Arguments for Considering Uncertainty in QSAR Predictions in Hazard and Risk Assessments

Ullrika Sahlin,[1,2] Laura Golsteijn,[3] M. Sarfraz Iqbal[1] and Willie Peijnenburg[4,5]

[1]Linnaeus University, School of Natural Sciences, Kalmar, Sweden; [2]Lund University, Centre of Environmental and Climate Research, Lund, Sweden; [3]Radboud University Nijmegen, Institute for Water and Wetland Research, Department of Environmental Science, Nijmegen, The Netherlands; [4]RIVM, Laboratory for Ecological Risk Assessment, Bilthoven, The Netherlands; [5]Institute of Environmental Sciences, Leiden University, Leiden, The Netherlands

Summary — Chemical regulation allows non-*in vivo* testing (i.e. *in silico*-derived and *in vitro*-derived) information to replace experimental values from *in vivo* studies in hazard and risk assessments. Although non-*in vitro* testing information on chemical activities or properties is subject to added uncertainty as compared to *in vivo* testing information, this uncertainty is commonly not (fully) taken into account. Considering uncertainty in predictions from quantitative structure–activity relationships (QSARs), which are a form of non-*in vivo* testing information, may improve the way that QSARs support chemical safety assessment under the EU Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) system. We argue that it is useful to consider uncertainty in QSAR predictions, as it: a) supports rational decision-making; b) facilitates cautious risk management; c) informs uncertainty analysis in probabilistic risk assessment; d) may aid the evaluation of QSAR predictions in weight-of-evidence approaches; and e) provides a probabilistic model to verify the experimental data used in risk assessment. The discussion is illustrated by using case studies of QSAR integrated hazard and risk assessment from the EU-financed CADASTER project.

Key words: *decision-making, non-testing information, probabilistic risk assessment, uncertainty analysis.*

Address for correspondence: *Ullrika Sahlin, Lund University, Centre for Environmental and Climate Research, SE 223 62 Lund, Sweden*.
E-mail: *Ullrika.Sahlin@cec.lu.se*

Introduction

Most people agree that better decisions are made when prevailing uncertainties are taken into account. This is especially true for decisions aimed at protecting health and the environment, for which there exist many gaps in the required background knowledge. In this context, the European Union Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) system attempts to use the scarce and scattered information that is available on the majority of substances as efficiently as possible. At the same time, risk management under REACH is urgent, as all the relevant industrial chemicals are due to be assessed before 2018 (1). In order to speed up the risk assessment process and reduce animal testing, part of the REACH strategy to fill knowledge data gaps is based on the optimisation of the use of non-*in vivo* testing information, which can be *in silico*-derived (i.e. obtained by computer-based modelling) or *in vitro*-derived (i.e. from experimental procedures with *in vitro* preparations) information on related compounds (2). *In vivo*-derived information is based on direct empirical observation and provides, compared to non-*in vivo* testing information, more-confident statements on regulatory endpoints. For this reason, the use of non-*in vivo* testing information is to be preceded by a proper validation of its "quality and appropriateness of the intended use" (3). In addition, uncertainty associated with non-animal testing information, which is often ignored, can affect the precision of the estimates produced or the accuracy of the classifications. However, decision makers are sensitive to the degree of uncertainty associated with the assessments based on the available scientific knowledge. In this paper, we argue why considering uncertainty in non-*in vivo* testing information might improve the way in which decision makers use risk assessment to demonstrate the safe use of chemicals.

One of the aims of the EU-funded CADASTER (**CA**se Studies on the **D**evelopment and **A**pplication of *in* ***S****ilico* **T**echniques for **E**nvironmental Hazard and **R**isk Assessment) project has been to integrate quantitative structure–activity relationships (QSARs) within risk assessment. A QSAR is a type of *in silico*-derived, non-*in vivo* testing information that provides information on a chemical's activity, or property, based on analogy reasoning. Such analogy reasoning uses mathematical modelling to link a chemical structure to its physical or

chemical properties, or to a measured biological activity (4, 5). The quantitative nature of a QSAR permits quantitative validation and assessment of uncertainty in the predictions, which can be probabilistic when based on a given probability model (6). The method of risk assessment is, in general, a structured way of describing uncertainty in unknown quantities (7), e.g. the future environmental impact from the use of a chemical, and is probabilistic in nature. One of the tasks within the CADASTER project has been to provide guidance on the integration of QSARs in probabilistic risk assessment (8), and thereby go beyond the use of QSARs as point predictions (5, 9).

