments in intact animals, in cultured tissues or cells, or in subcellular fraction or other cell-free systems. Based on approaches developed previously (Dekant and Bridges, 2016b), quality/reliability criteria were developed for these different study types. These criteria can be considered in an assessment of support for a mode of action. The quality/reliability scoring sheets were developed to assess mechanistic studies in intact animals (Table 1), mechanistic studies *in vitro* (Table 2), and genotoxicity studies (Table 3).

The scoring criteria for studies in intact animals are based on previously presented tables developed to assess quality of experimental animal studies with some modifications (Dekant and Bridges, 2016b). Scoring criteria for *in vitro* studies are new. Besides potential issues due to inadequately defined identity and purity of the chemical of interest, the quality/reliability criteria for *in vitro* studies also address issues of study design such as the number of independent repeats of an assay, inclusion of appropriate positive and negative controls, relevance of timing of the application and sampling of the test material, extent of quality assurance, and procedures used for statistical evaluation. Scores also need to reflect the extent of characterization of stability of the test chemical in the application medium. An important issue in *in vitro* toxicity studies is dosimetry since concentrations of the chemical of interest in the medium may differ from those reaching the target system in an intact animal due to factors such as solubility, loss of material due to volatility, and absorption to the surface of glass or plastic materials used in the experiments. Studies on the genotoxicity of a chemical also require specific scoring sheets, which were developed based on the available guidelines and some of the aspects also considered in scoring the *in vitro* studies. Scores from 0 to +4 were assigned for all of the study types considered with a score of +4 representing highest quality/reliability (Dekant and Bridges, 2016a).

2.4. Scoring system of individual studies for relevance/strength of adverse effects

Three criteria are proposed for scoring of individual studies for relevance/strength of effects: i) relevance of the model system to the assessment of the molecular initiating/key event or a biomarker of the key event, ii) relevance of the exposure conditions, e.g. concentrations at which there is a change in the key event or biomarker in the model system compared to exposure conditions where an adverse outcome relevant to C&L is produced in an experimental animal, and iii) strength and consistency of effects. Scores from 0 (no relevance) to three (highly relevant) apply to these criteria. Justification for the scores is given in Table 4.

Regarding "relevance of model system," experiments using the specific species and strain of animals in the toxicity studies under exposure conditions relevant to those inducing the adverse effect of interest are considered most appropriate since both toxicokinetics and toxicodynamics are represented (+3). However, well-justified studies in other model systems including genetically modified animals also may be appropriate. Many of the molecular initiating and early key events in a mode of action are difficult to assess in intact animals. Therefore, these events are often assessed *in vitro* using different systems ranging from freshly isolated cells from the target organ in experimental animals to tissue fractions and homogenates. When scoring for relevance/strength of effects, there is a need to assess whether the system is fit for the purpose of generating adequate data to support molecular initiating and early key events. Such an assessment needs to consider relevance of the *in vitro* system to target organ anatomy and physiology.

Relevance/strength of evidence scoring also needs to assess exposure conditions that result in an observable molecular initiating and early key events (Becker et al., 2017). Concentrations applied in an experiment to assess key steps may be many orders of magnitude above or below those in the target organ in an intact animal. Such experimental conditions have no relevance in supporting a step in a mode of action and receive a score of 0. Again, exposure of experimental animals under dosing conditions that have induced the adverse effect are most relevant here (+3), but there is a need to consider toxicokinetics regarding duration of the short-term exposure with respect to absorption, time to reach peak blood/tissue levels, and conclusions regarding time points of sampling. For many chemicals, early biomarkers for an effect may be observable after short-term exposures; the only exceptions to be considered here are bioaccumulating chemicals. In this case, short-term exposures may not result in tissue concentrations sufficient to trigger molecular initiating or early key events. Such cases will require specific considerations regarding time of exposure and

Table 1
Score sheet for quality assessment of data from mechanistic studies in intact animals.

Criterion	Score of 4	Score of 3	Score of 2	Score of 1	0
1. Chemical characterization including presence contaminants that may result in confounding. Avoidance of contamination from equipment/feed/dosing solutions;	Fully characterized by performing laboratory or analysis certificate available, source specified, CAS given, impurity analysis conducted, high purity (>99.8%), identified contaminants highly unlikely to interfere with assay. Suitable procedures in place and measurements to ensure compliance, methods well evaluated	Material adequately characterized based on supplier information, major contaminants identified and quantified, purity >99% Procedures well described to address these concerns but no demonstration of compliance	Reliance on supplier for information on identity and purity, purity >98%, little information on contaminants and their possible interference Compliance for only one of these factors	Not considered, source not or poorly defined, e.g. synthesis in lab and not adequately characterized, limited information on purity and contaminants Contamination appears likely and not considered in the design.	Not described
2. General experimental design (number of animals per dose group, controls, suitability of study duration, housing conditions)	Well designed for purpose including use of adequate positive and negative controls including information on historic controls. Adequate number of animals (n > 8), well justified sampling plan with adequate study duration	Suitable for purpose but some potentially significant limitations identified. Lower number of animals/group or no positive or historic control data available. Some aspects of study duration and sampling plan are questionable	Appears to be well described, but no consideration why design selected. Low number (n only 4 to 5) animals/group, lack of positive and historic controls, sampling and study duration not well justified	Potential flaws in the methodology, such as low numbers of animals (3)/ group, inappropriate controls, inadequate sampling plan and/or study duration	Not suitable
3. Assessment of possible interference from stress due to restraint, toxicity.	Well established exposure and sampling system, experienced facility and staff regarding animal handling, toxicity endpoints regarding cytotoxicity and irritation for test chemical well described, controls sham exposed	Experienced staff and appropriate exposure conditions, possible interference by slight irritation or cytotoxicity due to exposure regimen	Exposure conditions involve stressful handling, little experience in performing such exposures in performing laboratory, exposure conditions remain in an area of slight irritation	Significant stress due to exposure conditions such as evident irritation/cytotoxicity	No consideration of interference in study design and conduct
4. Mode of application of test item to animals (stability, vehicle used, route of administration, dosing intervals, if applicable)	Checks made on levels of test item in the feed/stability in vehicle if gavage used/intake by animals assessed, human relevant route of administration, application during sensitivity window	Mode, dose, route, medium, duration appear appropriate but no measurements to determine stability/ homogeneity, intake	No measurements but mode, duration and/or route raises no specific issues	Does not appear to be appropriate, insufficient justification	Unsuitable route of administration, insufficient reporting
5. Appropriate animal species and strain selection and historic data for effects incidences in controls	Appropriate species and strain used, assignment to controls and test groups randomized. Historic incidence data available for all endpoints	Some deficiencies identified such as not Sprague Dawley rats for reproductive toxicity studies. Full historic data available.	Insufficient information to judge fully or some deficiencies identified	Source not well described	Not mentioned
6. Suitability of sampling method, sampling times and procedures.	Complies with best practice for all sampling including adequate intermediary sampling times to test the hypothesis; sampling times included at several intervals	Some doubts regarding suitability of study design for picking up relevant effects	Substantial doubts about suitability of sampling scheme, e.g. single time point only	Major issues with experimental descriptions, but information provided remains interpretable	Not described or inappropriate
7. Suitability of biochemical measurements including quality control	Complies with best practice for all measurements Study blinded to assessor and replicates run, appropriate quality control	Selection not complying fully with best practice limiting inter parameter consistency to be assessed	Limited number of endpoints, considerable difficulty in understanding how methods applied	Limited description for informed conclusions, e.g. no description of QA procedures	Unclear how the analyses were performed
8. Suitability of pathological/ functional assessment	Complies with best practice for all assessments/ measurements Study blinded to assessor and some replicates also run blinded	Selection not complying fully with best practice; e.g., replicate samples not run or blinding not reported	Limited number of endpoints, considerable difficulty in understanding how methods were validated. No replicate samples or blinding	Only one endpoint assessed, methods not well described	Not suitable or insufficient for purpose
9. Accessibility of raw data	Complete access to all raw data	Limitations in access to data to identify details of methodology used or results.	Difficult to identify important methodological details	Summary data only reported	Data provided very limited
10. Statistical analysis	Appropriate statistical method suitable for analysis of endpoint. Checked for normal distribution	Statistical methodology not optimal but acceptable Normal distribution checked	Statistical methodology not appropriate although usable for some purposes	Findings too variable to be useful except for qualitative purposes	Not amenable to interpretation

achieved tissue concentrations. In cultured cells or tissue homogenates, the chemical of interest should be present in concentrations within an order of magnitude of those that have been measured or predicted in the target tissue of the adversely affected experimental animals to achieve a high score for "relevance of concentrations used." Effects elicited by unrealistic concentrations of the chemical of interest or nonspecific toxicity in an *in vitro* system may have little relevance for confirming a molecular initiating event in an intact animal. Exceptions are for very simple assays such as receptor binding or solely hazard-related endpoints such as

Table 2
Score sheet for quality assessment of data from *in vitro* toxicity studies.

Criterion	Score of 4	Score of 3	Score of 2	Score of 1	0
1. Chemical well characterized including presence contaminants Avoidance of contamination from equipment/dosing solutions. Control of adsorption onto glassware causing interference, appropriate caution to avoid loss of volatiles, limit of solubility	Fully characterized by performing laboratory or analysis certificate available, source specified, CAS given, impurity analysis conducted All media from certified sources, well characterized, concentrations applied below limit of solubility, specific design to avoid loss of volatiles from sampling to measurements	Material adequately characterized, major contaminants identified and quantified Well characterized media, design to avoid or limit loss of volatiles	Reliance on supplier for information on identity and purity Media from established supplier, but little consideration of factors influencing concentrations of chemical of interest in media	Not considered, source not or poorly defined, e.g. synthesis in lab and not adequately characterized Concentration of chemical of interest in medium not considered, use of media from established supplier	Not described
2. General experimental design such as number of assays per dose/ concentration, controls, suitability of exposure duration,	Well-designed including use of adequate positive and negative controls and information on historic controls. Adequate number of independent repeats (n > 6), well justified sampling plan with adequate study duration, dose-response assessed in detail	Suitable for purpose but some potentially significant limitations identified. Lower number of repeats or no positive or historic control data available. Some aspects of study duration and sampling plan are questionable. Only limited dose-response	Appears to be well described, but no consideration why design selected. Low number of repeats, lack of positive and historic controls, sampling and study duration not well justified, only one concentration used	Potentially flaws in the methodology such as only three determinations of endpoint in samples from same culture, inappropriate controls, inadequate sampling plan and/or study duration	Not suitable
3. Mode of application of test item system (stability, vehicle used, route of application, dosing intervals, estimation of actual concentration of chemical of interest in medium)	Concentrations of chemical of interest in medium over time determined by analytical procedures, stability of chemical of interest well assessed and solubility well characterized, vehicle controls	Concentrations only assessed at initial time point shortly after equilibrium, limited information on stability of chemical of interest in medium	Concentrations in media only calculated based on amount added to system	Limited information on concentration and fate of chemical applied	Not suitable
4. All assessments include determination of toxicity to model organism	Detailed toxicity assessment by appropriate methods in controls and exposed system, cytotoxicity of chemical of interest in system well defined and reported	Some limitations in toxicity assessment based on methodology or limited measurements, but generally considered acceptable	Toxicity assessment limited to few measurements under assay conditions, but toxicity not evident from other parameters	Very limited information on cytotoxicity	Not suitable
5. Suitability of sampling method, sampling times and procedures.	Sampling procedures well justified based on analogy/ previous experience and time course of response assessed in positive and negative controls, deviations well justified	Limited time course, but well justified based on response pattern observed with other relevant chemicals, sampling regimen remains well justified	Only few samples collected and evaluated, limited justification for sampling plan	Only few samples collected, no justification for sampling plan	Not suitable
6. Suitability of biochemical measurements including quality control	Complies with best practice for all measurements. Study blinded to assessor and replicates run, appropriate quality control	Selection not complying fully with best practice limiting inter parameter consistency to be assessed	Limited number of endpoints, considerable difficulty in understanding how methods applied	Limited description for informed conclusions, e.g. no description of QA procedures	Unclear how the analyses were performed
7. System used for biotransformation/cells have capacity to simulate relevant reactions that occur with chemical of interest in animals	Functional biotransformation system integrated, functionality assessed with adequate positive control, well characterized cell type with enzyme activities determined in conducting laboratory, biotransformation pathways for chemical of interest characterized	Positive control requiring biotransformation to induce effects included, but only limited information on biotransformation of chemical of interest. Capacity for biotransformation assessed with model compounds under experimental conditions	Positive control regarding biotransformation-mediated effect included, but no information on biotransformation of chemical of interest, reliance on suppliers, information regarding biotransformation capacities of system	Little experience regarding biotransformation capacities of system, only taken from literature	No relevant information
8. Accessibility of raw data	Complete access to all raw data	Limitations in access to data to identify details of methodology used or results.	Difficult to identify important methodological details	Summary data only reported	Data provided very limited
9. Statistical analysis	Appropriate statistical method suitable for analysis of endpoint. Checked for normal distribution	Statistical methodology not optimal but normal distribution checked	Statistical methodology not appropriate although usable for some purposes	Findings too variable to be useful except for qualitative purposes	Not amenable to interpretation

genotoxicity, where, in the absence of cytotoxicity, limited solubility or pronounced changes in osmolality may limit maximal concentrations applied. For genotoxicity, specific guidance from OECD is available regarding concentrations to be applied, and a score of +3 can be given when these requirements are fulfilled.

The 3rd criterion, "strength of effects," assesses the magnitude of change in a molecular initiating and early key events. A score of +3 (strong support) applies when the measured changes are consistent over time and highly significant, scores of +2 (moderate support) may be selected if the effect is less pronounced and has a lower statistical significance, and scores of +1 (weak support) can be given for effects that are measurable, but have limited

Table 3
Score sheet for quality assessment of genotoxicity studies.

