

EuroMix

European Test and Risk Assessment Strategies for Mixtures

Project number 633172

Collaborative project

H2020-SFS-2014-2

Deliverable D5.5 – Report

Methodology and results of the retain and refine approach

WP 5 – Aggregated and cumulative exposure assessment

Due month of deliverable: 44

Actual submission month: 48

Deliverable leader: Marc Kennedy (Fera)

Document status: version 1

Creation date: 29/04/19

Project co-funded by the European Commission within the H2020 Programme		
Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission)	
RE	Restricted to a group specified by the consortium (including the Commission)	
CO	Confidential, only for members of the consortium (including the Commission)	

Report on methodology and results of the retain and refine approach

May 2019

Authors: Marc Kennedy, Andy Hart (FERA), Johannes W. Kruisselbrink, Marco van Lenthe, Waldo J. de Boer, Hilko van der Voet (WUR Biometris), Emiel Rorije, Corinne Sprong, Jacob van Klaveren (RIVM)

Contents

1. Introduction	3
1.1. Retain and refine conceptual model	4
2. Model components developed within Euromix	5
2.1. Probability of causing a specified effect (CAG Level 2)	5
2.1.1. Estimating probabilities based on QSAR results	6
2.2. Uncertainties due to missing exposure data	13
2.3. Uncertainties due to missing toxicity data	14
2.4. Uncertainties affecting other parts of cumulative assessment	15
3. Implementation in MCRA	16
3.1. Tiered approaches	19
3.2. Uncertainty	20
3.2.1. Quantification of uncertainties within the assessment model	20
3.2.2. Uncertainty and deterministic tiers	20
3.2.3. Quantification of uncertainties outside the model	21
3.3. Use of tiered approach and uncertainty in retain and refine	21
3.4. Probabilistic approach to translate the uncertainties implied in the above low tiers to the final risk assessment	22
4. Worked example – steatosis	23
4.1. Retain step	23
4.2. The Refine step	34
4.3. Sensitivity analysis	35
5. Discussion and Conclusions	36
References	38
Appendix I. Bias as component of variability and uncertainty	40

1. Introduction

The Euromix project has developed new models and strategies to identify groups of chemicals to consider in a risk assessment. One of the aims is to develop practical tools that can be applied to larger groups than was previously possible. How to select the most important compounds that are relevant for a given health effect, from the many thousands in use, is an important but difficult problem. The purpose of this report is to:

- demonstrate the need to quantify uncertainty about which chemicals should be included in cumulative risk assessment, and about their combined effect on risk,
- demonstrate the need to quantify other sources of uncertainty, rather than exclude the chemicals they affect or make conservative assumptions,
- describe practical methods for doing this,
- illustrate application of the proposed approach with a worked example
- use the worked example to explore the potential magnitude of some key uncertainties.

There has long been concern about the potential cumulation of risk when considering chemicals collectively rather than individually (EFSA, 2008; Sarigiannis and Hansen, 2012). How risk cumulates across chemicals depends on their detailed mechanisms of action and how those interact. Research is ongoing to investigate those mechanisms and model them explicitly, but this is a complex and resource-intensive undertaking, especially when more than a small number of chemicals need to be considered. There is therefore substantial interest in whether empirical, rather than mechanistic, models may provide a more practical approach for regulatory assessment of cumulative risk, at least in the short to medium term.

Several types of empirical model have been considered: dose addition, effect or response addition (one version of which is referred to as independent action), synergistic action and antagonistic action. The European Food Safety Authority (EFSA) has recommended that dose addition be assumed as the default model for regulatory assessment of cumulative risks of pesticides in the EU (EFSA, 2018b) and this is assumed in the examples presented in this report.

Cumulative risk assessment requires specification of the group or groups of chemicals to be considered. First, cumulative assessment may be limited to chemicals that are subject to a common regulatory framework, e.g. pesticides. This will lead to underestimation of risk, if chemicals from multiple regulatory areas contribute to the same cumulative risk.

Second, it is necessary to assess which of the chemicals considered contribute to cumulative risk. For this purpose, the European Food Safety Authority (EFSA) has defined a hierarchy of 'cumulative assessment groups' (CAGs) with four Levels (EFSA, 2013). A Level 2 CAG comprises chemicals that cause the same phenomenological effect in the same organ.

In reality, it is uncertain which chemicals belong to each CAG. It cannot be stated with certainty which belong, and which do not; rather, some chemicals are more likely to belong, and other chemicals less likely. Current approaches deal with this uncertainty by using a combination of evidence, criteria and expert judgement to assign chemicals as belonging or not belonging. However, each chemical that is assigned as belonging has some probability of not belonging, and vice versa. A recent example where membership probabilities have been assigned based on expert judgement is described in EFSA (2018c) for a CAG of pesticides linked to effects on the nervous system. Rather

than assigning individual probabilities to each compound, experts were asked to assign a probability distribution to the number of substances in each sub-group causing the effect. Given the large numbers of chemicals that could potentially be included in some CAGs, the potential scale of under- or over-estimation of risk could be very substantial, leading to under- or over-regulation.

Analogous issues arise in relation to other sources of uncertainty, including missing data. For example, if chemicals for which either toxicity or exposure data are lacking are excluded from the assessment, their contribution will be excluded. This will tend to make the assessment unconservative, under-estimating cumulative risk. On the other hand, if a source of uncertainty is addressed by using conservative assumptions, this will lead to over-estimation of risk. Examples of this include the use of the threshold of toxicological concern (TTC) for chemicals that lack in vivo toxicity studies, and replacing censored (non-detect) concentration data by the limit of detection. When assessing risk for a single chemical, it may be possible for assessors to make some judgement about the degree of conservatism or unconservatism introduced by different assumptions, and their combined impact on the conservatism of the risk assessment as a whole. However, this is much more difficult in a cumulative assessment, involving multiple chemicals each subject to multiple assumptions, combined in a complex manner. In summary, dealing with uncertainty by excluding chemicals from the assessment, using conservative or unconservative assumptions, or a combination of these, produces assessments which have unknown levels of conservatism.

These problems can be avoided by quantifying the uncertainties affecting the assessment. Conservative or unconservative assumptions should be replaced, where possible, by probabilities or probability distributions that quantify the uncertainties they address (e.g. use the distribution the TTC is derived from, rather than the TTC itself). Quantify uncertainty about CAG membership using probability, retain all possible CAG members in the assessment, and conduct the assessment in a way that takes account of the probabilities. This leads to an estimate of cumulative risk with a distribution or confidence interval which includes the combined effect of the uncertainties about CAG membership and other quantified sources of uncertainty. Such an output can be used to derive point estimates of cumulative risk with specified levels of conservatism, which in turn enables risk managers to control the degree of precaution they apply in their regulatory decisions. It also makes it possible to analyse the contribution of each chemical to overall uncertainty and target refinement of the assessment (when needed) on those that contribute most.

We refer to the latter approach as ‘retain and refine’, since all possible CAG members are retained in the assessment, and refinement is targeted on those that contribute most to the uncertainty of the cumulative risk. This differs fundamentally from approaches that deal with uncertainty by excluding chemicals or making assumptions, and approaches that refine cumulative assessment by progressively excluding more chemicals from the CAG, which lead to unknown levels of conservatism. The following sections consider in more detail the conceptual model for the ‘retain and refine’ approach, describe practical methods for implementing the model, and illustrate it with a worked example.

1.1. Retain and refine conceptual model

EFSA address the question of which chemicals to include in a cumulative risk assessment through the concept of cumulative assessment groups (CAGs) with four hierarchical levels (EFSA, 2013):

- CAG Level 1 – chemicals affecting the same organ
- CAG Level 2 – chemicals within a Level 1 CAG that cause the same phenomenological effect

- CAG Level 3 – chemicals within a Level 2 CAG that cause the same effect through the same mode of action
- CAG Level 4 – chemicals within a Level 3 CAG that cause the same effect through the same mechanism of action.

The purpose of CAG Levels is to inform decisions about which chemicals to include in a cumulative assessment, and how to combine them. Uncertainty about this can be considered at two levels: first, uncertainty about whether a chemical causes a given phenomenological effect and, second, given a group of chemicals that do cause the same effect, uncertainty about whether their combined effect is consistent with the model of dose addition. Both these uncertainties can be quantified as probabilities (and will be discussed in more detail in a separate publication Hart *in prep*) however we focus here only on the first of these, i.e. CAG level 2 probabilities (Section 2.1). When large numbers of compounds are retained in the analysis, practical difficulties arise due to incomplete exposure and toxicity data for many of them. Proposed solutions are addressed in sections 2.2-2.3. Potential approaches for other sources of uncertainty in cumulative risk assessment are introduced in subsequent sections, 2.4 – 2.5 although these are outside the scope of EuroMix and not currently implemented.

2. Model components developed within Euromix

2.1. Probability of causing a specified effect (CAG Level 2)

In keeping with the general paradigm for chemical risk assessment, risk for different phenomenological effects are assessed separately, whether in the same or different organs. On this basis, cumulative assessment is restricted to members of the same Level 2 CAG. However, we note in passing that, if cumulative risks for individual effects were high enough, then it would become relevant to consider also combination at the organ or organism level.

In this report it is proposed to quantify the uncertainty about whether a chemical does or does not cause a specified effect, using probability. In principle, it may be possible to analyse study data using statistical methods. However, this would only take account of those uncertainties that are reflected in the variability of the data. It would not reflect important additional uncertainties, such as those relating to the quality and reporting of the study, and the relevance of the observed endpoints to the effect of interest, assessment of which requires expert judgement. Probabilities could in theory be obtained using structured methods for expert judgement (expert knowledge elicitation or EKE, EFSA 2014), taking account of both the reported results of the available studies and the additional uncertainties. Examples of this approach are developed in EFSA (2018c) in which experts considered the number of compounds within each of 7 subgroups that cause an effect. A distribution was also elicited for the total number of compounds causing the effect. Combining the resulting probability distributions, it would be possible to assign probabilities of CAG membership.

Such probabilities, whether from data analysis or expert judgment, can be used in cumulative risk assessment to take account of uncertainty about which chemicals really cause the effect of interest. This requires probabilistic calculations, which are simple to implement using Monte Carlo simulation. The calculation of cumulative risk is repeated multiple times, using the probability for each chemical to determine whether it is included or excluded in each iteration. When this is done for multiple chemicals in the same assessment, each iteration of the calculation will give a different cumulative risk, depending on which chemicals are included or excluded. Those results form a distribution

quantifying the uncertainty of the cumulative risk that results from the uncertainty about which chemicals really cause the effect of interest.

In principle, a separate probability is needed for each pesticide. However, conducting a detailed elicitation process separately for each chemical will require substantial time and resources when there are many chemicals to be considered. A practical alternative is to group the chemicals, based on criteria that the experts consider important for assessing the probability of causing the effect of interest. One could then elicit a probability distribution (representing uncertainty) for the proportion of chemicals in a specified group that cause the effect, or a discrete distribution representing uncertainty about the total number of chemicals in the group that cause the effect. Both options are feasible, with contrasting technical advantages and disadvantages that will not be discussed further here.

In vivo data may include false negatives as well as false positives, so it is also possible that the existing Level 2 CAGs exclude some pesticides that should be included. This could be addressed by the same method as described above. If experts considered that all pesticides with no effects reported in animal studies could be treated as a homogeneous group, i.e. that their individual probabilities would be identical or closely similar, then an uncertain proportion or discrete distribution could be elicited for the group as a whole. However, if there was some basis for differentiating these chemicals into smaller groups with different probabilities – e.g. based on chemical structure, or other effects seen in animal studies – or for assessing their probabilities individually – e.g. using in silico or in vitro data – this would be preferable to assessing them all collectively.

Within the EuroMix project, it was not feasible to elicit expert opinion on the CAG memberships of the CAG level2 steatosis compounds. A statistical model was used to estimate individual CAG membership probabilities from QSAR test data. Methods and example results are described in Section 2.1.1.

EFSA's work has thus far considered only pesticides. In future, it may be required to include other classes of chemicals in cumulative risk assessment. Animal studies may be available for some of those chemicals, but not others. It will therefore be necessary to consider what other types of evidence could be used for those. Similar approaches to those outlined above could then be applied: assess probabilities for individual chemicals where possible, e.g. using in silico methods; or form groups of chemicals, e.g. based on chemical structure, and elicit an uncertain proportion or discrete distribution for each group.

2.1.1. Estimating probabilities based on QSAR results

Data were compiled from runs of QSAR models on a large collection of training compounds with known in/out steatosis status (Ref to WP2 deliverable). From these data, summaries were calculated for each QSAR model as (TP, FP, TN, FN) = number of True Positives, False Positives, True Negatives, False Negatives.

To determine whether any given compound is relevant or not to the chosen health effect, a Bayesian weight of evidence approach similar to that already used in the OSIRIS project (Rorije *et al*, 2013) was applied. For a given health effect, we let I_i be the indicator that compound i is in the set of compounds leading to the effect (requiring inclusion in the CAG). Specifically, we have $I_i = 0$ if compound i is not relevant and $I_i = 1$ if it is relevant. Let $p_i = P(I_i = 1)$ denote the prior probability that compound i belongs in the CAG and let p'_i denote the posterior probability, conditioned on the observed QSAR model training data. Prior information about I_1, \dots, I_N might be obtained from

experts assigning probabilities. If compounds were selected at random with no knowledge at all whether they will influence the health effect, then the default $p_i = 0.5$ might be set. In the example presented in Section 4, this default is used simply for illustration. In a real risk assessment, if the compounds are selected from a list already assessed to be part of the steatosis CAG (DTU, 2012), a higher value may be appropriate.

For compound i , data are available from M QSAR models Q_1, Q_2, \dots, Q_M . In the simplest case, data are yes/no results¹ and we can express the result for the i th compound and k th model as

$$q_{ik} = \begin{cases} 0 & \text{negative result} \\ 1 & \text{positive result} \end{cases}$$

If we assume that the I_i are independent, i.e. information about one compound does not influence beliefs about the inclusion of any other, then Bayes' Theorem can be applied separately to each compound

$$p'_i = P(I_i = 1 | q_{i1}, q_{i2}, \dots, q_{iM}) = P(I_i = 1 | \mathbf{q}_i) = p_i P(\mathbf{q}_i | I_i = 1) / P(\mathbf{q}_i) \quad (1)$$

and p'_i can then be used in the Bernoulli simulation as part of the probabilistic exposure calculation.

Calculation of (1) requires

- i. The prior p_i which is set by default to 0.5 or is obtained from expert opinion and/or knowledge of the selection of compounds to be analysed by QSAR;
- ii. $P(\mathbf{q}_i | I_i = 1)$ which is the joint probability of obtaining the observed results when the correct classification is to include the compound;
- iii. $P(\mathbf{q}_i)$ which is the joint probability of obtaining the observed results.

Probabilities ii and iii were estimated using training data where compounds are known to be positive or negative, as explained below. The errors in the approximations of these probabilities depend on the number of positive and negative results in the training set.