The objective of this paper is to discuss why uncertainty associated with a QSAR prediction ought to be considered in hazard and risk assessments to support chemical safety assessment under REACH. In summary, the five arguments supporting the consideration of uncertainty in QSAR predictions discussed here are that:

a) it can assist in the evaluation of management strategies, in terms of a more balanced comparison of benefits and losses resulting from chemical regulation, by using uncertainty analysis instead of conservative safety factors (thus allowing the decision maker to be more risk-neutral);

b) it can be used to generate more-conservative (i.e. safer) hazard and risk assessments (compared to not considering uncertainty);

c) it can have an impact on decisions, which can be shown by uncertainty and sensitivity analysis;

d) it can provide useful additional information on the quality of a QSAR prediction in a weight-of-evidence (WoE) approach; and

e) it can aid the verification of experimental results, by providing a probabilistic model for their evaluation.

Firstly, we provide an introduction to the context and characteristics of the type of uncertainty that we are considering. Subsequently, we discuss the arguments as to why uncertainty in QSAR predictions ought to be considered. The arguments are illustrated with case studies from the CADASTER project.

Framing and the Context of Uncertainty

Uncertainty in a QSAR prediction is defined here as the added uncertainty associated with non-*in vivo* testing (i.e. *in vitro*/*in silico*-derived) information compared to *in vivo* testing-derived information. We acknowledge that there is also uncertainty associated with *in vivo* testing information, but that is not a reason to disregard uncertainty in non-*in vivo* information. Consider an assessment input parameter of either a point estimate based on, for example, an *in vivo* experimental test, or a prediction from a QSAR. The level of complexity that is required to ascertain uncertainty associated with a QSAR prediction is determined by how uncertainty in the test-based estimate is dealt with. Consider a situation where *in vivo* testing information is provided by a point estimate that corresponds to the most likely endpoints, or expected value. The practice, in this particular risk assessment, is then to disregard the magnitude of other sources of uncertainty, such as measurements errors, variability in experimental method or endpoint variability. Similarly, these sources of uncertainty then do not have to be considered for non-*in vivo* testing information. When variability is considered in *in vivo* testing information, for example in experimental data, this should also be reflected in the treatment of non-*in vivo* testing information. In the latter case, expert judgement can be employed to add variability to a QSAR prediction, since nowadays few QSARs actually model variability (but see Tebby and Mombelli [10]). In any case, there will always be an error associated with a QSAR prediction, because a prediction is derived from a simplification of reality (called a model), and models are always more or less 'imperfect'. Uncertainty in a QSAR prediction is, in this context, knowledge-based uncertainty that can be reduced by performing the experimental test that will remove the error associated with a model-based prediction and increase the confidence in the resulting assessments.

This paper focuses on uncertainty in a QSAR prediction from the perspective of the user (here a risk assessor or a person who performs chemical ranking or regulatory decisions). The risk assessor is obliged to follow principles to deal with uncertainty that were set up by the regulatory framework for risk management, which can involve cautious assessment or rules for information requirements. Within this framework, it is the risk assessor's uncertainty that is to be considered. Uncertainty associated with a QSAR prediction is both qualitative and quantitative, referred to as predictive reliability and predictive error, respectively (6). Uncertainty linked to the reliability of a model that is used for prediction can be described as a measure of extrapolation that can assist the risk assessor in judging the level of confidence in the use of the model for that purpose. The risk assessor's judgement of uncertainty associated with the predictive error can, for QSARs, be informed by predictive inference based on the QSAR data and a probability model of observed and not-yet-observed quantities (6). The advantage of QSARs compared to other non-*in vivo* testing

data is that the predictive error can be assessed by principles of statistical inference, thereby offering a quantitative and transparent treatment of uncertainty.