Criterion	Score of 4	Score of 3	Score of 2	Score of 1	Score of 0
1. Chemical well characterized including presence contaminants	Fully characterized, analysis certificate available, source specified, CAS given, impurity analysis conducted	Material adequately characterized, major contaminants identified and quantified	Reliance on supplier for information on identity and purity	Not considered, source not or poorly defined, e.g. synthesis in lab and not adequately characterized	No information
2. General experimental design such as number of replicates per concentration, and controls, suitability of study duration	Well-designed including use of adequate positive and negative controls and information on historic controls. Study design consistent with guideline including recommended concentrations of test chemical	Suitable for purpose but some limitations identified. Lower number of repeats or no historic control data available. Only limited dose-response	Appears to be well described, but no consideration why design selected. Low number of repeats, lack of historic controls, sampling and study duration not well justified,	Potentially flaws in the methodology such as inappropriate controls, inadequate sampling plan and/or study design	Not useful
4. Mode of application of test item to system and appropriate test system (stability, vehicle used, route of administration)	Recommended solvent, test system from established supplier as recommended by guideline	Some limitations such as application in less widely used solvent,	Use of rarely applied solvent, some uncertainty regarding addition of chemical of interest and resulting concentrations	Does not appear to be appropriate, insufficient justification	Unsuitable, insufficient reporting
5. Appropriate metabolic activation system when required	According to guideline and obtained from established supplier or specifically justified due to known pathways of bioactivation	Standard activation procedures applied, system generated in performing laboratory	Limitations regarding activation system	Possible issues with metabolism not considered	No information
7. Suitability of the procedures used to assess genotoxicity	Procedures consistent with respective guidelines	System used not covered by available testing guidelines, but significant experience with performance available	System not covered by guideline and limited information on performance available	Method and QA description has significant limitations	Not described
8. Accessibility of raw data	Full access	Only limited access to raw data	Raw data not accessible, but detailed description of results	Summary data only reported, questions if all generated data were reported	Data provided very limited
9. Statistical analysis	Fully appropriate	Some significant variations between observations, but appropriate statistical tests	Substantial variation between observations, limitations regarding statistical treatment of data	Findings too variable to be useful except for qualitative purposes	Not amenable to interpretation

Table 4
Scoring criteria for relevance/strength of effects.

Score	Weak, 1	Moderate, 2	Strong, 3
Concentrations applied and their relevance to dose/tissue concentrations of chemical of interest resulting in adverse effects in animals	Concentrations required to induce effect are at least two orders of magnitude above concentrations of the chemical of interest reasonably expected in tissue under exposure conditions causing the adverse effect	Concentrations required to induce effect are one order of magnitude above concentrations of the chemical of interest reasonably expected in tissue under exposure conditions causing the adverse effect	Concentrations required to induce effect are in the range of concentrations of the chemical of interest reasonably expected in tissue under exposure conditions causing the adverse effect
Relevance of model system and endpoint assessed to key events occurring in intact animals	Uncertainty regarding suitability of endpoint or biomarker to reflect critical endpoint in vivo, limitations of model system	Established model system, but some limitations regarding relevance of endpoint determined for sequence of events resulting in adverse effect in vivo	Endpoint or biomarker is clearly compatible with key event in vivo in mode of action, model system applied is highly relevant
Strength of effects	Changes in endpoint or biomarker observed, but no dose or time dependence and limited statistical significance (only $p \geq 0.05$)	Changes in endpoint or biomarker observed, but significant changes have limited dose or time-dependency	Consistent and time- and dose-related change in assessed endpoints, several measurements show significant changes ($p < 0.05$)

significance or are observed only at a single time point.

2.4.1. Deriving a score to support an individual step in a mode of action

Overall support for a molecular initiating/key event is obtained by multiplying the mean score for quality (0 to +4) with the mean overall score for relevance/strength of effects. Studies with adequate study design and detailed reporting usually give quality/reliability scores well above three, while average quality/reliability scores between two and three indicate some significant limitations. Quality/reliability scores below two indicate inadequate quality likely insufficient to come to valid conclusions based on study outcome. For transparency, the approach should integrate all studies, even when quality is well below an average of two. However, exclusion criteria for low quality studies may be applied if defined before conducting the QWoE.

The relevance/strength of effects score is obtained by multiplication of the scores obtained for relevance of exposure conditions, relevance of model system, and magnitude of effect (maximum of 27). Multiplication is used here because studies that use irrelevant exposure conditions, an irrelevant model system, or produce no response (scores of 0) do not support a molecular initiating/key event. In case several studies address a specific molecular initiating/key event with the chemical of interest, the mean of the overall study scores is used to calculate the scores for overall support of the mode of action.

2.4.2. Calculation of overall support for a mode of action in experimental animals

To obtain a summary score to assess experimental support for a

mode of action, weighting factors for molecular initiating/key events are needed. Usually, molecular initiating/key events are supportive for a specific mode of action, while late key events represent markers of a disease process that may have been initiated by upstream mechanisms (Becker et al., 2017). Therefore, late key events receive a relative weight of 0.33 in the final calculation of an overall support score for a mode of action whereas molecular initiating/early key events receive a weight of 1. The individual scores for the molecular initiating event and all weighted key events are added to give an overall score reflecting support for a mode of action. This procedure integrates supporting information and data that do not support a specific step, i.e., those that received a score of zero. Mode of action with good experimental support will receive scores close to the maximum score possible when high quality studies were performed in relevant model systems, applying relevant concentrations, and obtaining clear results. In contrast, datasets giving little support will receive a low overall score indicating that the mode of action is unlikely to account for the adverse effect induced by the chemical of interest. The overall confidence score of the dataset is compared to the maximum score achievable assuming that all studies have a quality score of +4 and all individual steps receive a score of 27 ($3 \times 3 \times 3$). A summary score for the mode of action of >75% of the maximal achievable score (quality scores of +4 for all studies and use of relevant models and concentrations with clear effects) is considered to provide very good support a specific mode of action in the test species since a quality rating of +4 for a study is not often obtained, especially for mechanism-oriented studies with many complex interferences and sometimes less pronounced changes in early endpoints. Summary scores between 50 and 75% of the maximum score achievable present moderate support indicating that data support for key events is only partially available, whereas summary score below 50 indicate only weak support for the hypothesized mode of action. Summary scores below 50 will results for datasets where several key events have limited support or where there is no support for one key event. In practice, scores for a mode of action below 30% of the maximal achievable score indicate absence of support. Low scores are mainly driven by the scores received for the late key events that are common to several modes of action and usually do not include mechanistic information. Since late key events are integrated in the calculation of summary scores, albeit with a reduced weight, summary scores of 0 for a mode of action cannot be achieved. Therefore, support scores up to the maximal score achievable for the late key events cannot be considered as support for a mode of action. The cut-offs of 30, 50, and 75% of the maximum score for the determination of support are arbitrary; an extension of the application of the QWoE-methodology to cover a larger dataset of model compounds may be needed to identify values that reliably identify the strength of support.

2.4.3. Identification of the best-supported mode of action in an animal model based on overall scores

Where there are two or more biologically plausible modes of actions that may account for an adverse effect, these are scored using identical criteria. The overall score and its relation to the maximal score achievable for a specific mode of action gives an indication of confidence and thus permits comparison of confidence in the different modes of actions. For the assessment of human relevance, the mode of action with best support in experimental animals (i.e. highest score) is selected for assessment of human relevance. Because the scoring system manipulates ranks (1, 2, 3) arithmetically as numbers and assumes equal strength of similar ranks in different domains, a comparison of scores will be most reliable when differences in scores are high. Small differences in scores may not be meaningful given the underlying assumptions about the ranks representing numerical quantities.

2.4.4. Score human relevance for best-supported mode of action in experimental animals

For scoring human relevance of the best supported mode of action in experimental animals, we propose using a simple scoring system with a score of zero for an individual step that does not occur in humans due to basic differences in anatomy, physiology, or biochemistry. A score of 1 is assigned if this step may occur in humans or if there is no relevant information to support a conclusion on human relevance. A score of zero breaks the key-event relationship. Thus, the mode of action for the resulting adverse effect induced by the chemical in experimental animals cannot be propagated to the apical endpoint, and the overall mode of action is not supported in humans (Bridges and Solomon, 2016).

3. Case studies

To demonstrate the applicability of the QWoE approach for assessment of human relevance, the methodology is applied in four case studies. Case 1, renal tumors in male rats induced by inhalation of MTBE due to accumulation of α_{2u}-globulin in the kidney, is a well described example for a mode of action that is recognized to be without human relevance (Swenberg et al., 1989; US-EPA, 1991). Cases 2 and 3 describe the application of the QWoE-methodology to octamethylcyclotetrasiloxane (D4) and to two separate effects induced by inhalation of D4, impaired female fertility and induction of benign uterine tumors observed in rats. Narrative assessments regarding the potential human relevance of these D4-induced have been published (Dekant et al., 2017). The mode of action for these two cases involving D4 are similar, and the cases are used here to demonstrate the differences in weighting of the same studies when considering different apical end-points. Case 4 represents diethylhexyl phthalate, a data-rich chemical that induces genital malformations in male offspring of rats with a specific mode of action for which human relevance is discussed (Boberg et al., 2011; Furr et al., 2014; Johnson et al., 2012; van den Driesche et al., 2015; Wilson et al., 2008).

3.1. Application of the QWoE methodology to assess human relevance of renal tumors in male rats induced by long-term inhalation of methyl tert butyl ether (MTBE)

3.1.1. Summary of the relevant toxicology of MTBE

MTBE is used as an additive in gasoline at concentrations between 2 and 15% and human exposures are likely (Stern and Kneiss, 1997; Vainiotalo et al., 1999). The toxicology of MTBE has been intensively investigated, and only studies relevant for a proof of concept regarding human relevance of a mode of action are reviewed here. Several studies have implicated the kidney as a target organ for toxicity after repeated exposure of male rats to MTBE. MTBE is consistently negative in genotoxicity testing (McGregor, 2006). In an inhalation study, Fischer 344 rats were exposed to 0, 400, 3 000, or 8000 ppm MTBE for 24 months (Bird et al., 1997). The kidney was the main target of MTBE toxicity, and the incidence of renal tubular cell tumors was increased in male rats at the intermediate dose but not at the highest dose. The lack of dose-dependency may be due to decreased survival in the high-dose group due to early death from progressive nephropathy.

Renal toxicity and renal tumor induction were not seen in female rats. MTBE was also tested for carcinogenicity in mice exposed to the same MTBE concentrations in air as rats, without effects on the kidney (Bird et al., 1997). Species differences in toxicokinetics and biotransformation of MTBE do not account for the specific effects of MTBE on the kidney of male rats (Hutcheon et al., 1996).

The developed QWoE-methodology involves a) defining mode of actions to account for the adverse effect of concern; b) scoring the quality of the experimental support for individual steps in the mode of action, c) comparing summary confidence scores for the mode of actions, d) selecting of the best supported mode of action based on the scores obtained, and e) scoring human relevance of the best supported mode of action for the adverse effect in animals. For renal tumors induced by MTBE, two possible modes of action were developed based on mechanisms of tumorigenicity of chemicals in the rodent kidney (Dekant and Vamvakas, 1992; Hard, 1998; Lock and Hard, 2004).

The first hypothetical mode of action is that the tumor induction by MTBE is related to the ability of MTBE to impair the degradation of the male rat -specific protein α_{2u}-globulin and induce a sequence of changes in renal pathology, finally ending in renal tumors. The sequence of individual key steps is shown in Table 5. This mode of action has been demonstrated with a variety of other chemicals that cause male rat-specific kidney tumors and impair the degradation of α_{2u}-globulin (Borghoff et al., 1990; Swenberg and Lehman-McKeeman, 1999; Swenberg et al., 1989). The alternative mode of action proposes tumor induction by DNA-damage and induction of mutations by MTBE. A genotoxic mode of actions has been implicated for some other chemicals that induce renal tumors in rodents (Dekant and Vamvakas, 1992; Hard, 1998; Lock and Hard, 2004).

Scoring of the role of α_{2u}-globulin-induced male rat-specific nephropathy gives a high overall score based on conclusive experimental support for the identified individual key steps based on targeted studies with high quality. The mode of action consists of a molecular initiating event, binding of MTBE to α_{2u}-globulin and a series of early and late key events (Table 5). In this case study, support for the individual key events is from high quality studies (quality score of 3.7) that were specifically targeted to the key events in the hypothesized mode of action in highly relevant system (intact rats) applying concentrations identical to those inducing the adverse effect. Therefore, scores of 3 for all aspects of relevance/strength of effects are obtained resulting in an overall relevance score of 27 ($3 \times 3 \times 3$) for all key steps. Multiplication of the relevance/strength of effects score by the quality scores then translates to a high weighted score for the individual key events. The summary score for the α_{2u}-globulin mode of action for MTBE is close to the maximal score achievable (90%, Table 5). Several other appropriately performed repeated dose studies also have described the characteristic pathology of α_{2u}-globulin nephropathy after MTBE exposures (Cruzan et al., 2007) and the available information on MTBE fulfills the US EPA criteria (Swenberg and Lehman-McKeeman, 1999) for establishing the role of α_{2u}-globulin nephropathy in male rat renal carcinogenesis. The very high score indicates strong support.

A genotoxic mode of action requires MTBE or its metabolites to interact with cellular DNA in the target tissues and cause damage to the genome. Inaccurate repair of DNA-damage causes mutations that may alter the course of cell differentiation The genotoxicity of MTBE has been intensively investigated (for overviews, see (Cruzan et al., 2007; McGregor, 2006; McGregor et al., 2005)). MTBE was not mutagenic in bacteria, did not induce mitotic gene conversion in a yeast, and did not induce chromosome damage, gene mutation or DNA-damage in mammalian cells or in intact animals (IARC-Monographs, 1999). The few positive studies (Cruzan et al., 2007) "were conducted with inappropriate methods, published only as abstracts, or not confirmed by other adequate studies".