The calculations can be expanded as

$$P(\mathbf{q}_i | I_i = 1) = p(q_{i1} | I_i = 1) p(q_{i2} | q_{i1}, I_i = 1) p(q_{i3} | q_{i2}, q_{i1}, I_i = 1) \dots p(q_{iM} | q_{i1}, \dots, q_{iM-1}, I_i = 1) \quad (2)$$

and

$$P(\mathbf{q}_i) = p(q_{i1}) p(q_{i2} | q_{i1}) p(q_{i3} | q_{i2}, q_{i1}) \dots p(q_{iM} | q_{i1}, \dots, q_{iM-1}) \quad (3)$$

so that (1) becomes

$$p'_i = p_i P(q_{i1} | I_i = 1) / P(q_{i1}) \prod_{m=2}^M P(q_{im} | q_{i1}, \dots, q_{im-1}, I_i = 1) / P(q_{im} | q_{i1}, \dots, q_{im-1}). \quad (4)$$

In general, useful information about dependence is not available, so the simplified versions of (2)-(4) that assume conditional independence are used:

$$P(\mathbf{q}_i | I_i = 1) = \prod_{m=1}^M p(q_{im} | I_i = 1) \quad (5)$$

$$P(\mathbf{q}_i) = p_i P(\mathbf{q}_i | I_i = 1) + (1 - p_i) P(\mathbf{q}_i | I_i = 0) \quad (6)$$

$$p'_i = p_i P(\mathbf{q}_i | I_i = 1) / P(\mathbf{q}_i). \quad (7)$$

The Cooper statistics *Sensitivity* (Se_m) and *Specificity* (Sp_m) are required for each test, where $Se = TP / (TP + FN)$ is the proportion of true positive test results when $I_i = 1$ and $Sp = TN / (FP + TN)$ is the proportion of true negative test results when $I_i = 0$.

¹ Alert based systems are also treated in this way, although it may be better to interpret the absence of an alert to be weaker information than a true negative. A problem in that case would be that the specificity and sensitivity measures are derived assuming an absence of an alert is a negative result.

These are translated into the required probabilities as follows.

$$P(q_{im}|I_i = 1) = \begin{cases} p(FN) = 1 - Se_m & q_{im} = 0 \\ p(TP) = Se_m & q_{im} = 1 \end{cases}$$

and

$$\begin{aligned} P(q_{im}) &= \begin{cases} p([I_i = 1 \text{ and } FN] \text{ or } [I_i = 0 \text{ and } TN]) & q_{im} = 0 \\ p([I_i = 1 \text{ and } TP] \text{ or } [I_i = 0 \text{ and } FP]) & q_{im} = 1 \end{cases} \\ &= \begin{cases} p_i(1 - Se_m) + (1 - p_i)Sp_m & q_{im} = 0 \\ p_iSe_m + (1 - p_i)(1 - Sp_m) & q_{im} = 1 \end{cases} \end{aligned}$$

We also use the fact that

$$P(q_{im}|I_i = 0) = \begin{cases} p(TN) = Sp_m & q_{im} = 0 \\ p(FP) = 1 - Sp_m & q_{im} = 1 \end{cases}$$

When considering many compounds, practical difficulties can arise where the model computations result in missing values for q_{im} . If model results are missing for all $m = 1, \dots, M$ then the posterior probability is set equal to the prior probability (no updating of information is possible). Otherwise, the products used in (5) and (6) simply omit the missing components (the updating for each compound i uses only the test data that are available for that compound).

In the description of the model, QSAR model outputs are assumed, but the method can be applied to any test results with binary outputs. In some of the examples below, we include molecular docking results where the output is converted to 0 or 1 depending on whether the docking energy exceeds a threshold.

As a test case of the method, we considered the liver steatosis effect which has been widely studied in Euromix. To estimate probabilities of steatosis CAG membership, a validation dataset containing 207 compounds (104 active and 103 inactive compounds) was used. As described in Euromix deliverable D2.4, this dataset comprises pharmaceuticals known to be positive/negative for steatosis in human studies (Donato et al. 2012; Benet et al. 2014; Jennings et al. 2014; Tolosa et al. 2016), PPPs from the EFSA CAG steatosis (mammals, as documented in the EFSA supporting publications on CAG Liver toxicity, and in EuroMix D2.1 the toxicity database) and from Al Eryani et al. 2015 (*in-vitro* studies). From these datasets the Cooper statistics were calculated to give sensitivity and specificity for each test (5 QSAR tests and 16 molecular docking tests).

Assuming independence between tests and fixed prior probabilities for inclusion of each compound, it is straightforward to derive posterior probabilities of inclusion for new compounds or to cross-validate the original data using the observed tests. Here, the prior was set as the proportion of compounds found to be positive for steatosis from a larger set of compounds ($79/327 = 0.242$)

Some results are shown in figures 1 - 2 showing the difference in probabilities under alternative strategies for selecting CAG members:

- Include if the majority of QSAR tests are positive, for each compound
- Include probabilities within the simulation to propagate the uncertainty (the method applied in retain and refine). Probabilities are calculated from binary QSAR outputs only
- Include probabilities calculated from both QSAR tests and (binary) molecular docking outputs

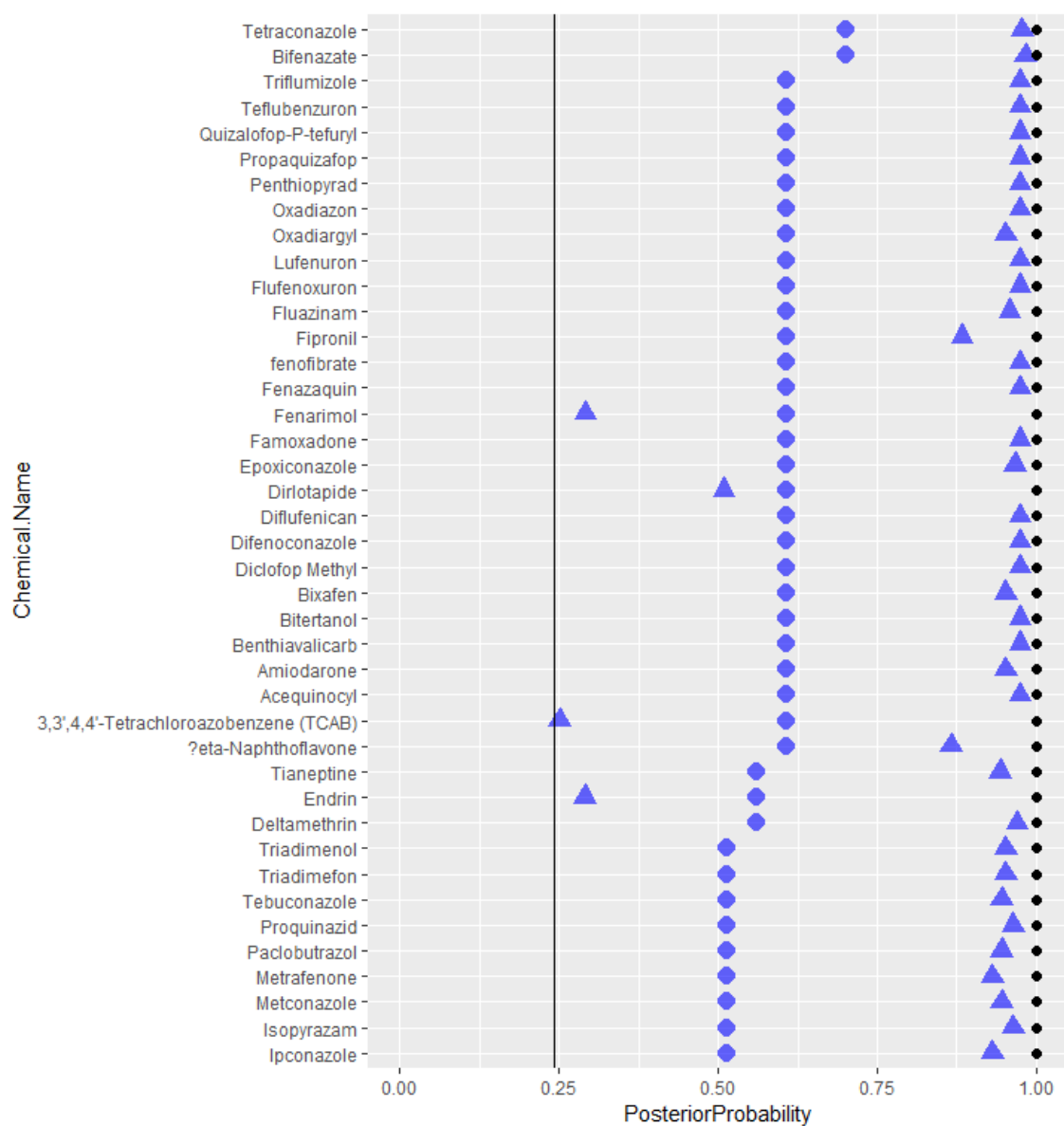
Further uncertainty could be added for the true unobserved test error rates, that accounts for the limited sample sizes of the validation sets. The simplest method would be to assume independent beta distributions per compound.

Steatosis validation set results:

In the following plot, blue points correspond to probabilities of inclusion for true steatotic compounds and red points are for true non-steatotic compounds. We would therefore expect the probabilities for blue points to be close to 1 and probabilities for red points to be close to zero. These calculations are a subset of the 207 compounds identified in the Euromix project as a validation set for steatosis QSAR model development and testing (3 compounds were excluded due to missing test results).

Results from 5 QSAR models were used to predict the steatosis status of each compound in the validation set. As a comparison, results from molecular docking tests were also included. These tests were treated as positive if the binding energy exceeded a set threshold. The wide spread of probabilities seen in figures 1-2 for steatosis CAG membership shows the importance of keeping these probabilities rather than setting them to 0 or 1. The majority rule is seen to be close to the probabilistic results in most cases, if we were to apply a simple threshold assignment depending on whether p is above or below 0.5. In most cases the inclusion of docking results generates a stronger result in the sense of shifting the probability closer to the correct 0 or 1 value. There are some exceptions, in which the inclusion of docking results appears to give a different result (e.g. phosmet in Figure 1 or ethametsulfuron methyl in Figure 2). Examples are also seen in which including docking scores leads to correct predictions whereas QSAR scores alone do not (flusilazole, cyproconazole, Figure 1).

Including docking scores for the inactive substances (Figure 2) leads to more false positives than using the QSAR tests alone, seen as more red triangles above 0.5 when the compounds listed in Figure 2 are the non-steatotic compounds (and therefore results should be closer to 0).



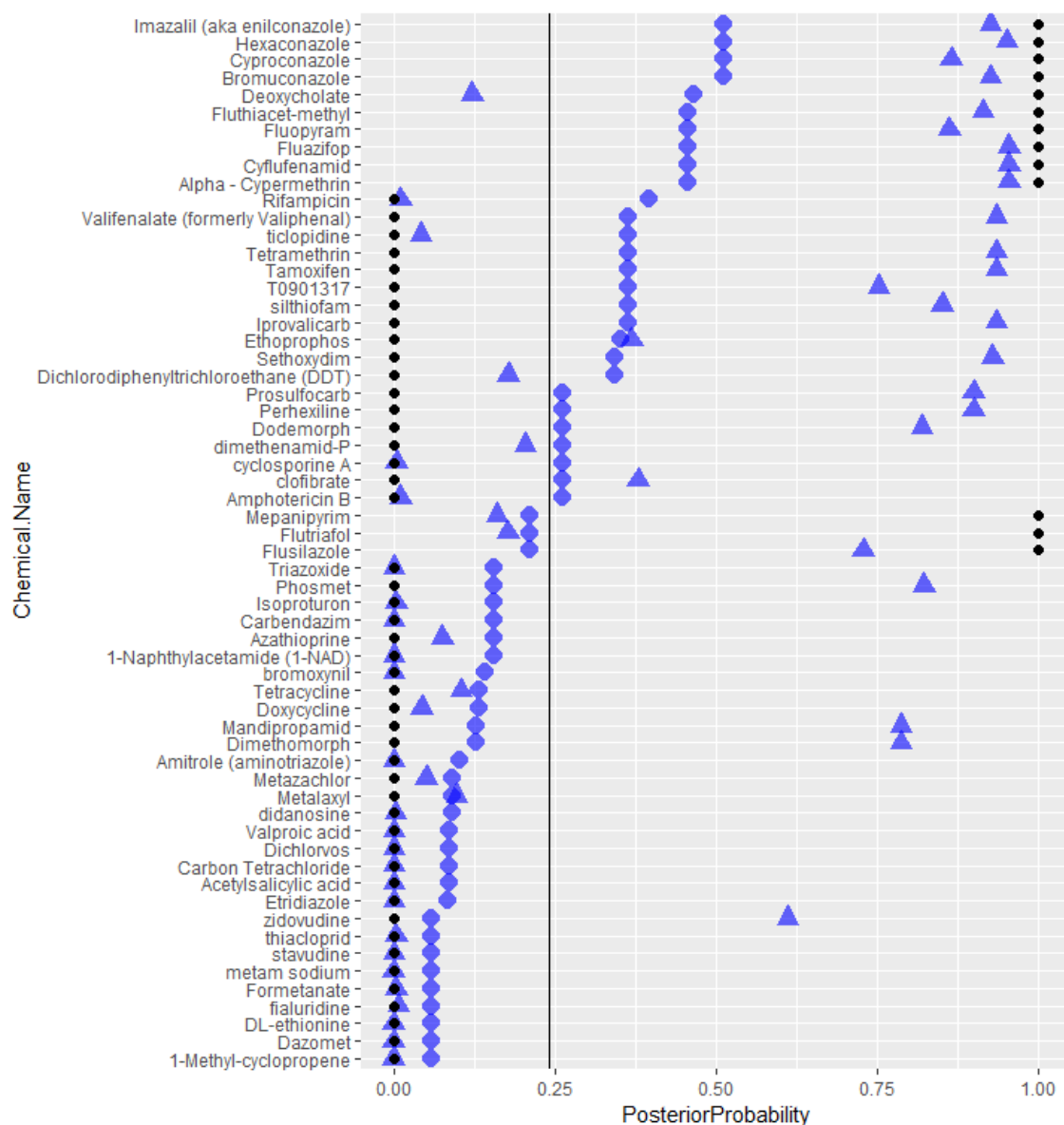


Figure 1: Assignment and Bayesian probabilities for each compound. The black dots are corresponding classifications using a simple majority scoring rule, where 1 is assigned if and only if more than half of the tests are positive. Blue points are for active steatosis compounds. Triangles indicate the use of QSAR + docking results combined. The vertical line at 0.242 indicates the prior probability in this case.