Case Studies

The discussion will be illustrated with three case studies of benzotriazoles (BTAZs), as examples of QSAR integrated chemical safety assessments performed in the CADASTER project. These studies cover QSAR integrated hazard and risk assessment and the validation of experimental (*in vivo*-derived) data, and are described in Appendices 1–3.

A QSAR integrated hazard assessment was performed to identify compounds classified as very toxic (Appendix 1). This case study demonstrates the impact on classification and, in particular, on the rate of type II errors (i.e. the failure to regulate a dangerous chemical), following different ways of considering uncertainty in QSAR predictions in the input, and attitudes to express uncertainty in the output of hazard assessment.

Non-safe emissions are those that cause the Environmental Concentration (EC) to exceed the No-Effect Concentration (NEC). QSAR integrated fate and effect assessments were made to find the Predicted EC (PEC) and Predicted NEC (PNEC) of eight BTAZs (Appendix 2). Uncertainty in these prediction values was expressed as probability distributions around PEC and PNEC. When there is strong evidence that the EC will be larger than the NEC, the decision is to apply risk management to reduce emissions, or refine the assessment to improve the evidence. We refer to these actions as 'to regulate'. In this case study, several QSARs were used to assess PECs and PNECs for risk assessment. Uncertainty in QSAR predictions was characterised by the predictive distribution following statistical inference or expert judgement. This case study demonstrates the effect of considering QSAR uncertainty in risk evaluation.

The last case study demonstrates a way of employing uncertainty in QSAR predictions for the validation of new experimental data (Appendix 3). In this case, the same QSARs as in the hazard assessment case study were used to validate experimental data on aquatic toxicity for previously untested BTAZs.