Scoring of experimental support for a mutagenic mode of action for MTBE shows that a mutagenic mode of action has no support due to the predominantly negative genotoxicity database on MTBE and its known metabolites (11.4%, Table 6). Due to the many studies available and the time consuming scoring of quality, an average quality of 3.3 (expected quality score for a database consisting both of genotoxicity studies performed according to OECD guidelines and other laboratory studies; most of the MTBE studies were performed following testing guidelines) was assumed for the database on the genotoxicity of MTBE.

Table 5
Scoring for quality and relevance/strength for the sequence of individual steps in the pathogenesis of α_{2u}-globulin nephropathy induced by methyl tert.butyl ether (MTBE) in male rats. MIE = molecular initiating event, KE = key event.

Key step in mode of action	Data support	Quality score(s)	Relevance/strength of evidence score evidence	Weighted total score for MIE/KE
Binding of MTBE/MTBE-metabolite to α_{2u}-globulin	Binding of MTBE to α_{2u}-globulin in vivo, displacement of MTBE-derived radioactivity from renal protein by more potent ligand (Prescott-Mathews et al., 1997; Williams and Borghoff, 2001)	3.7	Model 3 Concentrations 3 Strength of effect 3	99.9
Impaired lysosomal degradation of α_{2u}-globulin with bound chemical	Supported by accumulation of α_{2u}-globulin in male rat kidney after MTBE-exposure (Prescott-Mathews et al., 1997)	3.7	Model 3 Concentration 3 Effect 3	99.9
Accumulation of α_{2u}-globulin in proximal tubular epithelial cells	Demonstrated by observation of protein droplets after short term MTBE exposure (Prescott-Mathews et al., 1997),	3.7	Model 3 Concentration 3 Effect 3	99.9
Lysosomal swelling/cytotoxicity	Demonstrated in (Prescott-Mathews et al., 1997), and other studies summarized in (Cruzan et al., 2007)	3.7	Model 3 Concentration 3 Effect 3	99.9
Cell death observed by histopathology	Observed with MTBE after repeated exposure (Prescott-Mathews et al., 1997),	3.7	Model 3 Concentration 3 Effect 3	Relative weight 0.33 33.3
Regenerative cell proliferation, specific histopathology	Observed in several studies (Cruzan et al., 2007)	3.7	Model system 3 Concentration 3 Effect 3	Relative weight 0.33 33.3
Male rat specific renal tumors	Yes, but limited dose response due to early death in male rats (Bird et al., 1997)	3.8	Model system 3 Concentration 3 Effect 2	Relative weight 0.33 22.2
	Total score for mode of action Max. score achievable Percent of maximum			488.4 540 90.4

Table 6
Scoring for quality and relevance/strength for the sequence of individual steps in the pathogenesis of renal tumors by MTBE in male rats induced by a mode of action involving DNA-damage and induction of mutations. MIE = molecular initiating event, KE = key event.

Key step in mode of action	Data support	Quality score(s)	Relevance/strength of evidence score evidence	Weighted total score for MIE/KE
DNA reactivity leading to mutation	All high-quality genotoxicity studies are negative (Cruzan et al., 2007; McGregor et al., 2005)	3.3	Model system 3 concentrations applied 3 strength of effect 0	0
Insufficient repair of mutations	No DNA-repair induced (Cruzan et al., 2007; McGregor et al., 2005)	3.3	Model system 3 Concentrations applied 3 Strength of effects 0	0
Perturbation of cell growth and survival	Yes, (Prescott-Mathews et al., 1997), also supported by other studies	3.7	Model system 3 Concentrations applied 3 Strength of effects 1 *	33 × 0.33 (late KE) = 11
Male rat specific renal tumors	Yes, but limited dose response due to early death in male rats (Bird et al., 1997)	3.8	Model system 3 concentrations applied 3 strength of effect 2**	66 × 0.33 (late KE) = 22
Total score for mode of action				33
Max. score achievable				287.3
Percent of maximum				11.4

* low score for strength of effect due to highly specific pathology selectively observed in male rats, distinct from other pathologies resulting in kidney tumors; ** lower score due to non-linear increase in tumor incidence due to high mortality.

3.1.2. *Scoring of the best-supported mode of action for MTBE regarding human relevance*

The mode of action involving binding of MTBE to α_{2u}-globulin received very high support in the scoring of individual steps and in support of the overall mode of action since it scored much higher than the alternative mechanism. This mode of action is taken forward to human relevance assessment. Table 7 summarizes the results of the human relevance assessment and an analysis of concordance between renal tumor induction and the presence of α_{2u}-globulin in species where the tumorigenicity of MTBE has been assessed. The available information shows that α_{2u}-globulin is not biosynthesized in female rats and in mice that are insensitive to renal tumor induction by MTBE (IARC-Monographs, 1999). In addition, α_{2u}-globulin is not present in humans and proteins in human kidney do not interact with other chemicals that induce α_{2u}-globulin nephropathy (Cruzan et al., 2007; McGregor, 2006). Due to these basic differences in biochemistry and physiology regarding the molecular initiating event, the chain of key events operative in male rats with MTBE is not operative and thus cannot progress to the apical endpoint in humans. Therefore, the overall mode of action identified for male rats is not relevant to humans.

3.2. *Female rat-specific reproductive toxicity of octamethyltetracyclosiloxane (D4)*

Octamethylcyclotetrasiloxane (D4) is a cyclic siloxane primarily used as a monomer or intermediate in the production of silicone polymers resulting in potential exposure of workers and potential low level inhalation or dermal exposure for the general public. D4 is an odorless liquid that is highly volatile. Human exposure occurs by

Table 7
Human relevance scoring of the α_{2u}-globulin mode of action for MTBE-induced renal tumors in male rats (0, not relevant since specific step in mode of action is not possible in humans due to species differences in biochemistry/physiology/anatomy; 1, step possible based on human biochemistry/physiology/anatomy). If an early step in a mode of action is scored as 0, further downstream steps are not possible and the mode of action is therefore not relevant in humans.

Essential step in mode of action	Data support		Possible in Humans
	Male rats	Female rats, male and female mice	
Binding of MTBE/MTBE-metabolite to α_{2u}-globulin,	Binding of MTBE to α_{2u}-globulin in systemic circulation 1	Not possible since α_{2u}-globulin not expressed, structurally similar proteins do not bind chemicals that induce α_{2u}-globulin nephropathy in male rats 0	Not possible since α_{2u}-globulin not expressed, structurally similar proteins do not bind chemicals that induce α_{2u}-globulin nephropathy in male rats 0
Impaired lysosomal degradation of α_{2u}-globulin with bound chemical,	Accumulation of α_{2u}-globulin in kidney due to impaired renal degradation, 1	0	0
Accumulation of α_{2u}-globulin in proximal tubular epithelial cells	Protein droplets after short term MTBE exposure, 1	0	0
Lysosomal swelling/ cytotoxicity	Demonstrated after repeated exposure 1	0	0
Cell death observed by histopathology	Demonstrated after repeated exposure 1	0	0
Regenerative cell proliferation, specific hiostopathology	Demonstrated after repeated exposure 1	0	0
Male rat specific renal tumors	Demonstrated after repeated exposure 1	0	0
Total	**1**	**0**	**0**

inhalation and dermal contact, although the volatility of this compound makes inhalation the most important potential route of exposure. A detailed toxicity database for D4 is available including a number of mechanistic studies performed to address mode of actions for relevant adverse effects observed in the toxicity testing (for an overview, see Dekant et al., 2017; Franzen et al., 2017).

In a reproduction toxicity study, rats were exposed by inhalation to D4 concentrations of up to the maximum achievable vapor concentration, 700 ppm. Exposure of female rats by inhalation of D4 at 700 ppm resulted in a decrease in the number of corpora lutea, the number of uterine implantation sites, and litter size (Siddiqui et al., 2007). A more detailed reproduction study in female rats exposed by inhalation to D4 was reported (Kaufman, 1998) and subsequently published (Meeks et al., 2007). The sensitive time period for the decrease in implantations and litter size was isolated to the peri-fertilization phase, from three days prior to mating to gestation day three. There was a decrease in fertility with exposure for six hours on the day prior to mating. These data point to an inhibition of ovulation or corpus luteum function as a key event in the reduction of female fertility after exposure to D4. This mechanism was confirmed (Quinn et al., 2007a) when D4 was shown to inhibit the pre-ovulatory LH surge causing a delay in ovulation, persistent follicles, and a prolonged exposure to elevated estrogen in the adult Sprague-Dawley rat. The QWoE methodology was applied to two possible modes of action scenarios to assess their experimental support and to evaluate the human relevance. The competing scenarios propose molecular initiating events based either on dopamine activity by D4 or estrogenicity of D4. The chain of key events for these competing scenarios and their scores are shown in Tables 8 and 9. The quality assessments of all underlying studies are shown in the Supplemental Material. There are other conceivable mode of actions that could be proposed to explain ovulatory disturbance. For example, alterations by D4 in noradrenergic activity or receptor binding have been investigated *in vitro* (Elias, 2009; McMullin, 2009). The evidence for an adrenergic mechanism of ovulatory disturbance is scanty, and this mechanism was therefore not considered in this QWoE analysis. Because the QWoE is data-driven, it is limited to considerations of mode of actions for which there are adequate data.

3.2.1. Dopamine activity mode of action

This mode of action involves interaction of D4 with the dopamine system causing increased dopamine activity. Increased dopaminergic activity may result in decreased prolactin and impairment in ovulation and corpus luteum function in rats (Bachelot and Binart, 2007). Inhibition/delay of ovulation and/or inadequate corpus luteum formation results in decreased mating and decreased fertility. There is inadequate evidence for a direct interaction of D4 with dopamine receptor(s) suggesting that post-receptor events are more likely (Dekant et al., 2017; Franzen et al., 2017; Jean and Plotzke, 2017). However, a mode of action based on dopamine activity is supported by studies showing dopamine-like effects of D4 in *in vitro* systems (Dekant et al., 2017; Franzen et al., 2017; Jean, 2005; Jean and Plotzke, 2017) and an observed decrease in prolactin, secretion of which is inhibited by dopamine, in *in vivo* experiments (Dekant et al., 2017; Franzen et al., 2017; Jean, 2005; Jean and Plotzke, 2017; Quinn et al., 2007a). The downstream key events (decreased prolactin and LH surge) in this mode of action have been well established for D4 using in vivo studies. However, one of the available datasets on the prolactin decrease and/or the decreased LH surge did not demonstrate an effect (Dekant et al., 2017; Elias, 2010; Jean and Plotzke, 2017) resulting in reduced scores for some of the key events due to inconsistent data.

The QWoE scoring uses the quality and strength/relevance scores of the underlying studies in the calculation of a mean score for each key event. Table 8 shows the scoring for the dopamine agonist mode of action for female reproductive toxicity of D4 based on the quality and relevance/strength of effects scoring (see Supplemental Material). For several of the key events, different datasets are available and the scores integrated into calculated overall score for the key event are means of the scores for the individual studies and their results. The last two steps received a weighting of 0.33 because they are late events and appear in both mode of action tables. This mode of action received a score of 44% of the maximum score achievable, which is considered weak support.

3.2.2. Estrogenic activity mode of action

Interaction of D4 with the estrogen receptor and downstream consequences of this interaction can be a basis for adverse effects on female rat fertility (Table 9). With regard to experimental support for this mode of action, binding of D4 to estrogen receptor-α as molecular initiating event was demonstrated in cell-free systems (Quinn et al., 2007b); binding activated the receptor and resulted in estrogenic activity (He et al., 2003; McKim et al., 2001; Quinn et al., 2007b). However, the estrogenic activity of D4 is many orders of magnitude less than the reference estrogens ethinyl estradiol and diethylstilbestrol and about two orders of magnitude less than the common food estrogen coumestrol. Therefore, low scores were given for “relevance of concentrations” in the studies supporting the molecular initiating event and the 1st key step. Estrogens at certain exposure levels trigger release of LH from the pituitary, but high or prolonged estrogen exposure is expected to suppress pituitary LH by altering gonadotropin-releasing hormone (GnRH) production from the hypothalamus (Tng, 2015).

There is no data support for an effect of D4 on gonadotropin-releasing hormone production by rat hypothalamic explants (Meeker, 2009). The last two key events are identical in both a dopamine activity and an estrogenic mode of action for D4-induced effects on female rat fertility. In addition to the very low scores for experimental support, the estrogen mode of action pathway cannot be supported based on the break in the chain of key events. Even if the broken chain is ignored, this mode of action scored only 18.7% of the possible maximum, clearly inferior to the dopaminergic activity mode of action.

3.2.3. Human relevance

The next step is an evaluation of the human relevance of the dopamine activity mode of action, which is best supported (Table 10). While binding of D4 to the dopamine receptor may be considered possible in humans the available data do not support a direct interaction of D4 with the dopamine receptor (Dekant et al., 2017; Franzen et al., 2017; Jean and Plotzke, 2017). Regardless of the molecular initiating event, an increase in dopamine activity that decreases prolactin in humans is not relevant to human ovulation or corpus luteum maintenance, because prolactin is not important in ovulation in primates. Prolactin null mice have irregular estrous cycles and do not conceive (Bachelot and Binart, 2007). When these mice ovulate, the corpus luteum does not form normally and if conception occurs, pregnancy does not continue. By contrast, prolactin is not important in primate ovulation and, indeed, excessive prolactin interferes with ovulation, even if the excess is transient and clinically unapparent (Suginami et al., 1986). Dopamine agonist medications are used to treat ovulatory disturbances attributed to prolactin excess in women (Anon, 2004). Because there are no data suggesting that D4 binds to the dopamine receptor and because dopamine agonism does not interfere with ovulation in women, the species differences in this key event break the chain. Therefore, the mode of action that best explains the adverse effects of D4 on fertility in female rats is not relevant to humans.