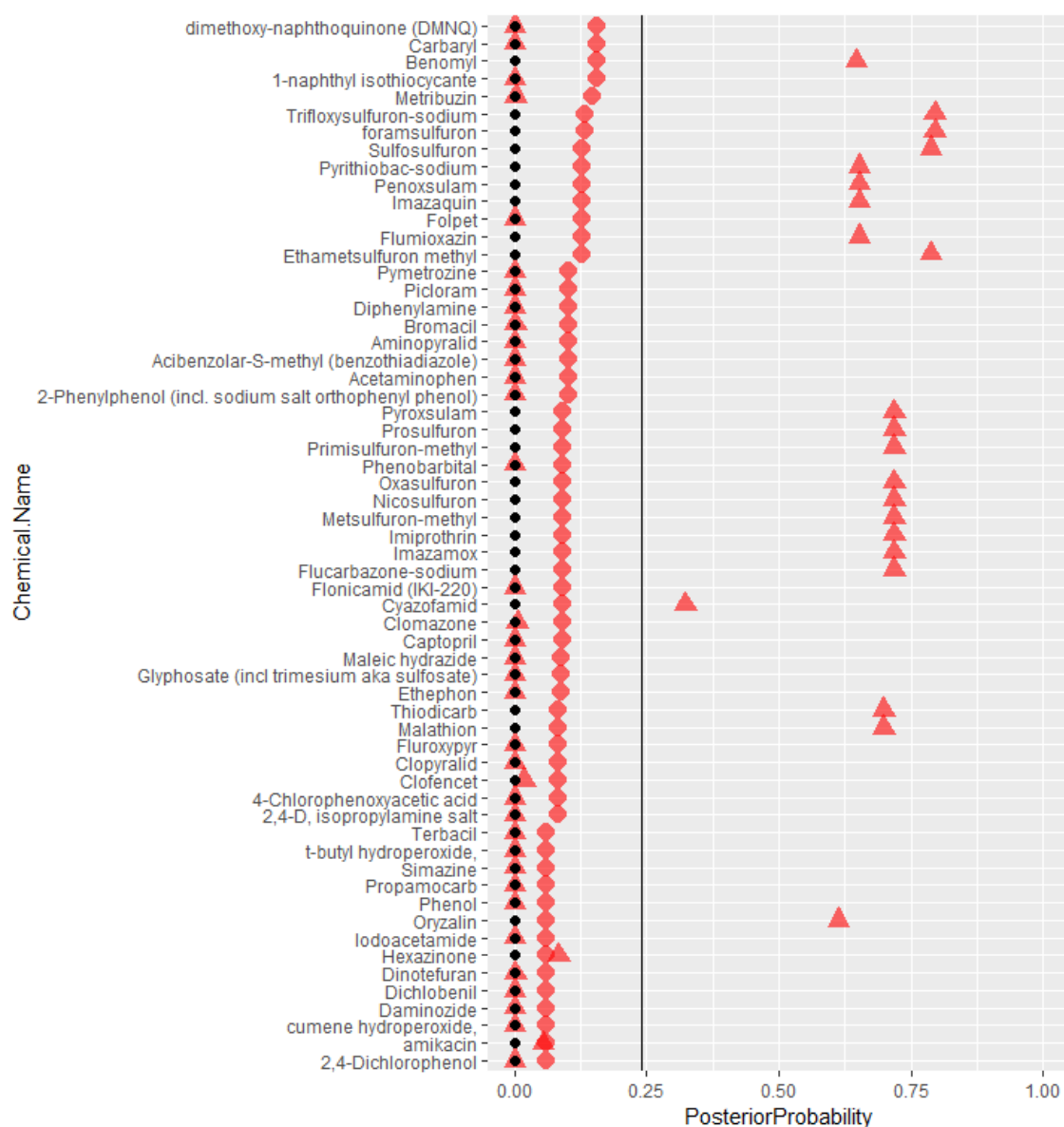


Figure 2: Assignment and Bayesian probabilities for each compound. The black dots are corresponding classifications using a simple majority scoring rule, where 1 is assigned if and only if more than half of the tests are positive. Red points are compounds that are inactive for steatosis. Triangles indicate the use of QSAR + docking results combined. The vertical line at 0.242 indicates the prior probability in this case.

2.2. Uncertainties due to missing exposure data

For some chemicals, exposure assessments may be lacking. This is true, for example, of pesticides for which no occurrence data are available. EuroMix has developed a simple approach for quantifying uncertainty about the missing exposure distribution, by taking exposure distributions for other pesticides and sampling whole distributions at random for use as surrogate distributions for the missing cases. This implies an assumption that the exposures for pesticides with missing data are exchangeable with (a random sample from) those for which data are available; uncertainty about this assumption also needs to be considered. It may be possible to apply a similar approach to missing

exposure distributions for other chemical classes, if sufficient actual distributions are available to sample from.

2.3. Uncertainties due to missing toxicity data

In some cases, there are substances that are known to cause (or may possibly cause) the effect of interest, but for which there are no data available for obtaining hazard characterisations. i.e. no points of departure or dose response models. Instead of excluding these substances in a retain and refine analyses, it is also possible to impute hazard characterisations for these substances based on hazard characterisations of other (similar) substances, and use these for calculating, e.g., relative potency factors or for risk assessment. Within MCRA 9 the following options are available for imputation of missing values:

Munro P5 (TTC approach):

The Threshold of Toxicological Concern (TTC) is an example of a tier for extrapolation of hazard characterisations from other compounds that is already in common use (see Munro, 1996). The Munro collection of NOELs/LOAELs is a collection of NOELs/LOAELs for chemicals for the critical (i.e., first occurring) effect. In the TTC approach, the toxicity of an unknown substance is imputed using the 5th percentile NOAEL of the sub-collection for chemicals in the same Cramer class (see Cramer, 1976).

Two variations of this approach are to use the empirical NOAEL distribution itself (just sample from the NOAEL data), or to fit a distribution (e.g. lognormal) to the empirical data and sample from the parametric distribution. MCRA provides an implementation of the TTC approach that uses the empirical distribution. In the nominal run, this implementation imputes the hazard characterisations with a value equivalent to the TTC. In the uncertainty runs, NOAELs are sampled from the empirical distribution.

The TTC is a conservative estimate of the NOAEL for at least two reasons:

1. TTCs are calculated from a collection of NOELs for the critical (i.e., first occurring) effect within each study and often the effect of interest will not be the critical effect, and therefore higher NOAELs are expected.
2. The TTC is a low percentile and therefore a conservative estimate for a random class member with unknown NOAEL.

Munro central value:

To avoid the conservatism of taking the 5th percentile in the Munro P5 approach, a nominal (or central) value can instead be taken from the Munro collection for each Cramer class. For a nominal run without uncertainty, the expected contribution of a substance with missing hazard characterisation is derived based on the risk as quantified in the hazard index

$$HI = SF \sum_{i=1}^n \frac{exp_i}{HC_i}$$

where SF are all combined safety factors (i.e. the product of safety factors), exp_i is exposure and HC_i is the hazard characterisation for compound i , e.g. the NOAEL. It follows from this equation that

an unbiased estimate for the contribution from a substance with missing hazard characterisations is obtained by taking the harmonic mean from the available NOAELs:

$$NOAEL_{central} = \left(\sum_{i=1}^n \frac{1}{NOAEL_i} \right)^{-1}$$

This is the value to use in a nominal run without uncertainty for the Munro central value approach. For the uncertainty runs, this approach also uses random sampling from the empirical distribution of the corresponding Cramer class.

Available hazard characterisations distribution P5:

Another way to avoid the conservatism of the TTC approach, is to use the effect-specific hazard characterisations of the substances for which these are available (rather than the full set of effect non-specific study NOAELs which is used for TTC). This collection will have on average higher NOAELs than those of the Munro NOAEL collection, because for many substances, the effect of interest will not be the critical effect.

Available hazard characterisations distribution central value:

Similar to the Munro central value approach, a nominal/central value could also be obtained from the set of effect-specific hazard characterisations distribution for imputation of hazard characterisations. This approach is considered to yield the most realistic, or unbiased imputation value for missing hazard characterisations.

2.4. Uncertainties affecting other parts of cumulative assessment

There will be sources of uncertainty in every part of the cumulative assessment. As explained in Section 1, the same principle applies to them all: dealing with uncertainty by excluding chemicals from the assessment, or using conservative or unconservative assumptions, produces assessments which have unknown levels of conservatism. Below, we list some other sources of uncertainty, and briefly indicate what type of approach could be considered for them. Methods for some of them are being developed in the EuroMix project or in other projects, and will be reported in detail elsewhere.

- *Uncertainty about CAG Level 1.* Section 2.1 discussed how to address uncertainty about membership of CAG Level 2 (causing the effect of interest, in the organ of interest). There can also be uncertainty about CAG Level 1 – whether the chemical affects the organ of interest. This may be taken into account within the approach used for Level 2. However, in some cases it may be useful or necessary to treat uncertainty about Level 1 separately, either for chemicals that lack animal data or to provide a prior for updating with animal data in a Bayesian approach. Many QSARs are available for predicting which chemicals cause effects on which organs, without specifying the type of effect.
- *Non-detects in occurrence data.* EFSA's (2012) guidance on probabilistic exposure assessment recommends using 'pessimistic' and 'optimistic' assumptions as a basic approach to non-detects. It is clear that the pessimistic assumption (setting NDs to the LOD) often generates extremely unrealistic results. Efforts are underway to specify some intermediate assumption, but this will suffer from the problem identified in Section 1: it will lead to exposure estimates

of unknown conservatism (unless the intermediate assumption is calibrated by a full probabilistic treatment of the uncertainty). What is needed instead is to represent the uncertainty about non-detects using appropriate distributions. Work on this is beyond the scope of the Euromix project but is considered in a related Framework Partnership Agreement between EFSA and RIVM (Kruisselbrink et al, 2018).

- *Other uncertainties in exposure assessment.* Many other uncertainties may arise in exposure assessment, e.g. processing factors, missing information on which foods additives and flavourings are used in and at what levels, use of surrogate data (e.g. for migration of chemicals from packaging), etc. Some of these are commonly treated with conservative assumptions in regulatory assessment, and need to be replaced with distributions quantifying variability and uncertainty. Where data to estimate distributions statistically are lacking, quantification of the resulting uncertainty by expert judgement will be better than using fixed assumptions (whether conservative or not).
- *Extrapolation of toxicity between and within species.* Methods for extrapolating potency (points of departure) from animals to humans and taking account of human variation have been developed by IPCS (2017). Work is needed to incorporate these or similar methods into cumulative risk assessment.

3. Implementation in MCRA

The EuroMix toolbox is developed as part of the EuroMix project, and builds on the Monte Carlo Risk Assessment (MCRA) platform (van der Voet et al. 2015, MCRA 2016), accessible at <https://mcra.rivm.nl>. Aspects that can be used as part of retain and refine are built into MCRA version 9 and these are demonstrated with a worked example in Section 4. The final software platform will be fully described in deliverable D6.4 (Scientific paper on the EuroMix toolbox and software).

The basic idea is that all entities (e.g. substances) which are potentially relevant should be included in the assessment (retain), but can be handled in different ways (more or less refined) while still being considered together in the same risk assessment. This functionality is provided by a tiered, modular design (Figure 3). Technical details are available from the MCRA website. The tiers can include uncertainty. The risk assessment (RA) framework in MCRA is intended to be flexible. RA is seen as a hierarchy of modules (Exposure and hazard assessment at the highest level, but these cover further trees of modules).

In Figure 3, each box represents a data type and an action type related to this data. The grey boxes depict **scoping modules**, having the primary entity definitions as data and selection of these primary entities as the action associated with this module. Both the green boxes and the orange boxes refer to data about these primary entities and the relationships between them. The green boxes depict the modules that perform the action of data collection/selection (**selection modules**), whereas the orange boxes depict modules that perform either collection/selection of data or calculation of such data based on other data (**selection or calculation modules**). For example, concentration models may be provided as data, but can also be calculated based on concentration data, residue limits, and agricultural use data. For risk assessments, the final output is the box labelled 'Risk', which summarises the risk of the specified health effect for the population of interest with respect to the exposure of the specified substances due to consumption or use of the specified foods/sources.

For each module multiple tiers are available, ranging from pre-sets (for inexperienced users) to completely custom settings (for experienced researchers). The tiers available are dependent on the data provided or selected by the user. For example, to include uncertainty of CAG inclusion as part of a retain and refine model, a file of QSAR or molecular docking results on training data must be provided to calculate probabilities as described in Section 2.1. Other uncertainties can also be provided by the user or calculated within MCRA. Examples are provided in Section 4.

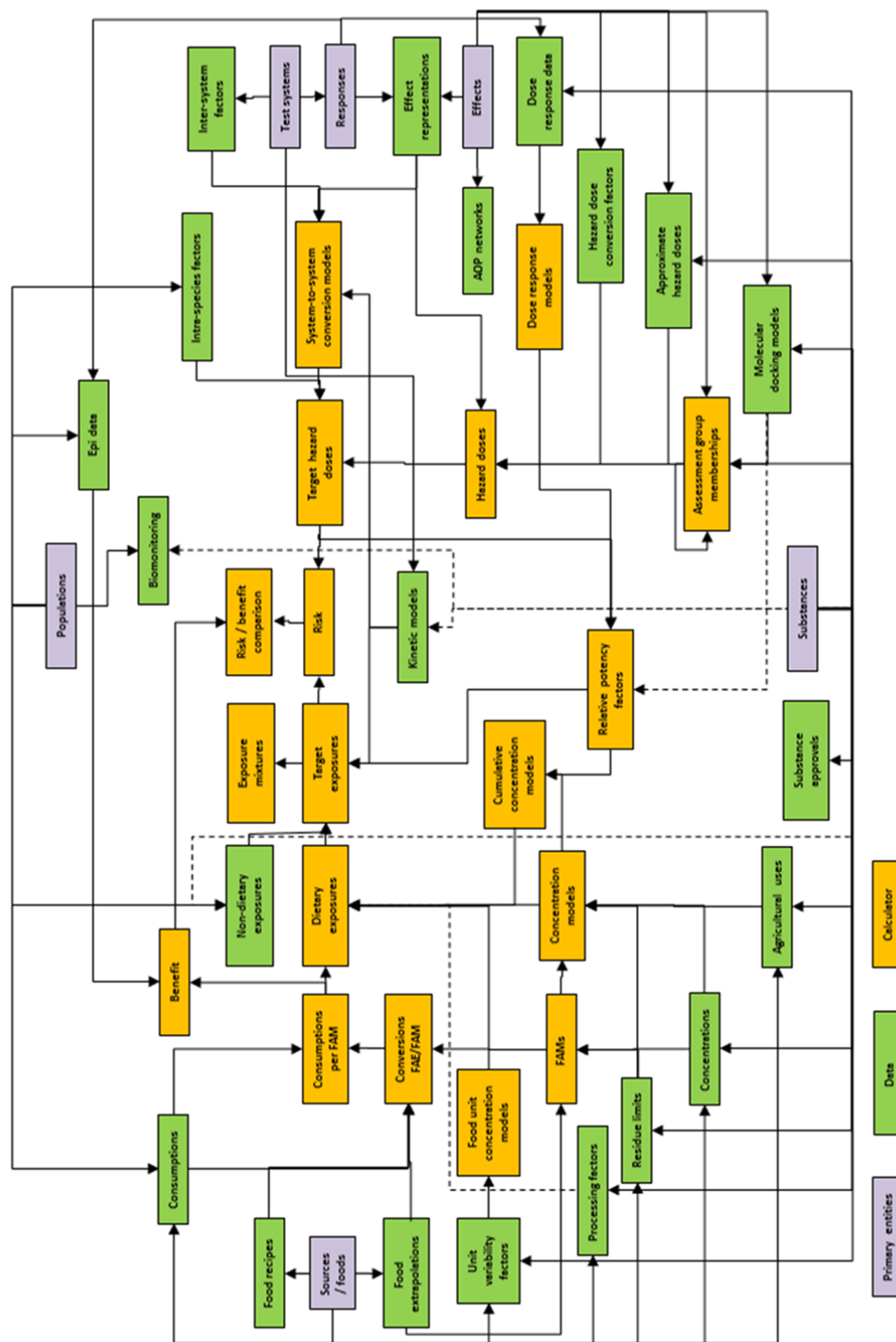


Figure 3. Modular structure of the data and actions of human health chemical risk (and benefit) assessment.