Arguments for Considering Uncertainty in QSAR Predictions

Rational decision-making

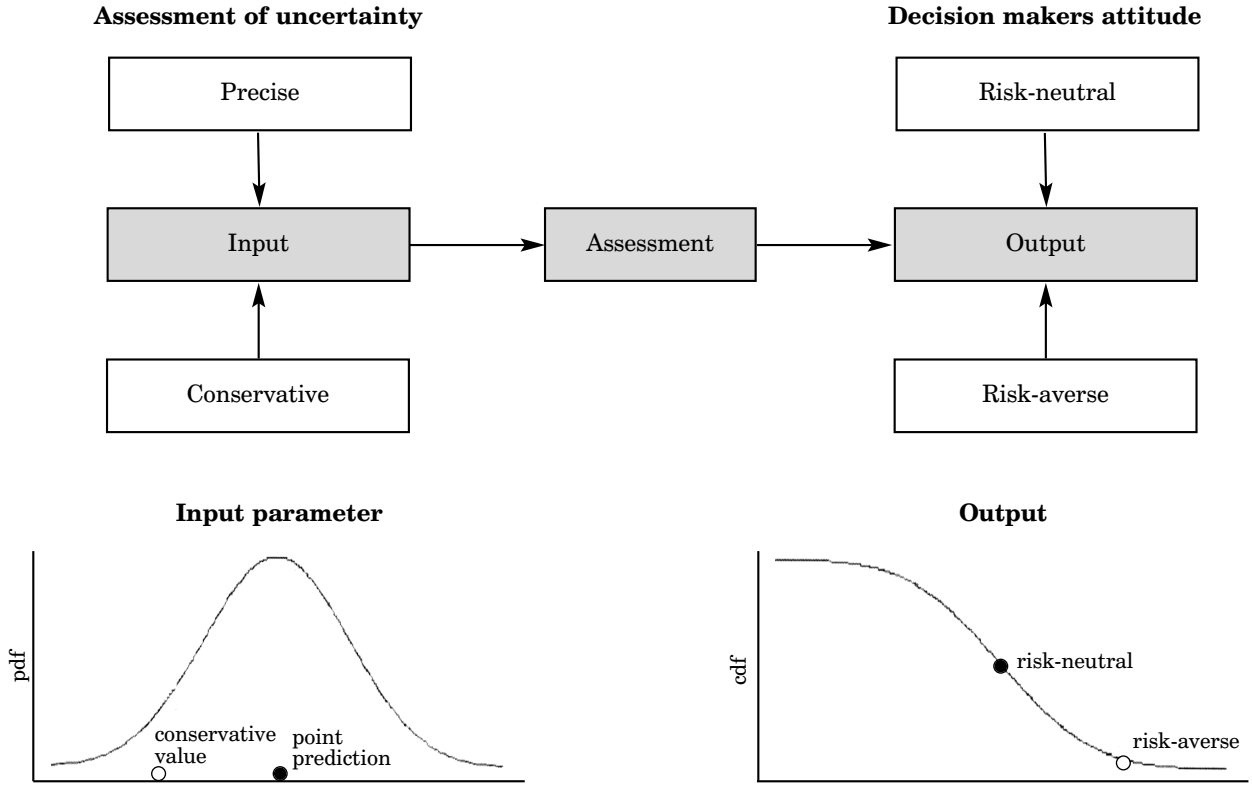
The decision maker's behaviour and attitude toward uncertainty and risk can be described or prescribed by decision rules. A common rule of rational decision-making is to maximise the expected utility, where utility expresses how much a decision maker values a future outcome. Consideration is given to the range of possible values of utility, derived from an assessment outcome for which uncertainty is characterised by a probability distribution. The range of likely values for utility depends on the inherent variation in the underlying processes that lead to the assessed outcome (variability), or on the variation that follows from our uncertainty in the characteristics of this process. The utility linked to the production and use of chemicals ought to decrease with increasing damage to health and the environment, and ought to increase when society benefits from the use of the chemicals. Utility in risk assessment is commonly the opposite of loss, whereas net benefit can be used as utility in a cost–benefit analysis. A risk-neutral decision maker is indifferent to a certain and an uncertain alternative with the same expected utility (Figure 1). If choosing between two chemicals, where chemical A is of medium risk with little uncertainty and chemical B is of low risk with high uncertainty, a risk-neutral decision maker would choose chemical B. In practice, however, it has been noticed that decision makers are often risk averse — they tend to dislike the more uncertain alternative, if everything else remains equal. When a conservative value on risk is higher

Table 1: The decision matrix for the simplified probabilistic chemical safety assessment

Decision	Outcome	Likelihood	Utility
Regulate: P (PEC > PNEC) > 0.05	EC > NEC, correct EC < NEC, type I error	P (correct) P (type I error)	No loss Missed opportunity loss or increased damage from substitute
No concern: P (PEC < PNEC) > 0.05	EC > NEC, type II error EC < NEC, correct	P (type II error) P (correct)	Damage to environment and health No loss

Decision alternatives, outcomes and their likelihood, and the factors influencing utilities are shown.
PEC = Predicted Environmental Concentration; PNEC = Predicted No-Effect Concentration.

Figure 1: The process of rational decision-making



Points at which to consider uncertainty in QSAR predictions are found in the specification of the input or in the response to the output of an assessment. Input can either be given a full probabilistic characterisation (precise; illustrated by the probability density function [pdf] of a continuous quantitative input parameter) or a conservative point value. Decision makers can be insensitive to uncertainty in the outcome of an assessment (risk-neutral; illustrated by the mean or median from the cumulative distribution function [cdf] of the assessment output) or dislike more uncertain outcomes (risk-averse; illustrated by an extreme value from the output cdf). Attitudes toward uncertainty can only be fully supported by precise characterisation (i.e. full probabilistic distribution) of uncertainty in input.

for chemical A, this means that the risk assessor might choose chemical A, since he/she wants to avoid the possibility that chemical B might be, in fact, a higher most-likely risk than chemical A. Thus, rational decision making is guided by utility and the decision makers' attitudes toward the uncertainty in utility. It should be noted that utility is not supposed to be a function of uncertainty *per se*. Rational decision making based on expectations is possible, if uncertainty reflects the likelihood of all possible outcomes (a precise characterisation is shown in Figure 1).