Table 8
Scoring for quality and relevance/strength for the sequence of individual steps for a dopamine agonism mode of action for inhibition of ovulation in female rats exposed by inhalation to D4 at 700 ppm. MIE = molecular initiating event, KE = key event.

Key steps in mode of action	Data support	Quality scores (from supplemental tables)	Relevance/ strength of evidence score	Weighted score for MIE/ KE
Increased dopamine activity	MMQ cells (a rat pituitary tumor cell line) produced less prolactin after exposure to D4 (Jean, 2005) and had decreased forskolin-stimulated cyclic AMP at D4 concentrations ≥ 25 μM without cytotoxicity (Domoradzki, 2011). Not mediated by D2 receptor (not blocked by raclopiride) or G-protein (not blocked by pertussin toxin) (Domoradzki, 2011).	(Domoradzki, 2011) 2.3	Model: 2 Concentrations: 2 Effect: 2	18.4
		(Jean, 2005) 2.2	Model: 2 Concentrations: 1 Effect: 1	4.4
			Mean	**11.4**
Decreased prolactin	Sprague-Dawley rats exposed to D4 700 or 900 ppm on diestrus 1–2 and proestrus had decreased prolactin at 1400 h on proestrus (Quinn et al., 2007a). This effect was attributable to the non-ovulatory animals. In another study, prolactin was decreased 18 h but not 8 h after inhalation of D4 700 ppm in Fischer 344 rats that had been dopamine-depleted by the administration of reserpine (Llames, 2010). However, (Elias, 2010) did not show a decrease in prolactin in 20-month-old female Fischer 344 rats exposed by nose-only inhalation to D4 700 ppm. In ovariectomized Sprague-Dawley rats, prolactin was decreased in response to estradiol implant (Stump, 2001).	(Quinn et al., 2007a) 3.6	Model: 3 Concentrations: 3 Effect: 2	64.8
		(Elias, 2010) 3.4	Model: 1 Concentrations: 3 Effect: 0	0
		(Llames, 2010) 3.4	Model: 2 Concentrations: 3 Effect: 3	61.2
		(Stump, 2001) 3.8	Model: 3 Concentrations: 3 Effect: 2	68.4
			Mean	**48.6**
Decreased LH surge	Ovariectomized Sprague-Dawley rats exposed to 700 or 900 ppm D4 for 6 h had a downward shift in distribution of LH values, although mean values were not changed from control (Stump, 2001). Sprague-Dawley rats exposed to D4 700 or 900 ppm on diestrus 1–2 and proestrus had decreased height of LH surge at 1800 h on proestrus (Quinn et al., 2007a). There was a 3–4% decrease in terminal body weight in the treated groups.	(Quinn et al., 2007a) 3.6	Model: 3 Concentrations: 3 Effect: 2	64.8
		(Stump, 2001) 3.8	Model: 3 Concentrations: 3 Effect: 2	68.4
			Mean	66.6
Inhibition/delay of ovulation and/or inadequate corpus luteum	In Sprague-Dawley rats, 700 or 900 ppm exposure on diestrus 1 and 2 and proestrus reduced the proportion of animals that ovulated (chi-squared statistically significant by us) and the number of oocytes in the oviducts on estrus (Quinn et al., 2007a). When exposure was restricted to a 6-h window, the periovulatory period was uniquely sensitive (Meeks et al., 2007). However, a decrease in oocytes in the oviducts was not seen (Quinn, 2006), although there were fewer animals. There was an increase in the number of animals with 5-day cycles (more time in diestrus) after 35 days of D4 inhalation treatment at 700 ppm. Also, an increase in resorptions seen in (Meeks et al., 2007) with exposure form 3 days before until 3 days after mating is not explained and might be spurious given lack of confirmation in longer exposures that also included this time period.	(Meeks et al., 2007) 3.5	Model: 3 Concentrations: 3 Effect: 3	94.5
		(Quinn et al., 2007a) 3.6	Model: 3 Concentrations: 3 Effect: 2	64.8
		(Quinn, 2006)	Model: 3 Concentrations: 2 Effect: 0	0
			Mean	**53.1 × 0.33 (late KE) = 17.7**
Decreased mating, fertility	Decrease in number of pups born in 2-generation study (Siddiqui et al., 2007), decrease in fertility in rats exposure for 6 h on the day prior to mating (Meeks et al., 2007).	(Meeks et al., 2007) 3.5	Model: 3 Concentrations: 3 Effect: 3	94.5
		(Siddiqui et al., 2007) 3.7	Model: 3 Concentrations: 3 Effect: 3	99.9
			Mean	**97.2 × 0.33 (late KE) = 32.1**
Total mean scores for mode of action				**176.4**
Maximum score achievable				**396**
Percent of maximum				**44%**

The exposure level at which there are adverse effects of D4 treatment in female reproduction in rats (500 and 700 ppm) could also be used to assess the relevance of this finding for human risk assessment. However, effective exposures inducing effects and their dose-response are often ignored in simplistic hazard identification schemes such as C&L. ECHA guidance (ECHA, 2015) and recent OSHA (US-OSHA, 2016) guidance for assessing carcinogenicity for C&L does, however, provide a list of factors that can be viewed as either increasing or decreasing the level of concern for human reproductive toxicity and carcinogenicity. One of these factors is “the possibility of a confounding effect of excessive toxicity at test doses”. The reproductive effects following D4 exposure were only seen at the two highest dose levels (500 and 700 ppm). It is possible that these doses may have exceeded the rat physiological capacity to handle the chemical thereby calling into question the relevance of this effect in humans at dose levels so unrealistic compared to human exposures.

3.3. Benign uterine tumors after D4 treatment

In a two-year bioassay with Fischer344 rats, inhalation of D4 at 700 ppm was associated with an increase in cystic endometrial hyperplasia and uterine adenomas. While these are benign findings, the observation of such changes might still be considered for human health risk assessments and/or C&L. An evaluation of potential mode of actions for uterine tumor formation by D4 was conducted previously but without application of a QWoE methodology (Dekant et al., 2017). The previous evaluation identified

Table 9
Scoring for quality and relevance/strength for the sequence of individual steps for an estrogenic mode of action for inhibition of ovulation in female rats exposed by inhalation to D4 at 700 ppm. MIE = molecular initiating event, KE = key event.

Key steps in mode of action	Data support	Quality scores	Relevance/ strength of effects scores	Weighted scores for MIE/KE
Binding of D4 to an estrogen receptor	Cell-free systems demonstrate displacement of 17β-estradiol from estrogen receptor-α by D4 (He et al., 2003; Quinn et al., 2007b)	(He et al., 2003) 0.9	Model: 1 Concentrations: 1 Effect: 2	1.8
		(Quinn et al., 2007b) 2.33	Model: 3 Concentrations: 1 Effect: 2	13.8
			Mean	**7.8**
Activation of downstream elements	D4 shows estrogenic activity in transactivation assay and uterotrophic assays (He et al., 2003; McKim et al., 2001; Quinn et al., 2007b; Turck, 1999). Potency is several orders of magnitude lower than ethinyl estradiol or diethylstilbestrol. Uterotrophic assay negative in one study of inadequate quality (Lee et al., 2015).	(Quinn et al., 2007b) 2.33	Model: 3 Concentrations: 1 Effect: 3	28.8
		(He et al., 2003) 2.2	Model: 3 Concentrations: 1 Effect: 1	6.6
		(McKim et al., 2001) 3.4	Model: 3 Concentrations: 1 Effect: 3	30.6
		(Lee et al., 2015) 2.1	Model: 3 Concentrations: 1 Effect:0	0
			Mean	**16.5**
Negative estrogenic feedback on hypothalamic kisspeptin or GnRH systems and/or on pituitary gonadotropins	D4 does not suppress rat hypothalamic GnRH *in vitro* (Meeker, 2009).	2.8	Model: 1 Concentrations: 1 Effect: 0	0
Inhibition/delay of ovulation	In SD rats, 700 or 900 ppm exposure on diestrus 1 and 2 and proestrus reduced the proportion of animals that ovulated and the number of oocytes in the oviducts on estrus (Quinn et al., 2007a). When exposure was restricted to a 6-h window, the periovulatory period was uniquely sensitive (Meeks et al., 2007). However, a decrease in oocytes in the oviducts was not seen, although there were fewer animals. There was an increase in the number of animals with 5-day cycles (more time in diestrus) after 35 days of D4 inhalation treatment at 700 ppm. Also, an increase in resorptions seen in (Meeks et al., 2007) with exposure form 3 days before until 3 days after mating is not explained and might be spurious given lack of confirmation in longer exposures that also included this time period.	(Meeks et al., 2007) 3.5	Model: 3 Concentrations: 3 Effect: 3	94.5
		(Quinn et al., 2007a) 3.6	Model: 3 Concentrations: 3 Effect: 2	64.8
		(Quinn, 2006)	Model: 3 Concentrations: 2 Effect: 0	0
			Mean	**53.1 × 0.33 (late KE) = 17.2**
Decreased mating, fertility	Decrease in number of pups born in 2-generation study (Siddiqui et al., 2007), decrease in fertility in rats exposure for 6 h on the day prior to mating (Meeks et al., 2007).	(Meeks et al., 2007) 3.5	Model: 3 Concentrations: 3 Effect: 3	94.5
		(Siddiqui et al., 2007) 3.7	Model: 3 Concentrations: 3 Effect: 3	99.9
			Mean	**97.2 × 0.33 (late KE) = 32.1**
Total score for mode of action				**73.9**
Max. score achievable				**396**
Percent of maximum				**18.7%**

three possible modes of action for which there were adequate data. The experimental support for these mode of actions, including dopaminergic agonism or dopamine agonist-like activity of D4, D4 estrogenicity, and D4 mutagenicity, is now subjected to the QWoE analysis. The first two of these mode of action are similar but not identical to the two mode of actions considered for female reproductive effects of D4.

The three competing possible mode of actions are presented in Tables 11–13 with the quality and relevance/strength of effects scoring of the underlying studies detailed in the Supplemental Material. Some of the scores allocated here differ from scores used for the modes of action for female reproduction due to the use of different strains of rats; the reproduction studies were performed with Sprague-Dawley rats and the two-year bioassay was performed with Fischer-344 rats. Because rat strains may differ in the sensitivity of response to D4, use of the strain in which the apical

Table 10
Human relevance scoring for a dopamine mode of action regarding female fertility for D4. 0, not relevant since specific step in mode of action is not possible in humans due to species differences in biochemistry/physiology/anatomy; 1, step possible based on human biochemistry/physiology/anatomy. If an early step in a mode of action is scored as 0, further downstream steps are not possible and the mode of action is therefore not relevant in humans.

Key steps in mode of action	Data support	Possible in humans
Increased dopamine activity	There are no data on D4 and dopamine in humans, but it is theoretically possible that a chemical exposure could increase dopamine activity, as do some pharmaceutical products.	1
Decreased prolactin	An increase in dopaminergic activity will decrease prolactin in humans	1
Decreased LH surge	A decrease in prolactin is not associated with a decrease in LH surge in humans.	0
Inhibition/delay of ovulation and/or inadequate corpus luteum	A sufficient decrease in LH surge will inhibit or delay ovulation in humans. However, control of LH is very different in rodents compared to primates and humans.	1
Decreased mating, fertility	Inhibition of ovulation will decrease fertility in humans.	1
Total		**0**

endpoint was identified (fertility in Sprague-Dawley, uterine tumors in Fischer-344) is considered to provide the most relevant information for support of a mode of action. In the estrogenicity assays, the strain of rat used would be expected to be less important, and scores are largely unchanged (OECD, 2007).

Although the percentage of the maximum possible score was comparatively low for all modes of action evaluated, the dopamine activity mode of action is the best supported of the three proposed reaching a score of 48% of the maximal score achievable (Table 11). The scores indicating only weak to moderate support are mainly driven by inconsistent datasets for the 2nd key step, decreased prolactin, and a reduced weight for an important study assigned to a late key event (Jean and Plotzke, 2017; Sloter, 2015).

An estrogen mode of action for induction of uterine tumors by D4 (Table 12) received considerably less support (22% of max. achievable score). The low score is mainly due to the very low estrogenic potency of D4 resulting in low scores for concentrations applied in the relevance/strength of effects scoring.

For a mode of action involving DNA-damage by D4 and induction of mutations to account for the induction of uterine tumors, the consistent absence of genotoxicity of D4 clearly drives the low score received (Table 13). The consistently negative genotoxicity studies with D4 result in a score of 0 for the molecular initiating even and the 1st key event. The overall score is due to late key events that are common to all three modes of actions.

Because the dopamine activity mode of action is best supported for development of uterine lesions after D4 inhalation in rats, it is taken forward to the assessment of human relevance (Table 14). When evaluating human relevance of the molecular initiating/key events, the chain of key steps is again broken at key step #3, decreased LH surge, due to the absence of an association between a decrease in prolactin and the LH surge in humans. Therefore, the dopamine activity mode of action for proliferative endometrial lesions is not relevant to humans, based again on lack of a role for prolactin in human ovulatory function. As in the discussion of the relevance of the rat reproductive effect, the exposure level at which there are adverse effects of D4 treatment in female reproduction in rats (700 ppm) could also be used to assess the relevance of this finding for human risk assessment. However, effective exposures inducing effects and their dose-response are ignored in simplistic hazard identification schemes such as C&L and we will not further discuss exposure level here.

3.4. Male developmental toxicity of diethylhexyl phthalate

In case 4, to demonstrate a quantitative WoE that supports human risk assessment, the impairment of male genital development by di-(2-ethylhexyl) phthalate (DEHP) in fetal and neonatal rats is subjected to the QWoE-procedure. Although there is evidence for other developmental alterations with DEHP treatment, and species other than rats are also sensitive, we have restricted our discussion due to the very large literature on the reproductive and developmental effects of this compound. Even within this restricted data set, we were selective in the papers that were included, bringing forward the studies that most clearly delineated the mode of action.