3.1. Tiered approaches

Each type of data and each calculator module can exist in several forms (or tiers). For data, these tiers correspond to different ways (i.e., formats or levels of granularity) in which the data is recorded/provided. For the calculator modules, these tiers correspond to the different models (and model options) available for performing the modelling task of the module. The overall risk assessment model is then specified as the collection of all data and sub-models.

For calculators, tiers are defined as a set of specified options. The calculator settings specify which data is used/required as input, which data (tier) is produced as output, and how the output is computed from the input. Tiers can differ in many respects, and there is no single dimension to rank tiers as low vs. high. In risk assessment, typical tiers contrast deterministic to probabilistic approaches, conservative to realistic approaches, approaches using restricted data to approaches using more extensive data, and approaches using different degrees of model complexity.

Figure 4 illustrates how a data/calculator module consists of several calculator tiers (in this case defined as method 1 to method n) and produces different data tiers (in this case containing different levels of detail with respect to variability/uncertainty).

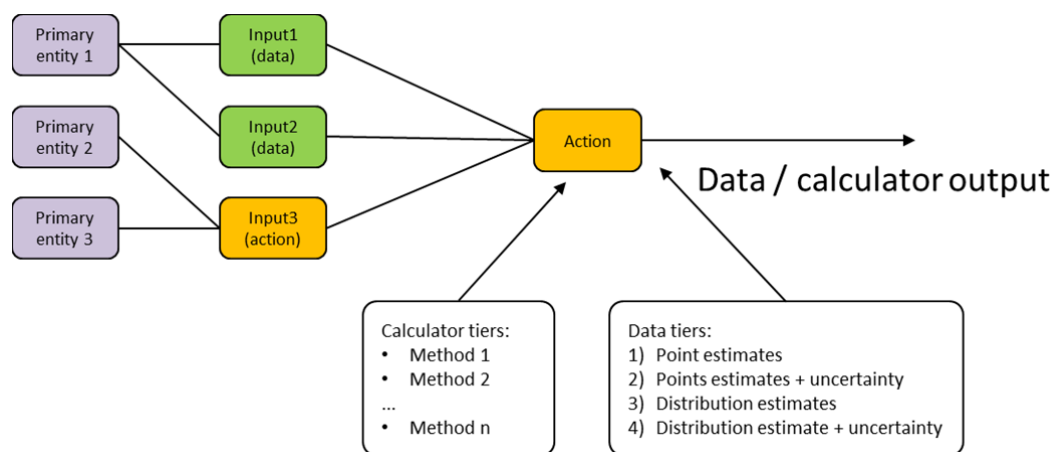


Figure 4 Illustration of how calculator tiers and data tiers belong to a data/calculator module. Data inputs are coloured green and calculators are coloured yellow

It should be noted that for a full risk assessment, involving multiple modules of the modular design, the “overall tier” is the collection of tiers chosen at the individual modules. Therefore, “overall tiers” are defined as the level of the central risk assessment module, with specific requirements on the tiers of the input-data and/or modules producing the input data.

Each calculator has as a main output an entity that can combine different tiers (tiered entities). For example, in a hazard assessment, some substances may be assessed using a tier ‘Hazard Dose from dose-response data’, other substances may be assessed using a tier ‘TTCx100’ or ‘sample from general NOAEL distribution x100’ (which only requires knowledge of the Cramer class of the substance). Here the tiered entity is substance. In dietary exposure assessment, some food x substance combinations may be recognised as risk drivers requiring a more complex approach (e.g. probabilistic modelling), whereas a simpler approach (e.g. deterministic modelling) may be sufficient for all other food-substance combinations. Therefore, in this case the tiered entity is ‘food-substance’.

3.2. Uncertainty

The true risk may be highly uncertain, and although each module can be addressed including some degree of uncertainty quantification, a flexible system is required in practice. Automatically excluding compounds due to lack of data, even when there is no strong evidence that they are safe, should be discouraged. In the retain and refine approach, all potentially-relevant compounds are retained in the assessment and refinement is targeted on those compounds with greatest uncertainty. While the nature of uncertainty varies per modelling task, ideally a consistent approach to treating uncertainties in the modelling approaches should be followed, so that uncertainty in the final risk outputs can be interpreted properly. Within EuroMix this has been implemented as far as pragmatically possible.

3.2.1. Quantification of uncertainties within the assessment model

All data are uncertain estimates of intended quantities. In some cases, part of the uncertainty (but not all, see below) can be quantified from the data themselves, e.g., by resampling techniques, in other cases a parametric specification of uncertainty can be given, and this will be part of the corresponding green (selection) box. For data that are generated by the calculators, and which may be single values or distributions representing variability between individuals or individual-days, the quantification of uncertainty can be included within the output. The uncertainty distributions can be generated by:

1. Resampling inputs with replacement but keeping the sample size fixed equal to the input sample (bootstrapping). This is designed to capture uncertainty due to limited sample size. It is used for e.g. consumption and concentration data.
2. Sampling from a parametric distribution or simulated realisations corresponding to a known degree of uncertainty in unknown input parameters. The degree of uncertainty may be obtained from statistical modelling of parameters (e.g. used for the binomial-lognormal concentration model), or it may be based on expert judgement using formal or semi-formal expert elicitation methodology (EFSA, 2014, 2018a). Elicitation may be an option where data are available but subject to limitations in reliability and/or relevance that require expert judgement (these uncertainties are not captured by the other options in MCRA). Examples might also include cases where estimates can be made based on indirect data, e.g. for other compounds (as in read across and the TTC approach).
3. Sampling different selections from pre-generated uncertainty sets (provided as additional data). This mechanism is already used for uncertainty in non-dietary exposures for a population of individuals (Kennedy *et al*, 2019). In this case, the population is assigned multiple sets of exposures corresponding to multiple uncertainty realisations.

3.2.2. Uncertainty and deterministic tiers

At deterministic tiers, where a single value is used instead of a distribution, users should have a choice whether to work with best estimates (e.g. mean values) or conservative estimates (e.g. tail percentiles). Users should be aware that it is extremely difficult to assess the degree of conservatism of deterministic cumulative assessments by expert judgement (see Section 3.2.3).

In a tiered approach it is accepted that some uncertain quantities are nevertheless modelled with fixed quantities. The uncertainties should then be handled collectively by characterising the overall impact on the assessment conclusion using a ‘combined assessment’ approach described in EFSA (2018a, section 16).

3.2.3. Quantification of uncertainties outside the model

In every assessment, all uncertainties that are not quantified within the assessment model should be assessed collectively outside the model, so that they can be included in a combined assessment of uncertainty at the end of the assessment. The rationale for this is set out by EFSA (2018, section 16), which also contains guidance on how it can be done. At simplest, the users will make a list of all the uncertainties not quantified within the model, quantify their combined impact on the assessment by expert judgement, and combine this with the uncertainty distribution output by the model. Any uncertainties which the users feel unable to include in the collective quantification can be excluded, but must be listed, described and reported alongside the quantitative assessment results (which are then conditional on the unquantified uncertainties having no impact).

In a tiered approach it is accepted that some uncertain quantities are nevertheless modelled with fixed quantities. The remaining unquantified uncertainties should then be handled collectively by the ‘combined assessment’ approach described above.

Combined uncertainty assessment should be included in every assessment that is conducted to support regulatory decision-making, because decision makers need to know how different estimated and actual risks might be (EFSA, 2018). Combined uncertainty assessment may not be needed when doing assessments for research purposes, although even here they are beneficial (e.g. to assess the reliability of research conclusions that depend on model outputs).

In view of the wide use anticipated for this methodology, a separate tool to support combined uncertainty assessment within the EuroMix toolbox has been developed (Kennedy *et al*, 2018). This replaces and improves on the ‘uncertainty table’ tool developed in the Acropolis project. It has the added benefit of supporting the assessment of uncertainty for model inputs by expert judgement, or by weight of evidence assessment (EFSA, 2017), as these can be done with the same functionality.

3.3. Use of tiered approach and uncertainty in retain and refine

In a retain and refine approach, a typical risk assessment will start at a tier that is simple to perform for all tiered entities (potential risk drivers). In the simplest possible assessment, every input would be set to a default value or distribution, representing the uncertainty of that input when no compound-specific data is available. However, based on data availability and ease of application, the initial assessment can already include more complex elements, such as probabilistic modelling. If the initial calculations produce risk estimates that do not exclude concern, refinement of the modelling for the perceived risk drivers is useful for checking whether this concern is real.

Quantification of uncertainty is inherent to the retain and refine approach, at least for CAG membership, because it allows each compound to be included in proportion to its probability of belonging (other inputs could be deterministic if the user so chose). The network shown in Figure 3 includes a node ‘Assessment Group Memberships’. As part of the input data, a probability of group membership at CAG Level 2 and/or Level 3-4 can be assigned to each compound in the list. Probability values might be generated from expert opinions or from QSAR test results, for example. In some cases, compounds might be grouped together to simplify the prior probability assessment of

group membership if data are not available. Another simple approach would be to use default probabilities based on proportions seen in previous studies.

When including uncertainty in the model run, each compound is independently simulated as being included or excluded from each CAG Level using its probability for each Level. By repeating the process multiple times, in the outer loop of a 2D Monte Carlo calculation, the impact of uncertainty is captured in the results. Note that the retain and refine concept is not compatible with total exclusion of any compound. In nominal runs, which lack an outer loop, each compound will be included but its contribution to exposure will be multiplied with its probability.

In a conventional risk assessment, another reason for excluding some compounds might be lack of data for one or more required inputs (e.g. dose-response, concentration data, etc.). In the retain and refine approach, defaults must be available for every input. Wherever possible, the defaults should be probabilistic, i.e. distributions quantifying uncertainty. Deterministic defaults can be used, but will result in the overall assessment having a degree of conservatism that is effectively unknown. This is because the complexity of cumulative risk assessment means that any expert judgement of the overall degree of conservatism deriving from many inputs with varying degrees of conservatism will be extremely unreliable.

Given that Euromix has a finite budget, it is not feasible to develop high quality defaults for every possible input. Therefore, functionality was first developed to enable users to populate defaults at a later stage, then some of these defaults were populated during the project (to enable case studies and demonstration assessments to be done). These are the uncertainties described in Section 2. For other aspects, simple 'placeholder' defaults (e.g. based on literature values or informal expert judgement) have been used in the Euromix tool for as many other inputs as possible. The basis of all included defaults is included in the documentation of the system.

3.4. Probabilistic approach to translate the uncertainties implied in the above low tiers to the final risk assessment

Given probabilities of CAG membership, when run 'with uncertainty' the outer loop of the RA model samples ones and zeroes (ins and outs) from independent Bernoulli distributions to reflect the uncertainty of membership. A different realisation of the cumulative risk is generated for each realisation in this outer loop (Figure 5).

Further at the lowest tier, hazard doses are sampled from the chemical-unspecific hazard dose (e.g. NOAEL) distribution behind the TTC approach, and exposures from a newly established chemical-unspecific exposure database.

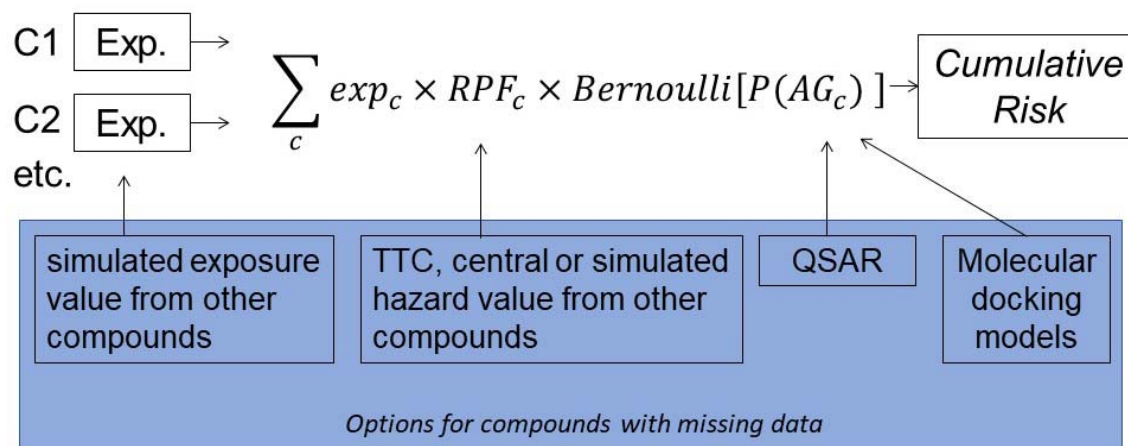


Figure 5. Conceptual model of the simulation implemented in MCRA, as used in a retain and refine analysis. In an uncertainty run, the process would be run repeatedly with potentially different realisations

For a nominal run (RA without uncertainty), which is often useful as a first approach, it is natural to multiply the estimated exposures per compound with the probability, and this has been implemented

4. Worked example – steatosis

In this section an example is presented based on one of the main case studies developed within the Euromix project. As explained in Section 3, MCRA has been updated to allow uncertainties to be quantified at each step of the calculation. This allows for retain and refine to be implemented as far as practically possible in any given scenario. The example presented below uses the data and information available within the Euromix scenario for liver steatosis plus some additional compounds to illustrate the impact of uncertainty on the analysis. The results are only included to demonstrate the retain and refine principles and implementation in MCRA and do not represent a real risk assessment.

4.1. Retain step

An example was implemented in MCRA version 9 using 100 compounds. Of these, 50 were selected as those with highest RPF values from the previously determined CAG steatosis compounds. The remaining 50 were selected randomly from the 573 compounds in the (PPP only) EuroMix substance inventory file. This second set was included to demonstrate what might happen if we retain a larger set of compounds. For the combined 100 compounds the Substances input file, together with a file of QSAR models and scores were compiled into the correct format and uploaded to MCRA. The remaining files were used from the shared EuroMix folder already uploaded in MCRA. Table 1 summarises the settings and data used in the example run with MCRA. These options correspond to some default uncertainty calculation methods built into MCRA. As explained in Section 3, more flexible options are provided for future extensions of retain and refine by allowing a user to upload externally calculated uncertainties in these or other model inputs. Imputation of missing exposures was not included in this example. For concentration models the EFSA guidance optimistic calculation

tier was used for this exercise. (Using the pessimistic model instead led to substantially higher exposure estimates, although this model has been found to generate extremely conservative results in earlier work.)