It is possible to distinguish the different types of errors that can be made in chemical safety regulation (see Table 1). Regulating a chemical when in fact it is not a risk — a type I error — creates losses due to missed opportunities or the increased use of a more dangerous substitute. The consequences of failing to regulate a dangerous chemical — a type II error — are associated with increased losses due to damage to the environment and health. In general, type II errors are less favoured than type I errors. Let us consider that the utility function represents the losses associated with different outcomes in Table 1. Since uncertainty in QSAR predictions might have an influence on the likelihood of different outcomes, and thereby on the expected utility, it should be considered in hazard and risk assessments (*Argument I*).

Let us consider a rational decision maker who is risk-neutral, i.e. is seeking to maximise the expected utility. This leads to a situation where risk and uncertainty are both reduced to an expected value. Does this imply that it is enough to provide a QSAR prediction as a point value? In some cases it does, as long as the influence of QSAR uncertainty does not affect the expected utility. However, in certain situations — for exam-

ple, when the distribution describing uncertainty in a parameter is highly skewed, or when there is a complex relation between the input parameter and the output of the assessment — that consideration may not hold. The removal of possible extreme values (or events) from the characterisation of QSAR prediction, and thereby from the assessment, might have a large influence on the final estimate of risk and expected utility. Uncertainty in a QSAR prediction used as input parameter in fate or effect assessment models may be important for the expected utility, when the most extreme values of the predictive distribution have a large influence on the final outcome.

Conservative risk assessment

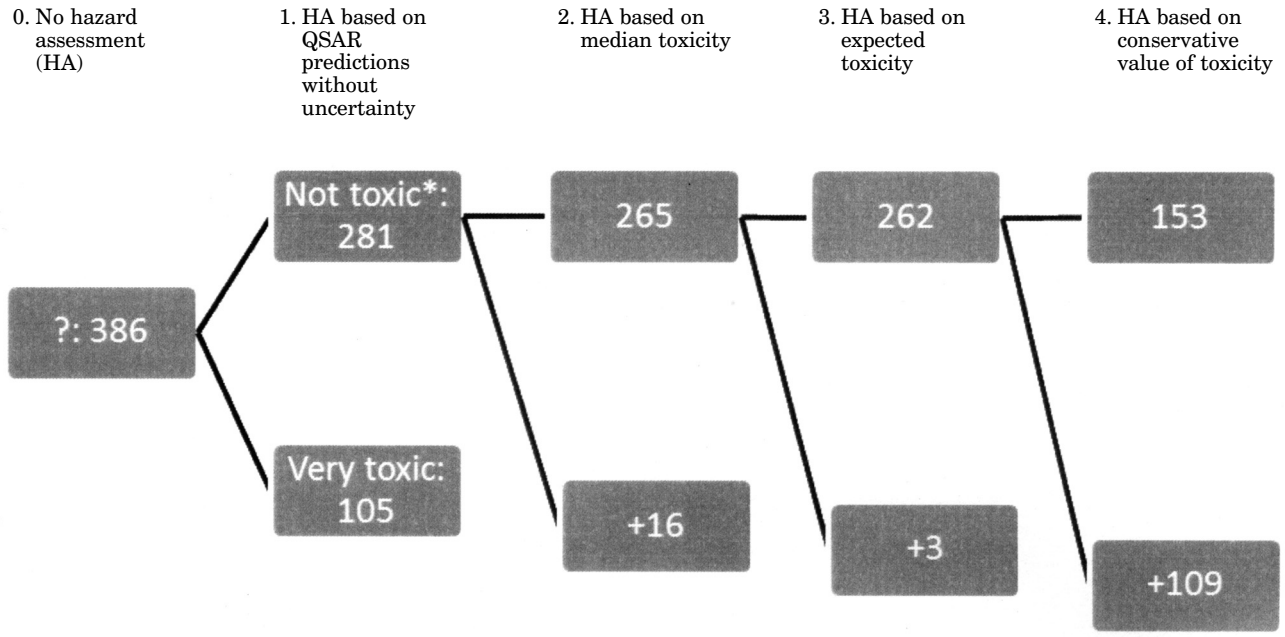
In environmental decision making, it is generally complicated to base decisions on expected utility, as utility is difficult to specify when effects are irreversible and the risk to humans and the environment is high. Limitations in the background knowledge for hazard and risk assessments make it difficult to produce precise estimates of the likelihood of different outcomes, and thereby, of risk. These difficulties have led to the emergence of conservatism in risk management, which is expressed as a higher acceptance of erring on the safe side. In practice, the decision to regulate is based on the probability of PEC being larger than PNEC, given a critical threshold ($p > 0.05$ in this case), without considering the likelihood of different outcomes associated with alternative decisions. Risk is said to be unacceptable when the probability of a given chemical to be a risk is considerable, e.g. if p (PEC > PNEC) > 0.05 (8; Table 1). QSAR uncertainty might still have an impact on the probability of PEC exceeding PNEC, but there is a need to find a proper balance between being cautious and being too conservative.