Euling et al. (2013) presented a qualitative WoE evaluation of the related phthalate diester dibutyl phthalate for which there are similar toxicogenomic considerations. Although the Euling et al. presentation was not quantitative, a quantitative approach could be supported by their analysis, and the proposed mode of action is similar to that proposed for DEHP in rats. Table 15 summarizes the mode of action for the anti-androgenic effects of DEHP on fetal and very young postnatal rats. A quantitative assessment of the underlying literature is summarized in the supplemental tables. Only weak support for this mode of action can be derived from the analysis, likely due to the complexity of the studies addressing the different endpoints and the large number of studies available.

Table 16 presents the human relevance scoring for this mode of action. Although the second step in the mode of action has not been evaluated in humans, it remains possible and the mode of action therefore is considered relevant to human risk assessment. However, a risk assessment must consider human exposures compared to the exposures in the experimental animal studies which are orders of magnitude above measured human exposures. Because the mode of action in rats (and other species) relies on conversion to the monoester, the kinetics of human compared to rat intestinal lipases is another factor to be considered in conducting the human risk assessment for the oral route of exposure. The fact that we have not proceeded beyond a determination that the rat data are relevant to human risk assessment does not mean that there is not considerable additional work to do in completing the risk assessment.

4. Discussion

4.1. General aspects

Criteria for C&L are hazard-based and unjustified from a scientific viewpoint (Barlow, 2016), but hazard assessment is a regulatory requirement for most uses of chemicals. Hazard assessment may be hampered by inconsistent explanations of the results of animal toxicity studies between regulators and disagreements on the relevance for humans of adverse effects in experimental animals (Golden et al., 2003; Ruden, 2001a, 2001b). Thus, a more harmonized framework based on current scientific understanding for hazard assessment and issues of extrapolation from animals to humans is needed (Schreider et al., 2010). While weight of evidence approaches are increasingly recognized, the advantage of QWoE methodology is that it provides a transparent numerical

Table 11
Scoring for quality and relevance/strength for the sequence of individual steps for a dopamine agonism mode of action for uterine tumors in female rats exposed by inhalation to D4 at 700 ppm for 24 months. MIE = molecular initiating event, KE = key event.

Key steps in mode of action	Data support	Quality scores (from supplemental tables)	Relevance/strength of evidence score	Weighted score for MIE/KE
Increased dopamine activity	MMQ cells (a rat pituitary tumor cell line) produced less prolactin after exposure to D4 (Jean, 2005) and had decreased forskolin-stimulated cyclic AMP at D4 concentrations $\geq$ 25 μM without cytotoxicity (Domoradzki, 2011). Not mediated by D2 receptor (not blocked by raclopiride) or G-protein (not blocked by pertussin toxin) (Domoradzki, 2011).	(Domoradzki, 2011) 2.3	Model: 2 Concentrations: 2 Effect: 2	18.4
		(Jean, 2005) 2.2	Model: 2 Concentrations: 1 Effect: 1	4.4
			Mean	**11.4**
Decreased prolactin	Sprague-Dawley rats exposed to D4 700 or 900 ppm on diestrus 1–2 and proestrus had decreased prolactin at 1400 h on proestrus (Quinn et al., 2007a). This effect was attributable to the non-ovulatory animals. In another study, prolactin was decreased 18 h but not 8 h after inhalation of D4 700 ppm in Fischer 344 rats that had been dopamine-depleted by the administration of reserpine (Llames, 2010). In ovariectomized Sprague-Dawley rats, prolactin was decreased in response to estradiol implant (Stump, 2001). However, (Elias, 2010) did not show a decrease in prolactin in 20-month-old female Fischer 344 rats exposed by nose-only inhalation to D4 700 ppm and (Sloter, 2015) did not show a decrease in prolactin in aged Fischer-344 rats exposed to D4.	(Quinn et al., 2007a) 3.6	Model: 3 Concentrations: 3 Effect: 2	64.8
		(Elias, 2010) 3.6	Model: 3 Concentrations: 3 Effect: 0	0
		(Llames, 2010) 3.6	Model: 2 Concentrations: 3 Effect: 1	21.6
		(Stump, 2001) 3.7	Model:2 Concentrations: 3 Effect: 2	44.4
		(Sloter, 2015) 3.8	Model: 3 Concentrations: 3 Effect: 0	0
			Mean	**32.7**
Decreased LH surge	Ovariectomized Sprague-Dawley rats exposed to 700 or 900 ppm D4 for 6 h had a downward shift in distribution of LH values, although mean values were not changed from control (Stump, 2001). Sprague-Dawley rats exposed to D4 700 or 900 ppm on diestrus 1–2 and proestrus had decreased height of LH surge at 1800 h on proestrus (Quinn et al., 2007a).	(Quinn et al., 2007a) 3.6	Model: 3 Concentrations: 3 Effect: 2	64.8
		(Stump, 2001) 3.7	Model: 2 Concentrations: 3 Effect: 2	44.4
			Mean	**54.6**
Inhibition/delay of ovulation, inadequate corpus luteum	In Sprague-Dawley rats, 700 or 900 ppm exposure on diestrus 1 and 2 and proestrus reduced the proportion of animals that ovulated (chi-squared statistically significant by us) and the number of oocytes in the oviducts on estrus (Quinn et al., 2007a). When exposure was restricted to a 6-h window, the periovulatory period was uniquely sensitive (Meeks et al., 2007).	(Meeks et al., 2007) 3.4	Model: 3 Concentrations: 3 Effect: 3	91.8
		(Quinn et al., 2007a) 3.6	Model: 3 Concentrations: 3 Effect: 2	64.8
			Mean	**78.3 × 0.33 (late KE) = 25.8**
Increase in estrogen-dominant cycle phase/Increase in circulating estradiol	There was an increase in the number of animals with 5-day cycles (more time in diestrus) after 35 days of D4 inhalation treatment at 700 ppm (Quinn, 2006). In aged Fischer-344 rats, exposure to D4 700 ppm produced an increase in number of days with estrogenic vaginal lavage, an increase in serum estradiol, and an increase in estrogen/progesterone ratio (Sloter, 2015).	(Quinn, 2006) 3.6	Model: 3 Concentrations: 3 Effect: 1	32.4
		(Sloter, 2015) 3.8	Model: 3 Concentrations: 3 Effect: 3	102.6
			Mean	**67.5 × 0.33 (late KE) = 22.3**
Increase in estrogen-dependent endometrial lesions	A 24-month inhalation study showed an increase in endometrial hyperplasia at 700 ppm and a statistically significant trend for endometrial adenoma,	(Batelle-Lee, 2004) 3.5	Model: 3 Concentrations: 3 Effect: 2	63 × 0.33 (late KE) = 20.8
Total mean scores for mode of action				**167.6**
Maximum score achievable				**432**
Percent of maximum				**38.8%**

Table 12
Scoring for quality and relevance/strength for the sequence of individual steps for an estrogenic mode of action for uterine tumors in female rats exposed by inhalation to D4 at 700 ppm for 24 months. MIE = molecular initiating event, KE = key event.

Key steps in mode of action	Data support	Quality scores (from supplemental tables)	Relevance/ strength of evidence score	Weighted score for MIE/KE
Binding of D4 to an estrogen receptor	Cell-free systems demonstrate displacement of 17β-estradiol from estrogen receptor-α by (He et al., 2003; Quinn et al., 2007b)	(He et al., 2003) 0.9	Model: 1 Concentrations: 1 Effect: 2	1.8
		(Quinn et al., 2007b) 2.33	Model: 3 Concentrations: 1 Effect: 2	13.8
			Mean	**7.8**
Activation of downstream elements	D4 shows estrogenic activity in transactivation assay and uterotrophic assays (He et al., 2003; McKim et al., 2001; Quinn et al., 2007b; Turck, 1999). Potency is several orders of magnitude lower than ethinyl estradiol or diethylstilbestrol. Uterotrophic assay negative in one study of inadequate quality (Lee et al., 2015).	(Quinn et al., 2007b) 2.33	Model: 3 Concentrations: 1 Effect: 3	28.8
		(He et al., 2003) 2.2	Model: 3 Concentrations: 1 Effect: 1	6.6
		(McKim et al., 2001) 3.4	Model: 3 Concentrations: 1 Effect: 3	30.6
		(Lee et al., 2015) 2.1	Model: 3 Concentrations: 1 Effect: 0	0
			Mean	**16.5**
Increase in estrogen-dependent endometrial lesions	A 24-month inhalation study showed an increase in endometrial hyperplasia at 700 ppm and a statistically significant trend for endometrial adenoma,	(Batelle-Lee, 2004) 3.5	Model: 3 Concentrations: 3 Effect: 2	63 × 0.33 (late KE) = 20.8
Total mean scores for mode of action				**45.1**
Maximum score achievable				**252**
Percent of maximum				**17.9%**

Table 13
Scoring for quality and relevance/strength for the sequence of individual steps for a mutagenic mode of action for uterine tumors in female rats exposed by inhalation to D4 at 700 ppm for 24 months. MIE = molecular initiating event, KE = key event.

Essential step in mode of action	Data support	Quality scores for key studies	Relevance/strength of evidence score evidence	Weighted total score for MIE/KE
DNA reactivity leading to mutation	All genotoxicity studies are consistently negative,	3.8	Model system 3 concentrations applied 3 strength of effect: 0	0
Insufficient repair of mutations	No data			0
Perturbation of cell growth and survival	No data			0
Cell proliferation and clonal expansion of neoplastic foci	Uterine hyperplasia in 24-month oncogenicity study	4	Model system 3 concentrations applied 3 Strength of effects 2	72 × 0.33 (late KE) = 23.76
Uterine tumors	Tumors at 700 ppm in 24-month oncogenicity study	4	Model system 3 concentrations applied 3 strength of effect 2	72 × 0.33 (late KE) = 23.76
Total mean scores for Mode of action				**47.52**
Maximum score achievable				**395.3**
Percent of maximum				**12%**

Table 14
Human relevance scoring for a dopamine mode of action regarding uterine tumors in rats after inhalation of D4. (0, not relevant since specific step in mode of action is not possible in humans due to species differences in biochemistry/physiology/anatomy; 1, step possible based on human biochemistry/physiology/anatomy). If an early step is scored as 0, further downstream steps are not possible and the mode of action is therefore not relevant in humans.

Key steps in mode of action	Data support	Possible in humans
Increased dopamine activity	There are no data on D4 and dopamine in humans, but it is theoretically possible that a chemical exposure could increase dopamine activity, as do some pharmaceutical products.	1
Decreased prolactin	An increase in dopaminergic activity will decrease prolactin in humans	1
Decreased LH surge	A decrease in prolactin is not associated with a decrease in LH surge in humans.	0
Inhibition/delay of ovulation, inadequate corpus luteum	A sufficient decrease in LH surge will inhibit or delay ovulation in humans.	1
Increase in estrogen-dominant cycle phase/ Increase in circulating estradiol	Inhibition or delay of ovulation in humans will increase exposure to endogenous estrogens.	1
Increase in estrogen-dependent endometrial lesions	Increase in exposure to endogenous estrogens in humans will increase estrogen-dependent endometrial lesions.	1
Total		**0**

Table 15
Scoring for quality and relevance/strength for the sequence of individual steps for the anti-androgenic mode of action for developmental toxicity in male rats exposed to DEHP. MIE = molecular initiating event, KE = key event.

Key steps in mode of action	Data support	Quality scores	Relevance/ strength of effects scores	Weighted scores for MIE/KE
Conversion to the monoester (MEHP)	Pregnant Sprague-Dawley rats convert DEHP to MEHP (Kessler et al., 2004)	(Kessler et al., 2004) 2.9	Model: 3 Concentrations: 3 Effect: N/A	26.1
Decreased activity of steroidogenic acute regulatory (StAR) protein and 5α-reductase activity	(Borch et al., 2006) reported treatment of pregnant Wistar rats with DEHP at 300 mg/kg/day to decrease mRNA and protein for StAR and other steroidogenic enzymes in fetal Leydig cells.	(Borch et al., 2006) 3.2	Model: 3 Concentrations: 3 Effect: 2	57.6
	(Svechnikov et al., 2008) showed decreased activity of StAR in Leydig cells cultured from mature and immature rats and decreased 5α-reductase activity in Leydig cells cultured from immature but not mature rats.	(Svechnikov et al., 2008) 2.2	Model: 1 Concentrations: 3 Effect: 3	19.2
	(Kariyazono et al., 2015) treated gestation-day 15 Wistar rats by gavage with DEHP and showed a decrease of StAR mRNA in fetal testes at a maternal dose level of 100 mg/kg.	(Kariyazono et al., 2015) 2.1	Model: 3 Concentrations: 3 Effect: 3	18.9
			Mean	31.9
Decreased transport of cholesterol across the mitochondrial membrane	(Svechnikov et al., 2008) showed decreased cholesterol transport associated with the decrease in StAR in rat Leydig cells	(Svechnikov et al., 2008) 2.2	Model: 1 Concentrations: 3 Effect: 3	19.2
Decreased synthesis of testosterone and dihydrotestosterone	(Akingbemi et al., 2001) showed a decrease in serum testosterone in juvenile Long-Evans rats after treatment during pregnancy of their dams with DEHP 100 mg/kg/day.	(Akingbemi et al., 2001) 2.9	Model: 3 Concentrations: 3 Effect: 3	54.8
	Borch et al., 2006 reported a decrease in fetal testis testosterone concentration and testosterone production in Wistar rats after treatment of their dams with DEHP 300 mg/gk/day.	Borch et al., 2006 3.2	Model: 3 Concentrations: 3 Effect: 2	57.6
	(Svechnikov et al., 2008) showed a decrease in hCG-stimulated testosterone production by cultured Leydig cells from adult and immature rats.	(Svechnikov et al., 2008) 2.2	Model:1 Concentrations: 3 Effect: 3	19.2
	(Culty et al., 2008) reported a decrease in fetal testis basal production of testerosterone and dihydrotestosterone after treatment of pregnant Sprague-Daley rats with 234 mg/kg/day (testosterone) or 117 mg/kg/day (DHT).	(Culty et al., 2008) 2.6	Model: 3 Concentrations: 3 Effect:: 3	23.4
			Mean	38.8
Abnormal genital development	(Moore et al., 2001) reported that treating pregnant and lactating Sprague-Dawley rats with DEHP produced male offspring with reduced anogenital distance, retained nipples and areolae, agenesis of the anterior prostate, undescended testes, incomplete preputial separation, and reduced weights of testes and testosterone-dependent sex organs beginning at 375 mg/kg/day.	(Moore et al., 2001) 3.1	Model: 3 Concentrations: 3 Effect: 3	27.9
	A multigeneration continuous breeding study by NTP (NTP, 2005) using Sprague-Dawley rats reported reduced male anogenital distance and delayed preputial separation and testis descent. Epididymides and testes were smaller. Effects were seen at a dietary level of 7500 ppm and higher, corresponding to a dose level of about 400–600 mg/kg/day.	(NTP, 2005) 3.9	Model: 3 Concentrations: 3 Effect: 3	35.1
			Mean	31.5 × 0.33 (late KE) = 10.5
Total score for mode of action				**126.5**
Max. score achievable				**467.64**
Percent of maximum				**27%**

assessment of study quality and of reported information that is transparent, consistent, and scientifically robust.