Table 1. Summary of uncertainties quantified and implementation in Euromix toolbox (MCRA version 9)

Retain and refine component	MCRA modules/actions	Data source, parameters and options used	Description
CAG membership	Active substances, with Settings = compute and select compute from QSAR membership scores	File Retain and Refine QsarMembershipModels.xlsx contains tables for QSAR membership models (Sensitivity & specificity of each QSAR test derived from training data with compounds of known CAG status) and QSAR membership scores data (QSAR results 0 or 1 for each combination of compound and QSAR test); Bayesian method used with prior probability = 0.5;	Bayesian update model computes posterior probability of CAG membership, that will be used for repeated random simulation of membership
Missing toxicity information	Hazard – Hazard doses and Target hazard doses calculation Action settings for point of departure	File Points of departure \HazardDoses.xlsx; Impute missing target doses by unchecking the option 'restrict to substances with known hazard'; Imputation method used = Munro p5 (unbiased for Bayesian method)	Missing toxicity data will be imputed based on the Munro distributions of compounds in the equivalent Cramer class. Using a p5 is a conservative option.

The steatosis CAG compounds have been used in other deliverables within EuroMix (Crépet et al, 2018; Kennedy et al, 2019).

An initial run was performed with the uncertainty option switched off. As uncertainty runs can be computationally intensive, this is useful as a first step to check all settings are appropriate and the nominal run results are sensible. Following this, an uncertainty run was performed. Results from both runs are compared below.

Results without uncertainty:

With uncertainty switched off, the point estimates from a nominal run use single imputed values for missing hazard data and the derived probability of CAG inclusion for each compound.

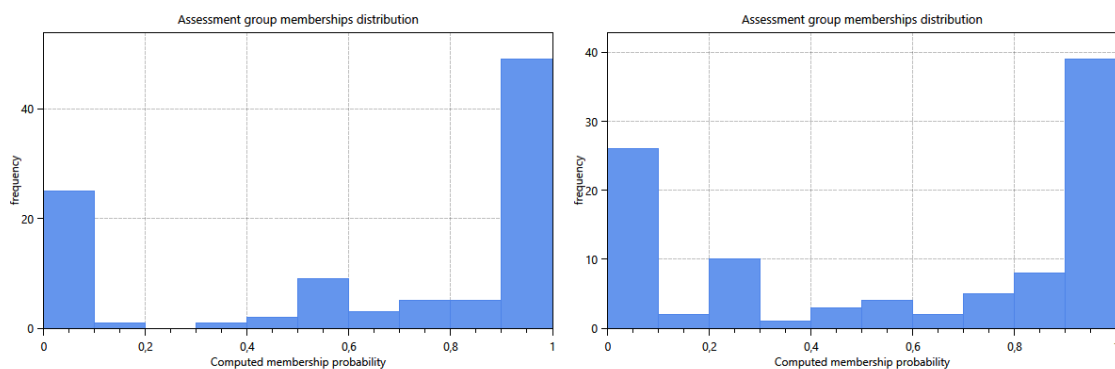
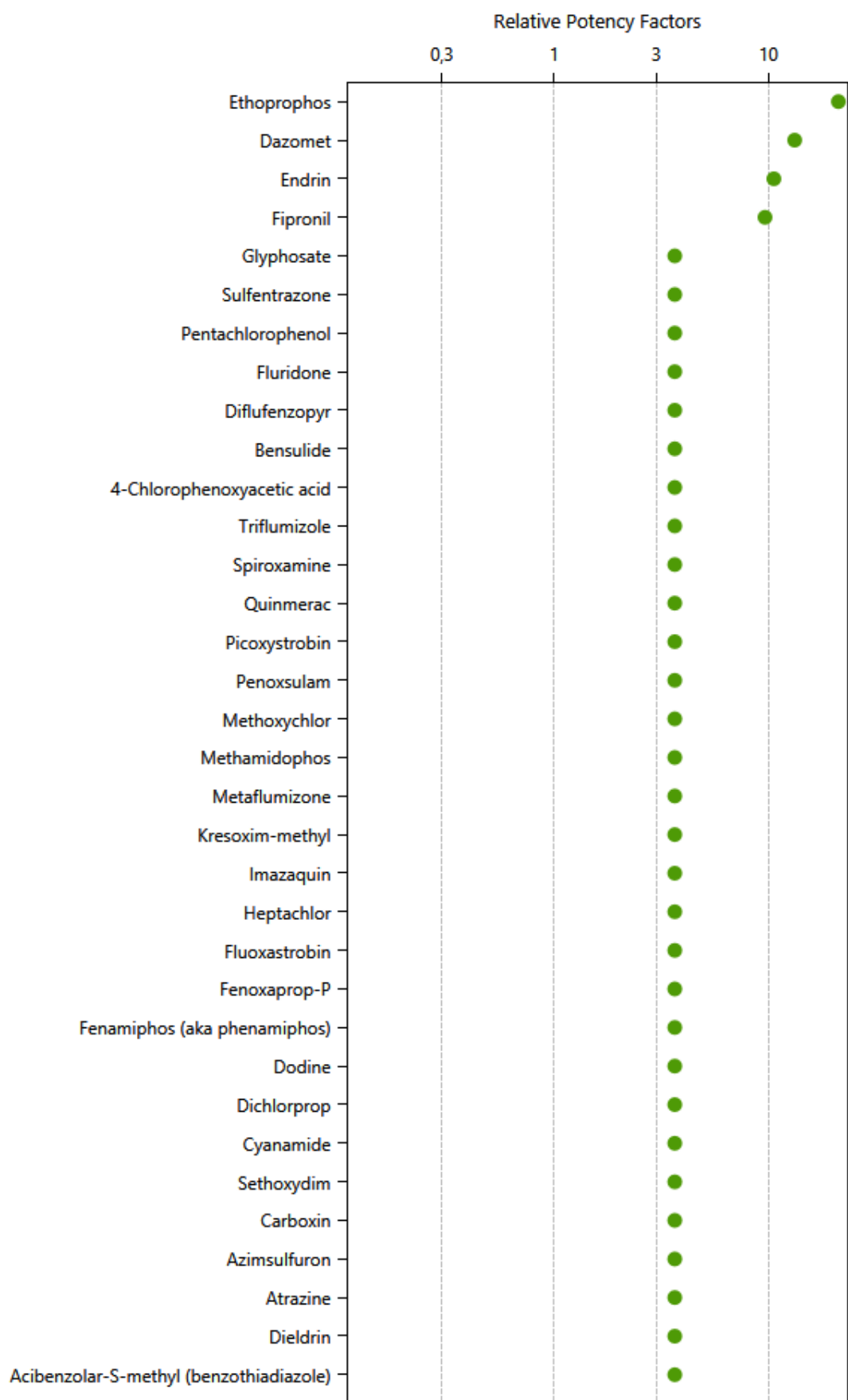
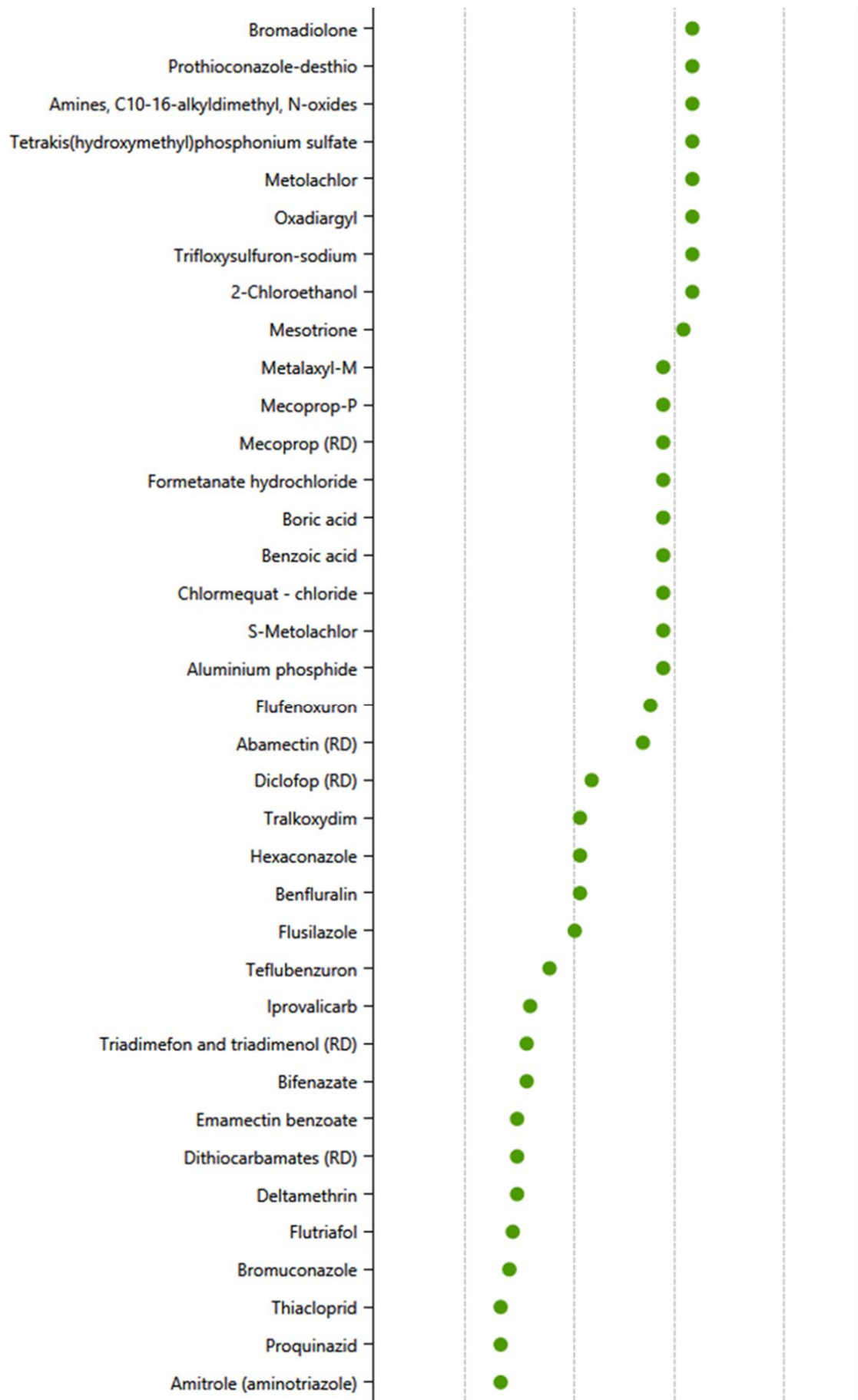


Figure 6: Histogram plot of P(CAG) probabilities of CAG membership for the 100 substances, using a default prior probability of 0.5 (left panel) or 0.252 (right panel)

The calculated probabilities for the 100 compounds are summarised in Figure 6. For the first exercise the values seen in the left panel were used, based on a default prior probability of 0.5. calculated RPF values are shown in Figure 7. Contributions to total exposure, by compound and by compound/food are shown in Figure 8. The Observed Individual Means model percentiles are shown in Table 2. The mean exposure is 0.426 ($\mu\text{g/kg bw/day}$).





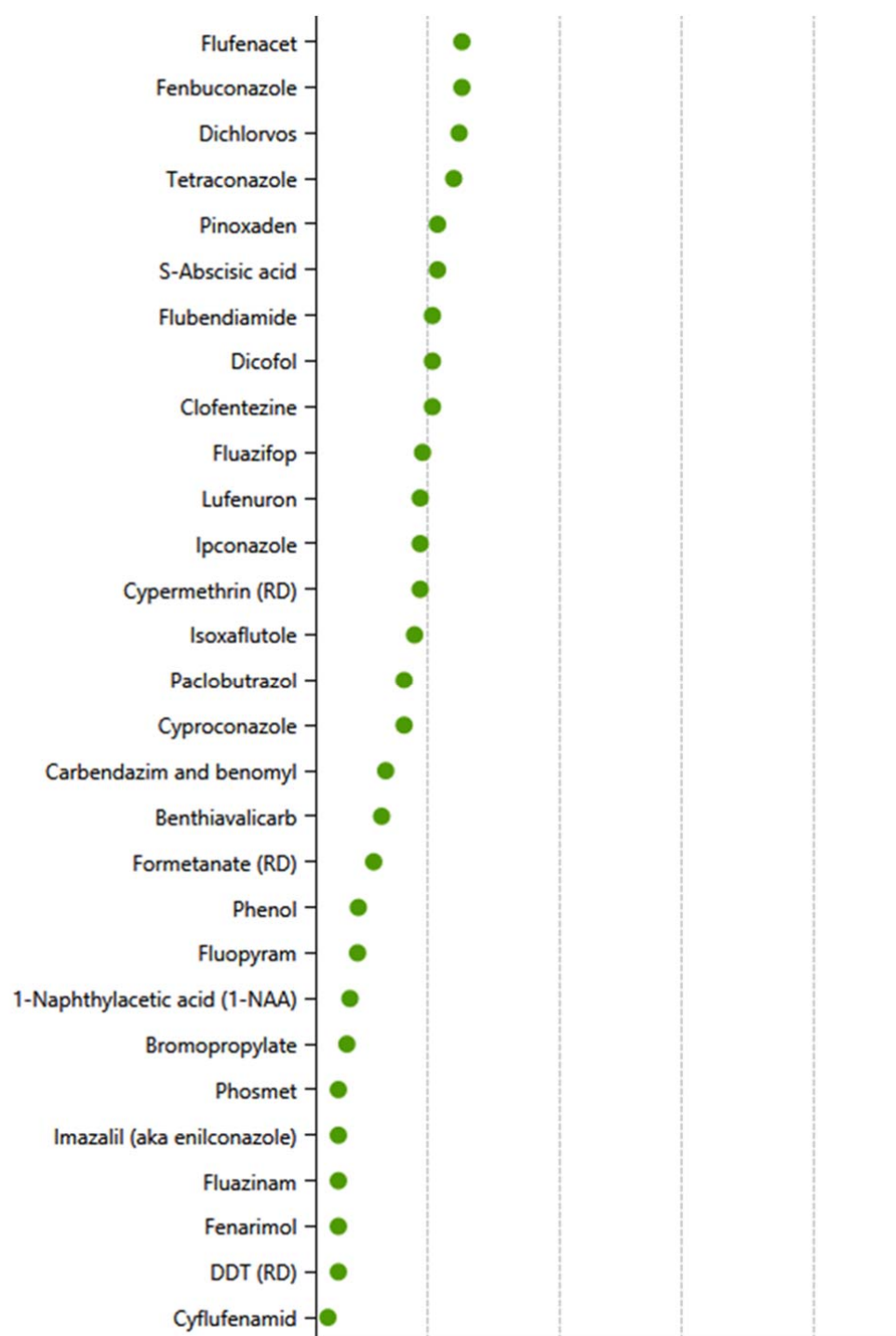


Figure 7: RPFs including imputed RPFs for 50 of the compounds – nominal run

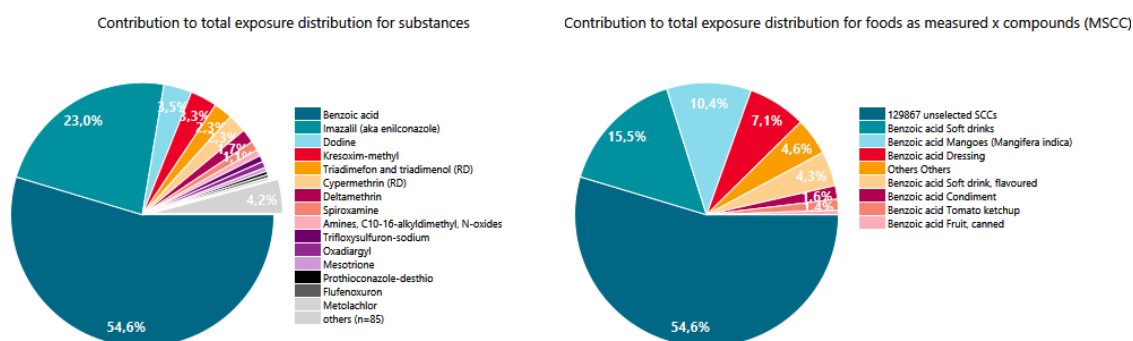


Figure 8: Main risk drivers by substance and by food/substance combination (risk drivers) in total exposure distribution – nominal run