The QWoE proposed uses the criteria in the CLP Regulation (CLP Regulation 1272/2008 1272/2008 part 3.7) as a basis for conclusions regarding C&L with a specific focus on the aspect "*mode of action differences are so marked that it is certain that hazardous effects seen in the animal model will not be seen in man*" (3.7.2.3.3). Even in the presence of adverse effects of a chemical of interest in an appropriate animal study, classification is not mandated under such circumstances. However, there are widely different opinions on human relevance even for well-characterized mode of actions that have good support for the absence of human relevance (Johnson et al., 2012; Mehlman, 2000; Melnick et al., 2013; van den Driesche et al., 2015). The QWoE methodology developed here may help to reduce controversial discussions in this area due to the transparent approach. A transparent QWoE-approach may also be used to reduce disagreements on the outputs of hazard assessment and risk characterization.

4.2. Advantages of QWoE

One of the challenges in the development of the QWoE methodology is that each of the endpoints addressed and the effects reported need to be integrated. The approach described here offers several advantages. Scores for each mode of action can be compared based on well-defined criteria to define experimental

Table 16
Human relevance scoring for interference with testosterone and dihydrotestosterone synthesis as the mode of action regarding developmental toxicity of DEHP.

Key steps in mode of action	Data support	Possible in human
Conversion to the monoester (MEHP)	After ingestion of labelled DEHP or intravenous infusion of DEHP, MEHP and other metabolites were present in urine (Koch et al., 2005a,b).	1
Decreased activity of steroidogenic acute regulatory (StAR) protein and 5α-reductase activity	There are no data on this step in humans; however, it remains possible. In such cases, human relevance has to be assumed	1
Decreased transport of cholesterol across the mitochondrial membrane	The activity of StAR in humans is similar to that in rats in facilitating transport of cholesterol across the mitochondrial membrane as an early step in steroid biosynthesis.	1
Decreased synthesis of testosterone and dihydrotestosterone	A decrease in cholesterol transport across the mitochondrial membrane will decrease steroid biosynthesis in humans, and a decrease in 5α-reductase activity will decrease dihydrotestosterone.	1
Abnormal genital development	The human embryo relies on adequate exposure to testosterone and dihydrotestosterone to complete development of the male urethra and external genitalia.	1
Total		**1**

support for all steps. Additionally, quality assessment covers all aspects that need to be considered and is based on best practice thus giving a more objective assessment. Quality criteria may help design of experiments optimized to assess molecular initiating/key events, and the detailed evaluation of support for molecular initiating/key events may indicate missing links in a mode of action that need to be addressed by specific experiments. The basis for scoring is clearly defined, can be adapted to changes in scientific understanding, and is broadly applicable. Scoring for relevance/effects provides a transparent approach to integrating complex and contradictory observations into a numerical score that can be used as a basis for conclusions on experimental support for a mode of action. As shown by the outcome of case study #1, targeted experiments to support a well-defined mode of action in a relevant system will results in a high confidence in the mode of action in animals.

4.3. Challenges

Scoring for widely different experimental approaches (specifically when assessing the ever-increasing number of *in vitro* systems), requires complex knowledge of advantages and pitfalls of such systems and may need input from specialized experts familiar with limitations of the systems and issues with dosimetry. Assignment of experimental data to support a key event is clear when a target experiment on one specific molecular initiating/key events is conducted and a simple system used. However, some experiments assess outcomes under more complex circumstances and use endpoints that may be considered as early or late key events. An example is the study of Jean and Plotzke, 2017, which assessed estrogen-dominant days during the rat estrous cycle and estrogen:progesterone ratio in aged Fischer 344 rats, important evidence to support a dopaminergic mode of action in the uterine tumor evaluation (Table 11). This experiment was carefully designed to assess a key step in the proposed mode of action, and it was given a high score for quality and relevance. Because it addressed a later step in the mode of action, however, its importance was reduced by 0.33, which reduced its contribution to the evaluation. This indicates that a lower score for late events may not be always appropriate and expert judgement may be required to justify a deviation from the general approach.

In this QWoE, only positive scores were applied. Negative scores for strength of effects may need to be included to balance positive and negative studies (Fenner-Crisp and Dellarco, 2016) since larger data sets on potential key events often contain some contradictory findings. Negative scores distinguish evidence of the absence of a response from the absence of evidence (score of 0). In case of negative scores for the criterion "strength of evidence," a negative overall score (clear evidence that the endpoint is not affected) in the model for a molecular initiating/key event will be obtained and the final calculation will generate a negative number suggesting that the mode of action is unlikely. However, the QWoE here applies scores of zero for both inconclusive experiments and for databases that can be interpreted to support absence of an effect. The exclusion of negative scores reflects our lack of confidence that insufficient sensitivity or model limitations may prevent detection of an effect. In this QWoE, scores of zero for molecular initiating/key events should be interpreted as absence of support.

The criteria applied focus on the plausibility of results to support a key event and consistency with a hypothesis. This is a challenge for QWoE since the rules for the scoring process may not adequately capture the complexities of the interpretation of evidence. This will still require significant expertise and further development of such schemes based on experience will be required. A possible solution for data-rich chemicals may be to perform a QWoE only on high-quality studies with clear results and little confounding to avoid "dilution" of a score for a key event by low quality studies in systems with limited relevance.

While the use of ranking systems for the quality and strength/relevance elements is reasonable, the treatment of these ranks as equivalent across domains and the arithmetic manipulation (addition, multiplication) of the ranks may be suboptimal. These ranks highlight the difference between high and low quality datasets, and studies that received high scores appear to have high scores in multiple domains. The comparison of mode of action with small differences in overall scores might, at least in theory, produce results different from those reached through scientific judgment.

Scoring of human relevance of a specific mode of action also is challenging. The procedure selected, for reasons of simplicity, applies a simple yes/no response regarding plausibility of a key step of the mode of action in humans without directly considering data support. An expansion of the assessment integrating data support for presence/absence of a key step in humans is may be developed applying a range of scores. However, this may represent a very tedious process that requires scoring of a large number of studies including those assessing basic human physiology performed decades ago often using simple experimental designs.

Quantitative aspects such as area-under-the curve (AUC) and non-linear toxicokinetics that may be important for the expression of toxicity after long term treatment also may require specific considerations in a QWoE, specifically if adverse effects occur after compensatory mechanisms are overwhelmed or only at very high doses may also need specific approaches.

4.4. Results of the case studies

The case study for MTBE yielded a very clear outcome due to the availability of a well characterized mode of action with molecular initiating/key events that can be experimentally addressed with

high precision and that are based on very well-evaluated outcomes that have few confounders. The QWoE of the available datasets shows that the α_{2u}-globulin mode of action receives a very high confidence score based on high quality experiments with consistent outcomes in relevant systems applying relevant concentrations and addressing all plausible key events. The outcome is consistent with regulatory practice that states that when these key events are adequately supported by data, renal tumors induced by this mode of action have no human relevance and should not be used as endpoints for assessing human risk or for classification and labelling.

The D4 studies are more complicated since some endpoints may be confounded by experimental conditions and good *in vitro* models for some molecular initiating/key events are not available. The data set is further complicated by use of Sprague-Dawley rats in the reproductive studies and F344 rats in the carcinogenicity study. Dekant et al. (2017) proposed that the uterine effects seen in F344 rats following D4 exposure were due to alteration of pituitary control of the estrous cycle as a result of dopamine agonist-like activity. Although cycle disruption was demonstrated in F344 rats, it was not possible to demonstrate LH modulation as a key event because the F344 rat is highly sensitive to stress that can be induced with such a complex study. If the key event of LH modulation accounts for the reproductive effects and altered cycles in F344 rats leading to the uterine effects, it is possible that the effects in the two studies are linked to the same molecular initiating event. Lastly, there may be kinetic differences between Sprague-Dawley and F344 rats in how D4 is processed following exposure in these two strains of rats. The uterine effects were only seen following exposure to the highest exposure concentration of D4 (700 ppm) and the reproductive effects were only seen at the top two dose levels (500 and 700 ppm). At air concentrations of D4 greater than ~300 ppm (Sarangapani et al., 2002) there was an apparent saturation of liver enzymes with subsequent decreasing liver metabolism suggesting that the high doses of D4 may exceed the physiological ability of the rat to handle the chemical. A similar assessment of the kinetics in Sprague Dawley rats could be done to assess if this kinetically maximum tolerated dose is in the same range or lower. Understanding this possible influence of strain would add to the overall weight of evidence for assessing relevance of the observed effects.

The narrative assessment concluded that there is adequate support for a dopamine-like mode of action regarding benign uterine lesions induced by D4. The QWoE produced only low to moderate support, albeit better than the competitive mode of actions involving estrogenicity and genotoxicity. Low to moderate support is due to the presence of inconsistent datasets for some key events in the dopamine mode of action. The evaluation might have been clearer had the Jean and Plotzke, 2017 study been conducted to evaluate an early key event, raising the question of whether the position of the key event in the mode of action should be characterized as early or late as opposed to specific or nonspecific. Late key events are generally applicable to more than one mode of action and are, therefore, regarded as nonspecific. In the dopaminergic mode of action for uterine tumors, the Jean and Plotzke, 2017 study supports a key event that, although late in the process, is highly specific to the mode of action and absent from the competing mode of actions. Fully weighting of this key event might be appropriate based on its specificity. Dekant et al. (2017) acknowledged that it is likely that cycle disruption occurred over time in F344 females exposed to D4 due to either an inhibition by D4 of pituitary prolactin production (via dopamine agonist like activity and/or through modulation of the LH surge by another non-specified molecular initiating event) leading to an increased endogenous estrogen signal to the uterus. However, they concluded that neither mechanism would be relevant to human risk due to differences between rat and human in pituitary control of the female reproductive cycle (Klaunig et al., 2016; Plant, 2012).

The DEHP case study was included to demonstrate a mode of action that is relevant to human risk assessment, although differences in effective dose levels in rats and human exposures remain to be evaluated in the risk assessment. There is a large literature on DEHP effects on the development of androgen-responsive tissues in a number of species. Our selective citation of some of this literature and restriction to the fetal and neonatal rat was intended only to illustrate the application of the process to a data set with human relevance.

The case studies demonstrate that such a simple approach may be sufficient to come to a conclusion on human relevance. For methyl *tert*-butyl ether (MTBE), the molecular initiating event of the best-supported mode of action in the test species does not occur in humans, mice, or female rats since neither humans nor female rats nor mice express α_{2u}-globulin or similar proteins. Thus, the molecular initiating event represented by the binding of MTBE to α_{2u}-globulin cannot occur and the overall mode of action is not applicable in humans. In cases where chemicals induce liver tumors in rats by interaction with peroxisome proliferator activated receptor alpha (PPARα) or the constitutive androstane receptor (CAR), both humans and rodents express the receptor indicating that the molecular initiating event of the mode of action may occur in humans (Braeuning, 2014; Corton et al., 2014; Klaunig et al., 2003; LeBaron et al., 2014). Thus, the molecular initiating event will receive a score of one in the human relevance assessment. However, there are several datasets that show that key events downstream from the molecular initiating event in both the PPARα and the CAR mode of action do not occur in human tissues resulting in a break in the mode of action chain, supporting the conclusion that both the PPARα and the CAR mode of action for liver tumors have no relevance in humans (Braeuning, 2014; Corton et al., 2014; Klaunig et al., 2003; LeBaron et al., 2014). However, this view has been challenged (Guyton et al., 2009).

5. Conclusion

The three case studies demonstrate the utility of the developed QWoE methodology presented here to 1) assess confidence in evaluating potential mode of action (MoAs) for adverse effects observed in animal toxicity studies and 2) assess the appropriateness of the adverse effects as relevant endpoints in human health risk assessments and for classification and labeling. The method can be applied to a range of different endpoints of common concern. Moreover, it is transparent and scientifically sound and therefore less likely that it will be used as a political tool. The major challenge is to define more fully the basis for which relevance of a mode of action to humans can be discounted. QWoE-approaches should be included as part of the guidance documents for C&L and risk characterization procedures to improve the credibility of the outcome of these processes and many regulations regarding chemical safety such as EUs Equivalent Concern Regulation, the Lautenberg Chemical Safety Act, IARCs assessment of carcinogenic hazard, and risk evaluations by EFSA, ECHA and other regulatory bodies.

Acknowledgment

Preparation of this review was supported in part through an honorarium to the authors from the American Chemistry Council. This review represents the individual professional views of the authors and not necessarily the views of the American Chemistry Council.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.yrtph.2017.08.012.

Transparency document

Transparency document related to this article can be found online at http://dx.doi.org/10.1016/j.yrtph.2017.08.012.

References

Akingbemi, B.T., Youker, R.T., Sottas, C.M., Ge, R., Katz, E., Klinefelter, G.R., Zirkin, B.R., Hardy, M.P., 2001. Modulation of rat Leydig cell steroidogenic function by di(2-ethylhexyl)phthalate. Biol. Reprod. 65, 1252–1259.

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Mount, D.R., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serrrano, J.A., Tietge, J.E., Villeneuve, D.L., 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem. 29, 730–741.

Anon, 2004. Managing anovulatory infertility. Drug Ther. Bull. 42, 28–32.

Bachelot, A., Binart, N., 2007. Reproductive role of prolactin. Reproduction 133, 361–369.

Barlow, S., 2016. Risk assessment of foods and chemicals in foods. Encycl. Food Health 653–658.

Batelle-Lee, M.K., 2004. Batelle, Toxicology Northwest, Technical Report for Dow Corning Corporation - 24-Month Combined Chronic Toxicity and Oncogenicity Whole Body Vapor Inhalation Study of Octamethylcyclotetrasiloxane (D4) in Fischer 344 Rats. Report number 2004-I0000-54091, HES/DCC Study number 9106, Batelle Study number N003441A.

Becker, R.A., Ankley, G.T., Edwards, S.W., Kennedy, S.W., Linkov, I., Meek, B., Sachana, M., Segner, H., Van Der Burg, B., Villeneuve, D.L., Watanabe, H., Barton-Maclaren, T.S., 2015. Increasing scientific confidence in adverse outcome pathways: application of Tailored Bradford-Hill considerations for evaluating weight of evidence. Regul. Toxicol. Pharmacol. 72, 514–537.

Becker, R.A., Dellarco, V., Seed, J., Kronenberg, J., Meek, B., Foreman, J., Palermo, C.M., Kirman, C., Schoeny, R., Dourson, M.L., Pottenger, L.H., Manibusan, M.K., 2017. Quantitative weight of evidence to assess confidence in alternative modes of action. Regul. Toxicol. Pharmacol. 86, 205–220.

Bird, M.G., Burleigh-Flayer, H.D., Chun, J.S., Douglas, J.F., Kneiss, J.J., Andrews, L.S., 1997. Oncogenicity studies of inhaled methyl tertiary-butyl ether (MTBE) in CD-1 mice and F-344 rats. J. Appl. Toxicol. 17 (Suppl. 1), S45–S55.

Boberg, J., Christiansen, S., Axelstad, M., Kledal, T.S., Vinggaard, A.M., Dalgaard, M., Nellemann, C., Hass, U., 2011. Reproductive and behavioral effects of diisononyl phthalate (DINP) in perinatally exposed rats. Reprod. Toxicol. 31, 200–209.

Borch, J., Metzdorff, S.B., Vinggaard, A.M., Brokken, L., Dalgaard, M., 2006. Mechanisms underlying the anti-androgenic effects of diethylhexyl phthalate in fetal rat testis. Toxicology 223, 144–155.

Borgert, C.J., Wise, K., Becker, R.A., 2015. Modernizing problem formulation for risk assessment necessitates articulation of mode of action. Regul. Toxicol. Pharmacol. 72, 538–551.

Borghoff, S.J., Short, B.G., Swenberg, J.A., 1990. Biochemical mechanisms and pathobiology of alpha 2u-globulin nephropathy. Annu. Rev. Pharmacol. Toxicol. 30, 349–367.

Braeuning, A., 2014. Liver cell proliferation and tumor promotion by phenobarbital: relevance for humans? Arch. Toxicol. 88, 1771–1772.

Bridges, J., Solomon, K.R., 2016. Quantitative weight-of-evidence analysis of the persistence, bioaccumulation, toxicity, and potential for long-range transport of the cyclic volatile methyl siloxanes. J. Toxicol. Environ. Health B Crit. Rev. 19, 345–379.

Corton, J.C., Cunningham, M.L., Hummer, B.T., Lau, C., Meek, B., Peters, J.M., Popp, J.A., Rhomberg, L., Seed, J., Klaunig, J.E., 2014. Mode of action framework analysis for receptor-mediated toxicity: the peroxisome proliferator-activated receptor alpha (PPARalpha) as a case study. Crit. Rev. Toxicol. 44, 1–49.

Cruzan, G., Borghoff, S.J., de Peyster, A., Hard, G.C., McClain, M., McGregor, D.B., Thomas, M.G., 2007. Methyl tertiary-butyl ether mode of action for cancer endpoints in rodents. Regul. Toxicol. Pharmacol. 47, 156–165.

Culty, M., Thuillier, R., Li, W., Wang, Y., Martinez-Arguelles, D.B., Benjamin, C.G., Triantafilou, K.M., Zirkin, B.R., Papadopoulos, V., 2008. In utero exposure to di-(2-ethylhexyl) phthalate exerts both short-term and long-lasting suppressive effects on testosterone production in the rat. Biol. Reprod. 78, 1018–1028.

Dekant, W., Bridges, J., 2016a. Assessment of reproductive and developmental effects of DINP, DnHP and DCHP using quantitative weight of evidence. Regul. Toxicol. Pharmacol. 81, 397–406.

Dekant, W., Bridges, J., 2016b. A quantitative weight of evidence methodology for the assessment of reproductive and developmental toxicity and its application for classification and labeling of chemicals. Regul. Toxicol. Pharmacol. 82, 173–185.

Dekant, W., Scialli, A.R., Plotzke, K., Klaunig, J.E., 2017. Biological relevance of effects following chronic administration of octamethylcyclotetrasiloxane (D4) in Fischer 344 rats. Toxicol. Lett. http://dx.doi.org/10.1016/j.toxlet.2017.01.010. S0378-4274(17)30010-3.

Dekant, W., Vamvakas, S., 1992. Mechanisms of xenobiotic-induced renal carcinogenicity. Adv. Pharmacol. 23, 297–337.

Domoradzki, J.Y., 2011. Dow Corning Corporation, Health and Environmental Sciences Technical Report - Non-regulated Study: In Vitro MMQ Cell-based Evaluation of the Potential for Dopamine Receptor Activation by Octamethylcyclotetrasiloxane (D4) and Decamethylcyclopentasiloxane (D5). Silicones Environmental, Health and Safety Council Study Number 11256–102.

EC-Regulation, 2008. Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on Classification, Labelling and Packaging of Substances and Mixtures, Amending and Repealing Directives 67/548/EEC and 1999/45/EC, and Amending Regulation (EC) No 1907/2006. available online at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:353:0001:1355:en:PDF.

ECHA, 2015. European Chemicals Agency - Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.7a: Endpoint Specific Guidance. http://echa.europa.eu/web/guest/guidance-documents/guidance-on-reach.

Elias, P.D., 2009. Potential for Octamethylcyclotetrasiloxane to Modulate the Norepinephrine Synthesis Pathway In Vitro. HES Study Number 10880–102. Dow Corning Corporation.

Elias, P.D., 2010. Dow Corning Corporation, Health and Environmental Sciences Technical Report - Non-regulated Study: Effect of Octamethylcyclotetrasiloxane (D4, CAS No. 556-67-2) and Decamethylcyclopentasiloxane (D5, CAS No. 541-02-6) on Circulating Prolactin Levels in the Aged Female Fischer 344 Rat (SEHSC Contract No. 09-030). DCC HES Study Number 11360-102 (2010-STEC-3686). Report No. 2010-I0000-62617, July 2010.

Euling, S.Y., White, L.D., Kim, A.S., Sen, B., Wilson, V.S., Keshava, C., Keshava, N., Hester, S., Ovacik, M.A., Ierapetritou, M.G., Androulakis, I.P., Gaido, K.W., 2013. Use of genomic data in risk assessment case study: II. Evaluation of the dibutyl phthalate toxicogenomic data set. Toxicol. Appl. Pharmacol. 271, 349–362.

Fenner-Crisp, P.A., Dellarco, V.L., 2016. Key elements for judging the quality of a risk assessment. Environ. Health Perspect. 124, 1127–1135.

Franzen, A., Greene, T., Van landingham, C., Gentry, R., 2017. Toxicology of octamethylcyclotetrasiloxane (D4). Toxicol. Lett. http://dx.doi.org/10.1016/j.toxlet.2017.06.007. S0378-4274(17)30232-1.

Furr, J.R., Lambright, C.S., Wilson, V.S., Foster, P.M., Gray Jr., L.E., 2014. A short-term in vivo screen using fetal testosterone production, a key event in the phthalate adverse outcome pathway, to predict disruption of sexual differentiation. Toxicol. Sci. 140, 403–424.

Golden, R., Doull, J., Waddell, W., Mandel, J., 2003. Potential human cancer risks from exposure to PCBs: a tale of two evaluations. Crit. Rev. Toxicol. 33, 543–580.

Guyton, K.Z., Chiu, W.A., Bateson, T.F., Jinot, J., Scott, C.S., Brown, R.C., Caldwell, J.C., 2009. A reexamination of the PPAR-alpha activation mode of action as a basis for assessing human cancer risks of environmental contaminants. Environ. Health Perspect. 117, 1664–1672.

Hard, G.C., 1998. Mechanisms of chemically induced renal carcinogenesis in the laboratory rodent. Toxicol. Pathol. 26, 104–112.

He, B., Rhodes-Brower, S., Miller, M.R., Munson, A.E., Germolec, D.R., Walker, V.R., Korach, K.S., Meade, B.J., 2003. Octamethylcyclotetrasiloxane exhibits estrogenic activity in mice via ERalpha. Toxicol. Appl. Pharmacol. 192, 254–261.

Hill, A.B., 1965. The environment and disease: association or causation? Proc. R. Soc. Med. 58, 295–300.

Hutcheon, D.E., Arnold, J.D., ten Hove, W., Boyle 3rd, J., 1996. Disposition, metabolism, and toxicity of methyl tertiary butyl ether, an oxygenate for reformulated gasoline. J. Toxicol. Environ. Health 47, 453–464.

IARC-Monographs, 1999. Some Chemicals that Cause Tumours of the Kidney or Urinary Bladder in Rodents, and Some Other Substances. IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Man. International Agency for Research on Cancer, Lyon.

Jean, P.A., 2005. Dow Corning Corporation - Non-regulated Study: Effect of Cyclic Siloxanes on Dopamine Receptor Regulation of Serum Prolactin Levels in Female Fischer F344 Rats. Final Study Report, Silicones Environmental, Health and Safety Study No. 9939–102.

Jean, P.A., Plotzke, K., 2017. Chronic toxicity and oncogenicity of octamethylcyclotetrasiloxane (D4) in the Fischer 344 rat. Toxicol. Lett. http://dx.doi.org/10.1016/j.toxlet.2017.06.003. S0378-4274(17)30228-X.

Johnson, K.J., Heger, N.E., Boekelheide, K., 2012. Of mice and men (and rats): phthalate-induced fetal testis endocrine disruption is species-dependent. Toxicol. Sci. 129, 235–248.

Kariyazono, Y., Taura, J., Hattori, Y., Ishii, Y., Narimatsu, S., Fujimura, M., Takeda, T., Yamada, H., 2015. Effect of in utero exposure to endocrine disruptors on fetal steroidogenesis governed by the pituitary-gonad axis: a study in rats using different ways of administration. J. Toxicol. Sci. 40, 909–916.

Kaufman, L.E., 1998. Dow Corning Corporation, WIL Research Laboratories, Inc. - an Inhalation Reproductive Toxicity Study of D4 in Female Rats Using Multiple Exposure Regimens. Report No. 1998-I0000-44490, May 22, 1998.

Kessler, W., Numtip, W., Grote, K., Csanady, G.A., Chahoud, I., Filser, J.G., 2004. Blood burden of di(2-ethylhexyl) phthalate and its primary metabolite mono(2-ethylhexyl) phthalate in pregnant and nonpregnant rats and marmosets. Toxicol. Appl. Pharmacol. 195, 142–153.

Klaassen, C.D. (Ed.), 2013. Casarett and Doull's Toxicology. The Basic Science of Poisons. McGraw Hill, New York.