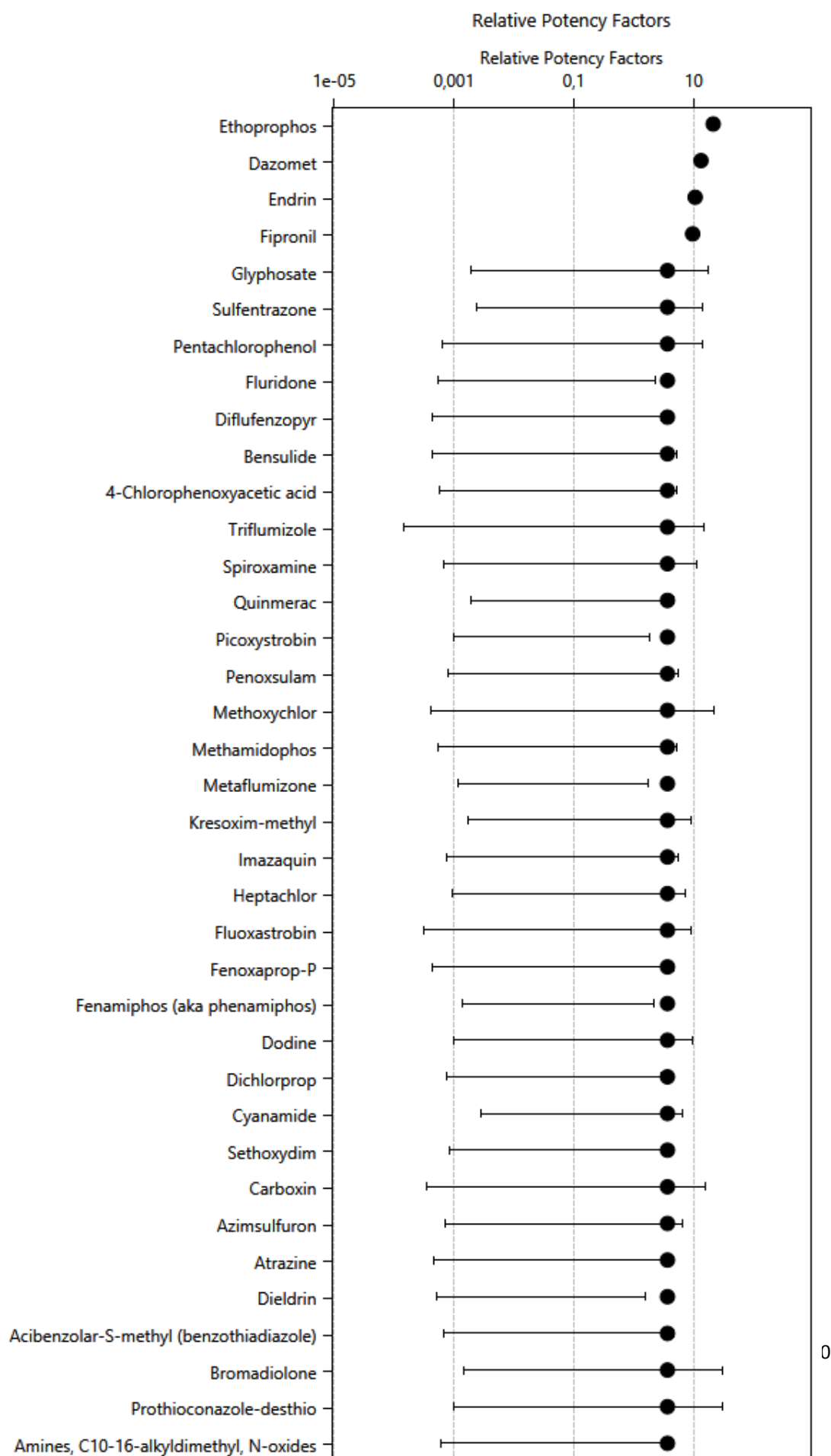
Table 2: OIM percentiles of exposure

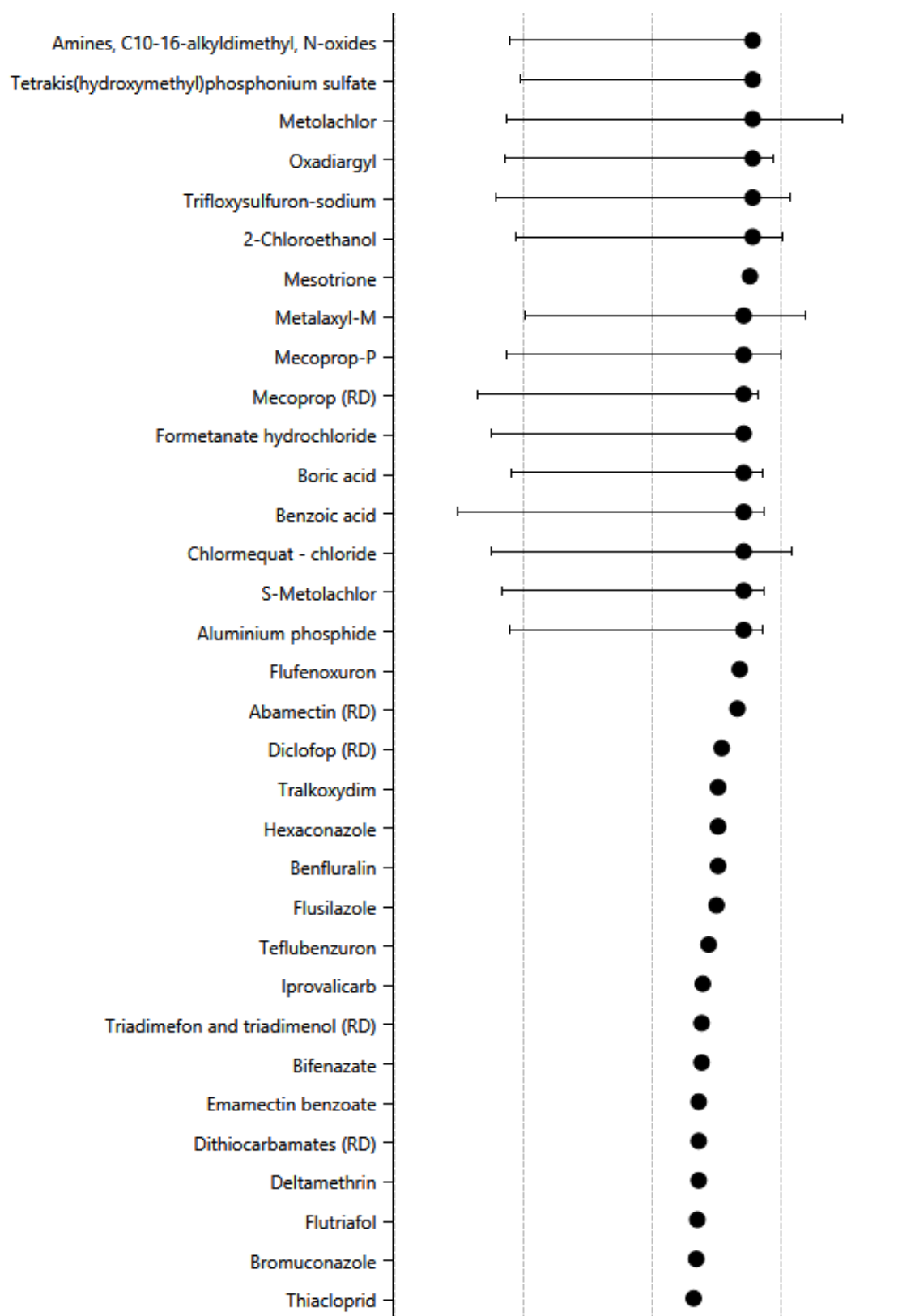
Percentage	Exposure (µg/kg bw/day)
50	0.24
90	0.76
95	1.15
99	3.54
99.9	11.89
99.99	24.93

It is interesting that benzoic acid emerges as a main risk driver, because the probability of CAG membership is below 0.001. By retaining this compound, the other values relevant to the exposure calculation have dramatically increased its influence. For example, the RPF is 2.65, the number of individuals exposed is 1645 (95.6%) and the mean exposure for all individuals is 90.11 µg/kg bw/day. By comparison, the second most influential compound is imazalil, with mean exposure 0.76 µg/kg bw/day, 1724 individuals exposed (100%), RPF = 0.1325, p(CAG) = 0.976. Therefore, the low membership probability of benzoic acid is more than compensated for by the higher exposure values and RPF.

Results with uncertainty:

The *retain* step is to quantify uncertainties about the contributions to risk from individual compounds. The idea is to assess which of these 100 should be targeted in any refinement given the quantified effect of the uncertainties. Therefore, as a more complex run the same options and compounds were used but uncertainty was included with 50 separate realisations. Point of departure information was missing from 50 of these compounds, so these values were imputed and resampled in the calculation of RPFs. Compounds were independently simulated as being included or excluded from each iteration according to their membership probabilities (Figure 6, left panel).





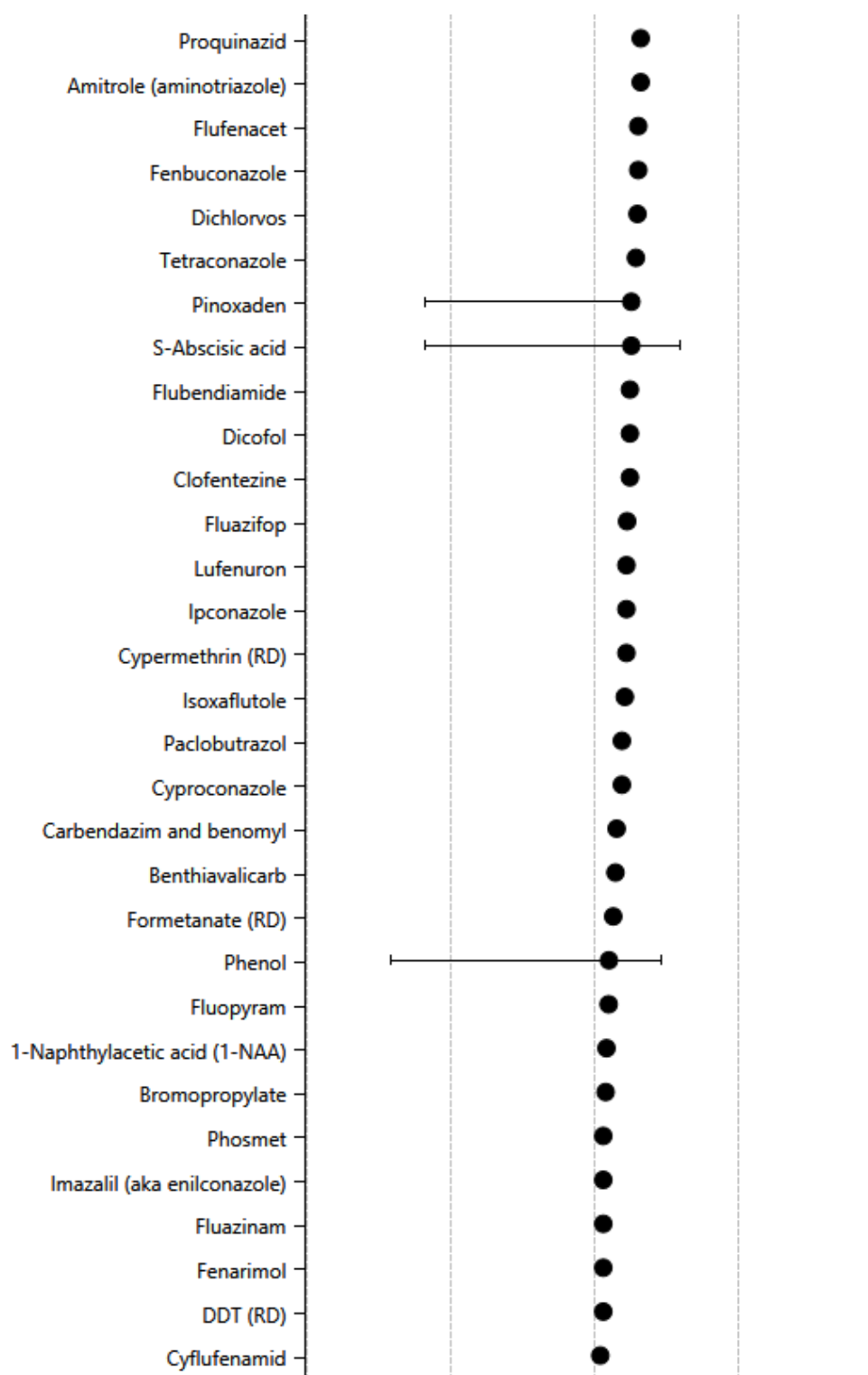


Figure 9: RPFs including imputed RPFs for 50 of the compounds – uncertainty run, including p2.5 and p97.5 uncertainty intervals for each compound

In Figure 9 we see the effect of uncertainties on the simulated RPF values of those 50 compounds with missing hazard data. Note that many of the distributions give probability to lower values than

those imputed in the nominal run (including the RPF value for benzoic acid: nominal value 2.65, interval 9.28E-05 – 5.41). The main risk drivers seen in Figure 10 are almost identical to the nominal run, but this is mainly because these summaries are based on the mean exposures per compound. To investigate the impact of uncertainty per compound, we can examine the output table of exposures. In this example, the compounds with the highest contribution in the nominal run also have the widest uncertainty intervals