Klaunig, J.E., Babich, M.A., Baetcke, K.P., Cook, J.C., Corton, J.C., David, R.M., DeLuca, J.G., Lai, D.Y., McKee, R.H., Peters, J.M., Roberts, R.A., Fenner-Crisp, P.A., 2003. PPARalpha agonist-induced rodent tumors: modes of action and human

relevance. Crit. Rev. Toxicol. 33, 655–780.
Klaunig, J.E., Dekant, W., Plotzke, K., Scialli, A.R., 2016. Biological relevance of decamethylcyclopentasiloxane (D5) induced rat uterine endometrial adenocarcinoma tumorigenesis: mode of action and relevance to humans. Regul. Toxicol. Pharmacol. 74 (Suppl. 1), S44–S56.
Koch, H.M., Bolt, H.M., Preuss, R., Angerer, J., 2005a. New metabolites of di(2-ethylhexyl)phthalate (DEHP) in human urine and serum after single oral doses of deuterium-labelled DEHP. Arch. Toxicol. 79, 367–376.
Koch, H.M., Bolt, H.M., Preuss, R., Eckstein, R., Weisbach, V., Angerer, J., 2005b. Intravenous exposure to di(2-ethylhexyl)phthalate (DEHP): metabolites of DEHP in urine after a voluntary platelet donation. Arch. Toxicol. 79, 689–693.
LeBaron, M.J., Rasoulpour, R.J., Gollapudi, B.B., Sura, R., Kan, H.L., Schisler, M.R., Pottenger, L.H., Papineni, S., Eisenbrandt, D.L., 2014. Characterization of nuclear receptor-mediated murine hepatocarcinogenesis of the herbicide pronamide and its human relevance. Toxicol. Sci. 142, 74–92.
Lee, D., Ahn, C., An, B.S., Jeung, E.B., 2015. Induction of the estrogenic marker calbindn-D(9)k by octamethylcyclotetrasiloxane. Int. J. Environ. Res. Public Health 12, 14610–14625.
Llames, L.T., 2010. Dow Corning Corporation, Health and Environmental Sciences Technical Report - Non-regulated Study: In Vivo Evaluation of the Impact of Exposure/endpoint Evaluation Timing on the Potential for Octamethylcyclotetrasiloxane and Decamethylcyclopentasiloxane to Affect Circulating Prolactin Levels in the Reserpine-treated Fischer 344 Rat. DCC Study no. 11257-102, Silicones Environmental Health and Safety Council, HES Study no. 11360–102.
Lock, E.A., Hard, G.C., 2004. Chemically induced renal tubule tumors in the laboratory rat and mouse: review of the NCI/NTP database and categorization of renal carcinogens based on mechanistic information. Crit. Rev. Toxicol. 34, 211–299.
Lutter, R., Abbott, L., Becker, R., Borgert, C., Bradley, A., Charnley, G., Dudley, S., Felsot, A., Golden, N., Gray, G., Juberg, D., Mitchell, M., Rachman, N., Rhomberg, L., Solomon, K., Sundlof, S., Willett, K., 2015. Improving weight of evidence approaches to chemical evaluations. Risk Anal. 35, 186–192.
McGregor, D., 2006. Methyl tertiary-butyl ether: studies for potential human health hazards. Crit. Rev. Toxicol. 36, 319–358.
McGregor, D.B., Cruzan, G., Callander, R.D., May, K., Banton, M., 2005. The mutagenicity testing of tertiary-butyl alcohol, tertiary-butyl acetate and methyl tertiary-butyl ether in Salmonella typhimurium. Mutat. Res. 565, 181–189.
McKim Jr., J.M., Wilga, P.C., Breslin, W.J., Plotzke, K.P., Gallavan, R.H., Meeks, R.G., 2001. Potential estrogenic and antiestrogenic activity of the cyclic siloxane octamethylcyclotetrasiloxane (D4) and the linear siloxane hexamethyldisiloxane (HMDS) in immature rats using the uterotrophic assay. Toxicol. Sci. 63, 37–46.
McMullin, T., 2009. Potential for Octamethylcyclotetrasoloxane to Bind Alpha- and Beta-adrenergic Receptors In Vitro. Dow CCorning Corporation. Study No. 10879–102.
Meek, M.E., Bolger, M., Bus, J.S., Christopher, J., Conolly, R.B., Lewis, R.J., Paolini, G.M., Schoeny, R., Haber, L.T., Rosenstein, A.B., Dourson, M.L., 2013. A framework for fit-for-purpose dose response assessment. Regul. Toxicol. Pharmacol. 66, 234–240.
Meek, M.E., Boobis, A., Cote, I., Dellarco, V., Fotakis, G., Munn, S., Seed, J., Vickers, C., 2014a. New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. J. Appl. Toxicol. 34, 1–18.
Meek, M.E., Palermo, C.M., Bachman, A.N., North, C.M., Jeffrey Lewis, R., 2014b. Mode of action human relevance (species concordance) framework: evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. J. Appl. Toxicol. 34, 595–606.
Meeker, L.S., 2009. Dow Corning Corporation - Non-regulated Study: Potential for Octamethylcyclotetrasiloxane to Bind Alpha- and Beta-adrenergic Receptors in Vitro Study Number: 10879–102. Sponsor:. Silicones Environmental, Health and Safety Council.
Meeks, R.G., Stump, D.G., Siddiqui, W.H., Holson, J.F., Plotzke, K.P., Reynolds, V.L., 2007. An inhalation reproductive toxicity study of octamethylcyclotetrasiloxane (D4) in female rats using multiple and single day exposure regimens. Reprod. Toxicol. 23, 192–201.
Mehlman, M.A., 2000. Misclassification of carcinogenic methyl tertiary butyl ether (MTBE) by the National Toxicology Program Board: smokescreen in, science out? Arch. Environ. Health 55, 73–74.
Melnick, R.L., Ward, J.M., Huff, J., 2013. War on Carcinogens: industry disputes human relevance of chemicals causing cancer in laboratory animals based on unproven hypotheses, using kidney tumors as an example. Int. J. Occup. Environ. Health 19, 255–260.
Moore, R.W., Rudy, T.A., Lin, T.M., Ko, K., Peterson, R.E., 2001. Abnormalities of sexual development in male rats with in utero and lactational exposure to the antiandrogenic plasticizer Di(2-ethylhexyl) phthalate. Environ. Health Perspect. 109, 229–237.
NTP, 2005. National Toxicology Program. Diethylhexylphthalate: Multigenerational Reproductive Assessment by Continuous Breeding when Administered to Sprague–Dawley Rats in the Diet. NTP Technical Report. National Toxicology Program, Research Triangle Park, NC. https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB2005107575.xhtml. Tech Rep Ser 2012.
OECD, 2007. Test Guideline for the Testing of Chemicals: Uterotrophic Bioassay in Rodents: A Short-term Screening Test for Oestrogenic Properties. OECD/OCDE Guideline No. 440, adopted 16 October 2007, available online at: https://ntp.niehs.nih.gov/iccvam/suppdocs/feddocs/oecd/oecdtg440.pdf.
OECD, 2016. Users' Handbook Supplement to the Guidance Document for Developing and Assessing 1076 Adverse Outcome Pathways. OECD Series on Adverse Outcome Pathways, No. 1, OECD Publishing, 1077 Paris. Available online at: http://www.oecd-ilibrary.org/environment/users-handbook-supplement-to-the-guidance-document-for-developing-and-assessing-adverse-outcome-pathways_5jlv1m9d1g32-en. Last accessed February 2017.
Patlewicz, G., Simon, T., Goyak, K., Phillips, R.D., Rowlands, J.C., Seidel, S.D., Becker, R.A., 2013. Use and validation of HT/HC assays to support 21st century toxicity evaluations. Regul. Toxicol. Pharmacol. 65, 259–268.
Plant, T.M., 2012. A comparison of the neuroendocrine mechanisms underlying the initiation of the preovulatory LH surge in the human, Old World monkey and rodent. Front. Neuroendocrinol. 33, 160–168.
Prescott-Mathews, J.S., Wolf, D.C., Wong, B.A., Borghoff, S.J., 1997. Methyl tert-butyl ether causes alpha2u-globulin nephropathy and enhanced renal cell proliferation in male Fischer-344 rats. Toxicol. Appl. Pharmacol. 143, 301–314.
Quinn, A.L., 2006. Dow Corning Corporation, Health and Environmental Sciences, Non-regulated Study: Effects of Octamethylcyclotetrasiloxane (D4) on Estrous Cyclicity, Estradiol Levels and Ovarian Endpoints in the Female Fischer 344 Rats. HES Study number 10045–10102, January 9, 2006.
Quinn, A.L., Dalu, A., Meeker, L.S., Jean, P.A., Meeks, R.G., Crissman, J.W., Gallavan Jr., R.H., Plotzke, K.P., 2007a. Effects of octamethylcyclotetrasiloxane (D4) on the luteinizing hormone (LH) surge and levels of various reproductive hormones in female Sprague-Dawley rats. Reprod. Toxicol. 23, 532–540.
Quinn, A.L., Regan, J.M., Tobin, J.M., Marinik, B.J., McMahon, J.M., McNett, D.A., Sushynski, C.M., Crofoot, S.D., Jean, P.A., Plotzke, K.P., 2007b. In vitro and in vivo evaluation of the estrogenic, androgenic, and progestagenic potential of two cyclic siloxanes. Toxicol. Sci. 96, 145–153.
Rhomberg, L.R., Goodman, J.E., Bailey, L.A., Prueitt, R.L., Beck, N.B., Bevan, C., Honeycutt, M., Kaminski, N.E., Paoli, G., Pottenger, L.H., Scherer, R.W., Wise, K.C., Becker, R.A., 2013. A survey of frameworks for best practices in weight-of-evidence analyses. Crit. Rev. Toxicol. 43, 753–784.
Ruden, C., 2001a. Interpretations of primary carcinogenicity data in 29 trichloroethylene risk assessments. Toxicology 169, 209–225.
Ruden, C., 2001b. The use and evaluation of primary data in 29 trichloroethylene carcinogen risk assessments. Regul. Toxicol. Pharmacol. 34, 3–16.
Saghir, S.A., 2015. Rethinking guideline toxicity testing. Regul. Toxicol. Pharmacol. 72, 423–428.
Sarangapani, R., Teeguarden, J., Plotzke, K.P., McKim Jr., J.M., Andersen, M.E., 2002. Dose-response modeling of cytochrome p450 induction in rats by octamethylcyclotetrasiloxane. Toxicol. Sci. 67, 159–172.
Schreider, J., Barrow, C., Birchfield, N., Dearfield, K., Devlin, D., Henry, S., Kramer, M., Schappelle, S., Solomon, K., Weed, D.L., Embry, M.R., 2010. Enhancing the credibility of decisions based on scientific conclusions: transparency is imperative. Toxicol. Sci. 116, 5–7.
Siddiqui, W.H., Stump, D.G., Plotzke, K.P., Holson, J.F., Meeks, R.G., 2007. A two-generation reproductive toxicity study of octamethylcyclotetrasiloxane (D4) in rats exposed by whole-body vapor inhalation. Reprod. Toxicol. 23, 202–215.
Sloter, E.D., 2015. WIL Research Laboratories, LLC. A Dietary and Inhalation Vaginal Cytology Study of Chronically Administered Pergolide, Octamethylcyclotetrasiloxane (D4) or Decamethylcyclopentasiloxane (D5) in Aging Fischer 344 Rats. Project ID: WIL-401010, December 2015.
Stern, B.R., Kneiss, J.J., 1997. Methyl tertiary-butyl ether (MTBE): use as an oxygenate in fuels. J. Appl. Toxicol. 17 (Suppl. 1), S1–S2.
Stump, D.G., 2001. Dow Corning Corporation, Health and Environmental Sciences. An Inhalation Study of the Effects of Octamethylcyclotetrasiloxane (D4) Exposure on the Preovulatory LH Surge in Ovariectomized Female Rats. Study number 9377, Technical Report No. 2001-I0000-50592, September 7, 2001.
Suginami, H., Hamada, K., Yano, K., Kuroda, G., Matsuura, S., 1986. Ovulation induction with bromocriptine in normoprolactinemic anovulatory women. J. Clin. Endocrinol. Metab. 62, 899–903.
Svechnikov, K., Svechnikova, I., Soder, O., 2008. Inhibitory effects of mono-ethylhexyl phthalate on steroidogenesis in immature and adult rat Leydig cells in vitro. Reprod. Toxicol. 25, 485–490.
Swenberg, J.A., Lehman-McKeeman, L.D., 1999. Alpha 2-Urinary globulin-associated nephropathy as a mechanism of renal tubule cell carcinogenesis in male rats. IARC Sci. Publ. 95–118.
Swenberg, J.A., Short, B., Borghoff, S., Strasser, J., Charbonneau, M., 1989. The comparative pathobiology of alpha 2u-globulin nephropathy. Toxicol. Appl. Pharmacol. 97, 35–46.
Tng, E.L., 2015. Kisspeptin signalling and its roles in humans. Singap. Med. J. 56, 649–656.
Turck, P.A., 1999. Dow Corning Corporation, Health and Environmental Sciences. Estrogenic and Antiestrogenic Activity of Octamethylcyclotetrasiloxane (D4) in Sprague–Dawley and Fischer 344 Immature Female Rats Using a Uterotrophic Assay. MPI Study No. 416–148, DC Report No. 1998-I0000-45425, May 1999.
US-EPA, 1991. Report of the EPA Peer Review Workshop on Alpha2u-Globulin: Association with Renal Toxicity and Neoplasia in the Male Rat. Risk Assessment Forum U.S. Environmental Protection Agency. Available online at: https://archive.epa.gov/raf/web/html/rpt-peer-review-workshop-alpha2u-globulin.html.
US-EPA, 2005. Guidelines for Carcinogen Risk Assessment. Risk Assessment Forum of the U.S. Environmental Protection Agency (EPA/630/P-03/001F, March 2005). Available from: http://www.epa.gov/raf/publications/pdfs/CANCER_GUIDELINES_FINAL_3-25-05.PDF.
US-OSHA, 2016. Occupational Safety and Health Administration: Guidance on Data

Evaluation for Weight of Evidence Determination. available online at: https://www.osha.gov/weightofevidence/.
Vainiotalo, S., Peltonen, Y., Ruonakangas, A., Pfaffli, P., 1999. Customer exposure to MTBE, TAME, C6 alkyl methyl ethers, and benzene during gasoline refueling. Environ. Health Perspect. 107, 133–140.
van den Driesche, S., McKinnell, C., Calarrao, A., Kennedy, L., Hutchison, G.R., Hrabalkova, L., Jobling, M.S., Macpherson, S., Anderson, R.A., Sharpe, R.M., Mitchell, R.T., 2015. Comparative effects of di(n-butyl) phthalate exposure on fetal germ cell development in the rat and in human fetal testis xenografts. Environ. Health Perspect. 123, 223–230.
Van Der Kraak, G.J., Hosmer, A.J., Hanson, M.L., Kloas, W., Solomon, K.R., 2014. Effects of atrazine in fish, amphibians, and reptiles: an analysis based on quantitative weight of evidence. Crit Rev Toxicol. 44 (Suppl. l 5), 1–66.
Williams, T.M., Borghoff, S.J., 2001. Characterization of tert-butyl alcohol binding to alpha2u-globulin in F-344 rats. Toxicol. Sci. 62, 228–235.
Wilson, V.S., Blystone, C.R., Hotchkiss, A.K., Rider, C.V., Gray Jr., L.E., 2008. Diverse mechanisms of anti-androgen action: impact on male rat reproductive tract development. Int. J. Androl. 31, 178–187.