Benzoic acid (0 – 70.8)

Imazalil (17.7 – 70.8)

Dodine (0 – 22.1)

Kresoxim-methyl (0 – 18.4)

but there are also cases in which, due to uncertainty, compounds may have higher priority than the nominal run suggests. For example,

Metolachlor: Mean contribution = 0.42 %, p97.5 = 13.26 %

Prothioconazole-desthio: Mean contribution = 0.4 %, p97.5 = 8.16 %

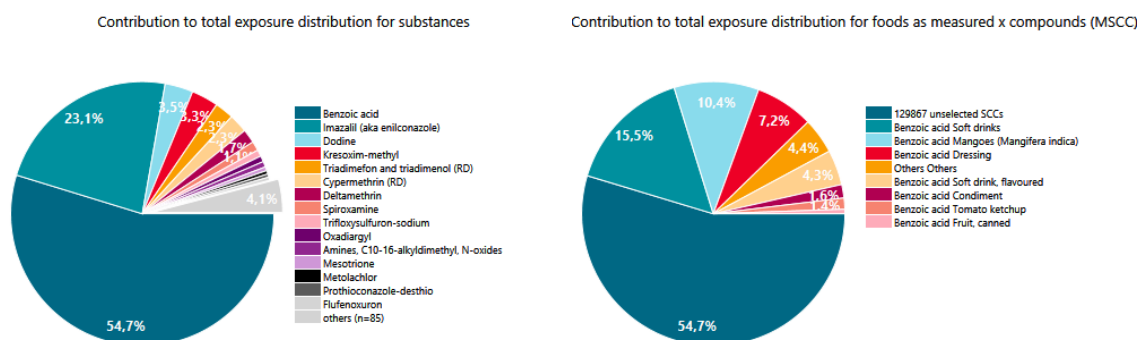
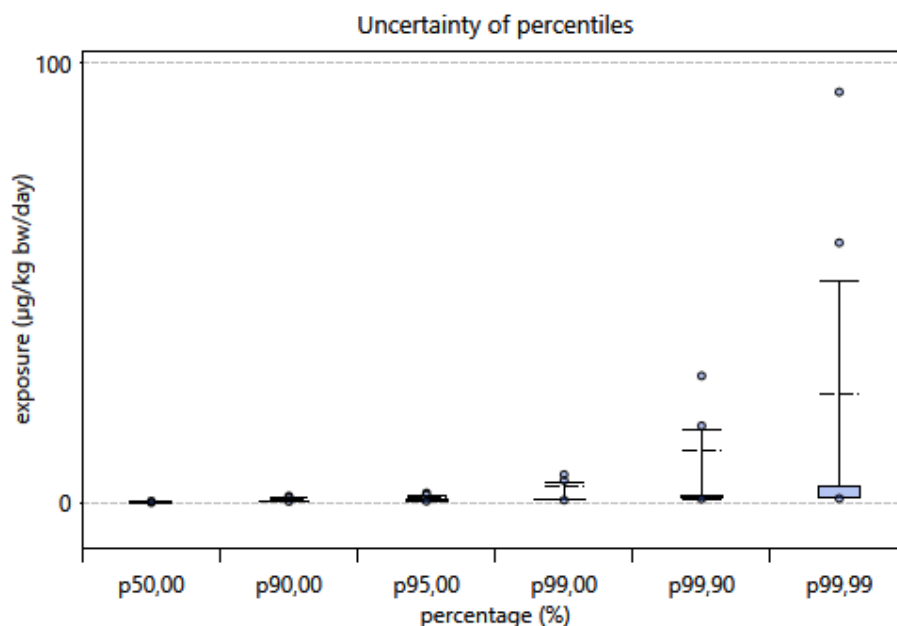


Figure 10: Main risk drivers by substance and by food/substance combination (risk drivers) in total exposure distribution – uncertainty run

The Observed Individual Means model percentiles are shown in Table 3. The mean exposure is 0.425 ($\mu\text{g/kg bw/day}$) and the uncertainty interval for the mean is 0.1365 – 0.58 ($\mu\text{g/kg bw/day}$). From Table 3 we see that for the 99.99% exposure the upper bound uncertainty is around twice the exposure for the nominal run. For other percentiles the uncertainty has more limited impact (see also Figure 11).

Table 3: OIM percentiles of exposure – uncertainty run

Percentage	Exposure (µg/kg bw/day)	Median (p50)	Lower bound (p2,5)	Upper bound (p97,5)
50	0.24	0.10	0.07	0.26
90	0.76	0.39	0.34	1.03
95	1.13	0.49	0.43	1.67
99	3.54	0.82	0.64	4.43
99.9	11.90	1.14	0.93	16.74
99.99	24.94	1.38	1.00	50.39

**Figure 11: Uncertainty in exposure percentiles**

4.2. The Refine step

For a subsequent *refine* step, focus should be on the most likely main contributors to overall risk, taking account of the quantified uncertainty, which in this case is uncertain group membership and uncertain hazard. For the compounds that require refinement, for example, the NOAELs might be replaced with BMD values derived from dose-response modelling, or replacing imputed values with compound-specific values. This involves changing settings in the Target Hazard Doses module and ensuring the necessary data are included in the input files for these substances. In the current example, a more appropriate refinement is to focus on benzoic acid, which appears to be the most important contributor to dietary exposure based on the imputed hazard data.

The following (nominal run, without uncertainty) results were obtained after adding a single line of data to the HazardDoses.xlsx, to specify this new NOAEL information for benzoic acid. The calculated RPF was reduced from 2.65 to 0.00014 which dramatically reduced the relative contribution of benzoic acid to the exposures. Benzoic acid no longer appears in the list of main contributors and the overall exposure estimates are reduced (Table 4).

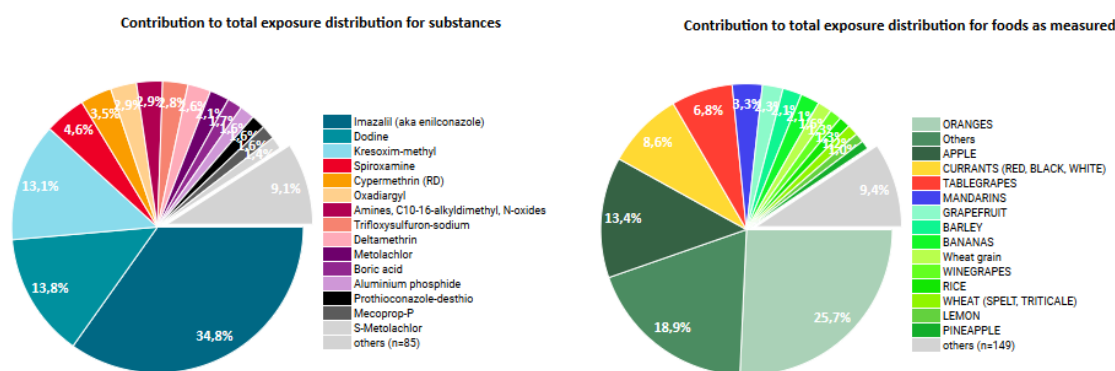


Figure 12. Main risk drivers by substance and by foods (as measured) in total exposure distribution – nominal run

The mean exposure is reduced from 0.426 to 0.282 ($\mu\text{g/kg bw/day}$). The percentiles are also reduced, although this is most evident for the higher percentiles.

Table 4: OIM percentiles of exposure

Percentage	Exposure ($\mu\text{g/kg bw/day}$)
50	0.206
90	0.5945
95	0.7396
99	1.085
99.9	2.164
99.99	5.057

4.3. Sensitivity analysis

As described above, different tiers can be used in the exposure and hazard assessment, and it is often necessary to use default parameters when data are missing. The retain and refine approach might lead to surprising results as illustrated in the worked example of Section 4, where benzoic acid emerged as an important contribution to total risk even though it had such a low probability of CAG membership. When this happens, the factors that lead to high exposure or hazard should be considered more carefully (the refine step) as illustrated in Section 4.2. When there is not such an obvious refinement to pursue, a relatively simple initial approach is to generate some new model runs with plausible alternative default values. This can help identify which of the possible defaults should be targeted first in the refinement. As an example, a prior distribution of 0.252 was used in the Bayesian probabilistic model for CAG membership instead of 0.5. The probability of CAG membership for Benzoic acid was reduced from 0.000975 to 0.000329.

Next, the Munro P5 imputation method for missing hazard data was replaced by the Munro unbiased method (see Section 2.3) so that a central value was used instead of the 5th percentile value within the Cramer class of compounds. For benzoic acid the RPF was reduced from 2.65 to 1.05 as a result. These changes result in a reduced contribution from benzoic acid relative to imazalil (Figure 13). However, benzoic acid remains one of the most important contributors, therefore stronger evidence

should be collected on its likely contribution to the effect, its toxicity or its exposure levels. Reducing the prior probability of group membership further to 0.084 reduced the contribution of benzoic acid further to 6.5% (figures not shown). Even with these alternative default prior options, benzoic acid remains one of the most important contributions to the estimated exposure. As seen in Section 4.2, a more substantial change in the estimation is available when the refined NOAEL is used for benzoic acid.

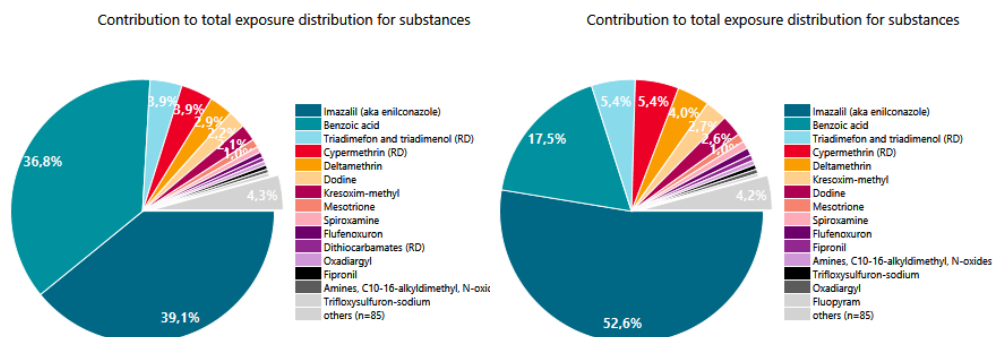


Figure 13. Contribution by substances under alternative model assumptions (left: Munro unbiased imputation; right: Munro unbiased and $p = 0.252$) – nominal run

5. Discussion and Conclusions

The purpose of this report is to highlight the need to quantify uncertainty in any risk assessment associated with chemical mixtures, and to retain as many compounds as possible during the initial analysis, rather than excluding them. If compounds are excluded it is unknown how protective the resulting regulations might be, so where possible a more rigorous approach is proposed.

We have described the concept of retain and refine, in which iterative refinement is targeted on those compounds for which the contribution to the risk may be greatest, when accounting for uncertainty. This is consistent with the tiered approach to risk assessment more generally and can provide a detailed and transparent treatment of uncertainty as proposed in the guidance of EFSA (2018a). As explained throughout, there are many challenges in implementing retain and refine, mainly due to limited information when considering very large sets of substances across different regulatory silos, e.g. PPP, veterinary medicines, environmental contaminants, packaging migrants and consumer products. Some of the practical difficulties associated with retain and refine have been discussed and practical solutions have been suggested, based on experience within the EuroMix project. The worked example illustrates how the EuroMix toolbox implemented in MCRA already includes the necessary functionality to conduct a retain and refine analysis with some of the uncertainties included. The software has been designed to allow for many more modules to include uncertainty, so that retain and refine methods can be developed further in future.

Some of the simple assumptions underlying the model to calculate probabilities of CAG inclusion (Section 2.1) are convenient in practice but may be questioned. Independence of the QSAR models is unlikely because data are difficult to obtain and therefore QSARs are often built using some shared data. (See also Section 4.2 of Rorije *et al*, 2013). The specificity and sensitivity approximations are based on a limited set of training data for compounds known to be real positives/negatives for the steatosis CAG. Sampling uncertainty due to the small sample and questions about whether the compounds really contribute to steatosis are not quantified. Future work could be to investigate

these in more detail if it is possible to obtain more information about real effects for the listed compounds and/or better QSAR models specifically for steatosis. Expert opinion may be included as part of the improved information, but is difficult to arrange in practice. The prior probability of steatosis for each individual compound in our example was selected using a default assumption. It was assigned a fixed value of 0.5. This too should be refined if the method is used for real risk assessment. For example, an alternative value of $79/327 = 0.264$ could be used as this is the proportion of steatotic compounds from an earlier set of tested compounds. Trying alternative plausible values can be used to investigate how robust the final results are to the selected prior, as illustrated in Section 4.3. If it makes a substantial difference, then more effort should be allocated to obtain a prior to reflect the true state of knowledge as accurately as possible. Ongoing work to address questions of this type includes work of EFSA, e.g. EFSA (2018c).

Input and output bias

A related concept that has been considered within Euromix is that bias in input values typically leads to bias in the outputs of interest. In the same way that uncertainty is propagated through the chain of calculations during retain and refine, it could be useful to develop a practical method to estimate the impact of individual sources of input bias on the risk outputs. Some preliminary work has been carried out, but this is not currently implemented in the Euromix toolbox. An example is provided in Appendix I

Uncertainties about how effects combine (CAG levels 3 and 4)

Ideally, the combined effects of chemical mixtures might be modelled mechanistically, i.e. by quantifying the mechanisms of action for multiple chemicals in sufficient detail to predict their combined effect. This is impractical for the number of chemicals that need to be assessed for regulatory purposes.

Dose addition, independent action, response addition, synergism and antagonism are empirical models describing different ways in which chemicals may combine in mixture experiments. As noted earlier, available data suggest that dose addition generally provides a better description of combined effects and, in other cases, is conservative.

There are several possible options for quantifying uncertainty about the combined effect of a group of chemicals, which all cause the same effect but have not been tested as a mixture. One option is to take dose addition as a general model for all mixtures, as proposed by EFSA, but in addition specify a distribution quantifying uncertainty about the extent to which the actual mixture effect will deviate from precise dose addition. In this case, response addition, synergism and antagonism are not modelled explicitly but treated as deviations from dose addition. This approach is simple, requiring only a distribution for deviations from dose addition. A general distribution for deviations might be derived by analysis or expert judgement from data on deviations from dose addition in the literature. However, it is not obvious how this could be combined with mechanistic information, which is also relevant for predicting how chemicals combine.

A second option is to specify a probability quantifying uncertainty about whether dose addition or response addition better describes the combined effect of a group of chemicals, and then specify two separate distributions quantifying uncertainty about deviations from precise dose addition or response addition (again including deviations that might be large enough to be considered as antagonism or synergism). A potential advantage of this option is that it results in a model structure

that aligns with the sources of evidence available to inform the required parameters. Information on similarity of mechanism of action can inform expert judgements of the probability of the mixture effect being better described by dose addition than response addition, while results from historical mixture experiments can be used to inform distributions for deviations from dose or response addition, either by statistical analysis or expert judgement. However, there are two severe obstacles to this approach. First, probabilities are required for every possible combination of the chemicals being considered, which in most cases will amount to an impractically large number of probabilities to assess. Second, the approach requires a specific model for response addition; a common model might be proposed for all combinations, since deviations from it will be quantified, but it may be difficult to decide what that common model should be.

A third option is a combination of the first two: use the first option, dose addition with deviations, for most of the possible combinations of chemicals; and the second option only for those combinations where there is sufficient mechanistic knowledge to specify a plausible alternative model, assess the probability it provides a better description than dose addition, and specify a distribution for deviations from the alternative model². This option has the twin advantages of avoiding the need to assess an impractical number of combinations individually, while enabling mechanistic knowledge to be used where it is informative.

It is suggested that assessors use either the first option described above or, when they have sufficient mechanistic knowledge, the third option (which is a combination of the first and second). How the alternative options would be implemented in cumulative risk assessment calculations, together with possibilities to incorporate mechanistic information when it is available, will be considered in a separate paper (Hart, in prep).

References

- Al-Eryani L, Wahlang B, Falkner KC, Guardiola JJ, Clair HB, Prough RA, Cave M. 2016. Identification of Environmental Chemicals Associated with the Development of Toxicant-associated Fatty Liver Disease in Rodents. *Toxicol Pathol.* 2015 Jun;43(4):482-97. <https://doi.org/10.1177/0192623314549960>
- Benet M1, Moya M, Donato MT, Lahoz A, Hervás D, Guzmán C, Gómez-Lechón MJ, Castell JV, Jover R. 2014. A simple transcriptomic signature able to predict drug-induced hepatic steatosis. *Arch Toxicol.* 2014 Apr;88(4):967-82. <https://doi.org/10.1007/s00204-014-1197-7>
- Bokkers, B., Slob, W., 2007. Deriving a data-based interspecies assessment factor using the NOAEL and the benchmark dose approach. *Critical Reviews in Toxicology* 37, 355–373. <http://dx.doi.org/10.1080/10408440701249224>
- Cramer, G.M., Ford, R.A. and Hall, R.L. 1976. Estimation of toxic hazard—a decision tree approach. *Food and cosmetics toxicology*, 16(3):255–276.
- Crépet, A., Vanacker, M., Sprong, C., de Boer, W., Blaznik, U., Kennedy, M. C., Anagnostopoulos, C., Christodoulou, D., Ruprich, J., Rehurkova, I., van Klaveren, J., Senaeve, D., van der Voet, H., Metruccio, F., Jacxsens, L., Jensen, B. H., Moretto, A., Domingo, J. L., Spanoghe, P. (2018). Selecting mixtures on the basis of dietary exposure and hazard data: application to pesticide exposure in the European population in relation to steatosis. *International journal of hygiene and environmental health*, 222, 291-306.
- Donato MT, Tolosa L, Jiménez N, Castell JV, Gómez-Lechón MJ. 2012. High-content imaging technology for the evaluation of drug-induced steatosis using a multiparametric cell-based assay. *J Biomol Screen.* 2012 Mar;17(3):394-400. <https://doi.org/10.1177/1087057111427586>

² In principle, assessors might specify more than one alternative model and assign probabilities to each of them, however we suggest it is more practical to specify one alternative to dose addition and take account of other possible models within the distributions for deviations from each of those.

- DTU (2012). Identification of Cumulative Assessment Groups of pesticides.
<http://www.efsa.europa.eu/en/supporting/pub/269e.htm>
- EFSA (2008). Scientific Opinion of the Panel on Plant Protection Products and their Residues (PPR Panel) on a request from the EFSA evaluate the suitability of existing methodologies and, if appropriate, the identification of new approaches to assess cumulative and synergistic risks from pesticides to human health with a view to set MRLs for those pesticides in the frame of Regulation (EC) 396/2005. The EFSA Journal (2008) 704, 1-85.
<https://doi.org/10.2903/j.efsa.2008.705>
- EFSA (2012). Panel on Plant Protection Products and their Residues (PPR); Guidance on the Use of Probabilistic Methodology for Modelling Dietary Exposure to Pesticide Residues. EFSA Journal 2012;10(10):2839. [95 pp.] Available online: <https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/j.efsa.2012.2839>
- EFSA (2013). Scientific Opinion on the identification of pesticides to be included in cumulative assessment groups on the basis of their toxicological profile. EFSA Journal 2013;11(7):3293.
- EFSA (2014). Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal 2014;12(6):3734. [278 pp.] Available online: <https://doi.org/10.2903/j.efsa.2014.3734>
- EFSA (2017). Scientific Opinion on the guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal 2017;15(8):4971, 69 pp. <https://doi.org/10.2903/j.efsa.2017.4971>
- EFSA (2018a). EFSA (European Food Safety Authority) Scientific Committee, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Younes M, Craig P, Hart A, Von Goetz N, Koutsoumanis K, Mortensen A, Ossendorp B, Martino L, Merten C, Mosbach-Schulz O and Hardy A. Guidance on Uncertainty Analysis in Scientific Assessments. EFSA Journal 2018;16(1):5123, 39 pp.
<https://doi.org/10.2903/j.efsa.2018.5123>
- EFSA (2018b) EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Younes M, Benfenati E, Castle L, Hougaard Bennekou Z, Laskowski R, Leblanc JC, Kortenkamp A, Ragas A, Posthuma L, Svendsen C, Testai E, Tarazona J, Dujardin B, Kass GEN, Manini P, Dorne JL and Hogstrand C, 2018. Draft guidance on harmonised methodologies for human health, animal health and ecological risk assessment of combined exposure to multiple chemicals. EFSA Journal 20 81pp.
<https://www.efsa.europa.eu/sites/default/files/consultation/consultation/180626-1-ax1.pdf>
- EFSA (2018c). EFSA Draft guidance. Structured EKE used in draft EFSA guidance “Establishment of cumulative assessment groups of pesticides for their effects on the nervous system”
https://www.efsa.europa.eu/sites/default/files/engage/180508_draftreport.pdf [Accessed 15/11/2018]
- IPCS. Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterization. Harmonization Project Document 11 (Second edition). Geneva:World Health Organization. 2017. Available:
http://www.who.int/ipcs/methods/harmonization/areas/hazard_assessment/en/ [accessed 5 October 2018]
- Jennings P, Schwarz M, Landesmann B, Maggioni S, Goumenou M, Bower D, Leonard MO, Wiseman JS. 2014. SEURAT-1 liver gold reference compounds: a mechanism-based review. Arch Toxicol. 2014 Dec;88(12):2099-133.
<https://doi.org/10.1007/s00204-014-1410-8>
- Kennedy, Marc, Hart, Andrew, Roelofs, Victoria, & Roelofs, Willem. (2018). Elicitation Tool 3 (Version v1). Software for expert elicitation. Available online on Zenodo. <https://zenodo.org/record/1243619>
- Kennedy, M.C., Garthwaite, D. G., de Boer, W.J. and Kruisselbrink, J.W. (2019). Modelling aggregate exposure to pesticides from dietary and crop spray sources in UK residents. Environmental Science and Pollution Research, 26(10), 9892-9907. <https://doi.org/10.1007/s11356-019-04440-7>
- Kruisselbrink JW, van der Voet H, van Donkersgoed G, van Klaveren JD, 2018. Proposal for a data model for probabilistic cumulative dietary exposure assessments of pesticides in line with the MCRA software. EFSA supporting publication 2018:EN-1375. 35pp.
- MCRA (2019). Euromix toolbox MCRA 9.0 Reference manual. <https://mcra-test.rivm.nl/EuroMix/WebApp/manual/index.html>.
- Munro, I.C., Ford, R.A., Kennepohl, E. and Sprenger, J.G. (1996). Correlation of structural class with no-observed-effect levels: a proposal for establishing a threshold of concern. Food and Chemical Toxicology, 34(9):829–867.
- Rorije, E. Aldenberg, T., Buist, H., Kroese, D. and Gerrit, S. (2013). The OSIRIS Weight of Evidence approach: ITS for skin sensitisation. Regulatory Toxicology and Pharmacology, 67, 146-156.

- Sarigiannis, D. A., & Hansen, U. (2012). Considering the cumulative risk of mixtures of chemicals - a challenge for policy makers. *Environmental health : a global access science source*, 11 Suppl 1(Suppl 1), S18
<http://dx.doi.org/10.1186/1476-069X-11-S1-S18>
- Tolosa L, Gómez-Lechón MJ, Jiménez N, Hervás D, Jover R, Donato MT. 2016. Advantageous use of HepaRG cells for the screening and mechanistic study of drug-induced steatosis. *Toxicol Appl Pharmacol*. 2016 Jul 1;302:1-9.
<https://doi.org/10.1016/j.taap.2016.04.007>
- van der Voet, H., van der Heijden, G.W.A.M., Bos, P.M.J., Bosgra, S., Boon, P.E., Muri, S.D. and Brüschweiler, B.J. (2009). A model for probabilistic health impact assessment of exposure to food chemicals. *Food and Chemical Toxicology*, 47: 2926-2940. <http://dx.doi.org/10.1016/j.fct.2008.12.027>
- van der Voet H, de Boer WJ, Kruisselbrink JW, Goedhart PW, van der Heijden GWAM, Kennedy MC, Boon PE, van Klaveren JD (2015). The MCRA model for probabilistic single-compound and cumulative risk assessment of pesticides. *Food and Chemical Toxicology*, 79: 5-12. <http://dx.doi.org/10.1016/j.fct.2014.10.014>.

Appendix I. Bias as component of variability and uncertainty

In population risk assessment input data may be biased or unbiased. Biased (conservative) values are often used with the intention to provide an appropriate level of protection. Biased inputs will typically lead to biased outputs (e.g. exposures, health-based guidance values or target margins of exposure).

In qualitative uncertainty assessments it is common to estimate the degree of bias together with the degree of uncertainty in an ordinal way (e.g. between --- and +++). It would be better 1) to disentangle bias and variation, and 2) to provide quantitative estimates.

Alternatively, the uncertainty distribution of data (input or output), can be expressed at an appropriate scale (e.g. logarithmic), by

- 1) the expected bias $\Delta = X - E(X)$
- 2) the distribution F_{θ_u} around $E(X)$, for example a normal distribution $N(0, \sigma_u^2)$

The aim is to estimate the bias of the output data from each calculator based on the biases specified for the input data. This should be simple: in principle a parallel run of the calculator can be made with bias-corrected inputs (faster short-cuts may be possible). Pragmatically, datasets would by default be assumed to be unbiased, but in certain tiers (e.g. TTC, IESTI) automatically labelled to be 'conservative'. For those cases, the risk assessor would be urged (but not obliged) to assess the bias. If no bias is provided for a 'conservative' dataset, the corresponding entity will appear in an uncertainty table as part of the final output of the risk assessment to remind the risk assessor of the unquantified biases.

In practice, the bias may be known or unknown, and it may be related to intended conservativeness or not. Intended conservativeness can relate to

- a) variability, i.e. the wish to protect P% of a population, or
- b) uncertainty, i.e. the wish to be reasonably certain to be on the safe side, or
- c) both variability and uncertainty.

The proposal is to characterise the bias component of uncertainty with the bias estimate itself (at an appropriate scale), and/or additionally with the intended levels of protection (LOPs). Levels of protection may be the LOP_v (level of protection for variability) and/or the LOP_u (level of protection

for uncertainty). As usual in uncertainty assessments, estimates can come from data or from expert opinions.

The LOPs can be converted to estimated bias (or vice versa) if the appropriate distributions are known (or can be estimated). For entities with only variability this is the distribution $F_X(x; \theta_v)$, for entities with only uncertainty it is the distribution $F_X(x; \theta_u)$, for entities with both variability and uncertainty two distributions are relevant, $F_X(x; \theta_v)$ and $F_{\theta_v}(\theta_v; \theta_u)$.

Table 3 shows four examples of entities and the corresponding uncertainty specifications. For simplicity, we assume in all cases that a normal distribution at the log scale is appropriate to characterise the stochastic component of uncertainty.

Table 3. Examples of specifying the bias and variance components of uncertainty.

	value	LOP_v	Δ_v	LOP_u	Δ_u	F_{θ_u}
body weight	70 kg	0.5	0	0.5	0	σ_u^2
(Large Portion) consumption	120 g	0.975	$F_X^{-1}(LOP_v; \theta_v)$	0.5	0	σ_u^2
inter-species factor	10	0.5	0	0.9	$F_X^{-1}(LOP_u; \theta_u)$	σ_u^2
intra-species factor	10	0.95	$F_X^{-1}(LOP_v; \theta_v)$	0.975	$F_{\theta_v}^{-1}(LOP_u; \theta_u)$	σ_u^2

The first example is the mean body weight in a population of interest. The value provided (70 kg) is considered to be an unbiased estimate. Likely the estimate for σ_u^2 will be small.

The second example is the Large Portion estimate of consumption of a food. At least conceptually it is a high percentile (e.g. p97.5) in the population of consumptions. If the distribution of log-consumption would be a normal distribution with SD 0.2, then the bias could be characterised by $\ln(120) - 2 \cdot 0.2 = \ln(80)$, or an overestimation factor $\frac{120}{80} = 1.5$. Likely the estimate for σ_u^2 will again be small.

The third example is the inter-species factor used when a rat study is used for human health risk assessment. In the example the common value of 10 is used. This factor addresses average differences between animals and humans in toxicological sensitivity. Variability is not involved in this concept. Uncertainty has both bias and stochastic components. An example can be found in Bokkers & Slob (2007) and van der Voet et al. (2009). Typically, there is no prior reason to expect different sensitivities between animals and humans apart from allometric scaling, so an unbiased estimate could be a value based on allometric scaling, for example an allometric scaling factor 5.13 based on rat body weight (300 g), human body weight (70 kg) and a power 0.7. For uncertainty a geometric standard deviation (GSD) 2.05 (or, equivalently, $\sigma_u^2 = [\ln(2.05)]^2 = 0.52$) was derived from an assumed GSD = 2 for toxicodynamic and toxicokinetic uncertainty and an uncertainty in the power 0.7 described by a 95% interval 0.65–0.75. The default factor of 10 is then a factor $\frac{10}{5.13} = 1.95$ higher, or $\frac{\ln(10) - \ln(5.13)}{\ln(2.05)} = 1.3$ standard deviations higher than the unbiased value, corresponding to a LOP of around 0.9.

The fourth example is the intra-species factor used to extrapolate from the ‘average human’ to the ‘sensitive human’. In the example this is again set to the common value 10. This example involves both a bias component for variability (the sensitive human w.r.t. the average human) and an

allowance for uncertainty. In Appendix A2 of van der Voet et al. (2009) a proposal was made how this can be modelled, based on a specification of the type: P95-sensitive individuals are between $EF_{\text{intra},p95,p2.5}$ (e.g. 2) and $EF_{\text{intra},p95,p97.5}$ (e.g. 10) times more sensitive than the average human. In this example θ_v would be the variance of the lognormal variability distribution, and θ_u the degrees of freedom of a chi-square distribution describing the uncertainty of θ_v .