Uncertainty and variability in chemical safety assessment are typically treated conservatively through assumptions, defaults or safety factors (11; Scenario 3 in Figure 1). Table A4.1 (Appendix 4) highlights conservative assessments, such as a high probability of PNEC being lower than NEC. The difference between a conservative PNEC and the actual NEC value increases with the level of caution. Erring on the safe side results in a higher probability of making a correct risk classification, at the cost of a type I error. The gain is a lower probability of type II errors. Setting regulatory emission levels based on conservative risk assessments results in lower permissible emission levels, and thereby reduced benefits to society from the use of the chemicals (a type I error). This means that assigning conservative PNECs values is the same as giving higher weight to losses associated with type II errors compared to type I errors. Conservatism inserts a bias toward regulation in the face of uncertainty. Conservative estimates of risk may overestimate the benefit of risk reduction (11), and force the decision maker to deviate from risk neutrality, as a precise estimate of the risk is not given and the maximum expected utility is unknown (12).

A practical problem resides in where and how a level of caution should be implemented into the assessment (Figure 1). The case study on the QSAR integrated hazard assessment (Appendix 1) showed that considering QSAR uncertainty results in a lower rate of type II errors (Figure 2). The approach to use the worst conservative, but plausible, point estimates on parameter values may seem advantageous from a practical point of view, because it is easy to communicate and simplify calculations in assessments. However, employing conservative values of input parameters can be dangerous, and may cause unwanted bias in the assessment of risk (Appendix 1). Conservative estimates of risk are to be based on uncertainty as close as possible to the final output of an assessment. We cannot evaluate the influence or the kind of influence that a conservative value of an input parameter might have on the final output. Reducing the information on uncertainty before the assessment seems inefficient as compared to uncertainty analysis in probabilistic risk assessment, i.e. when uncertainty is quantified by probabilities. Furthermore, the combination of several independent conservatively assigned input parameters may lead to compounding or 'cascading' conservatism (11).

Cautious risk management uses safety factors to compensate for data with less confidence (13). For example, it is recommended (but is not a requirement) that when an obtained value on aquatic toxicity is provided by a QSAR instead of an *in vivo* test, the estimated value is divided by a safety factor of ten (14). Safety factors may result in overly conservative assessment, or uncontrolled deviations from the desired attitude toward risk in decision making. The level of caution is a matter of risk management. The price of over-regulation also needs to be considered, as it forces decision makers to become overly (and perhaps unwillingly) risk-averse. A too-conservative assessment of risk makes it difficult to balance other societal concerns in any socioeconomic analysis related to the use of chemicals. The reverse could also occur, as conservative assumptions on input parameters may lead to unknown underestimates of risk, which cause unknown deviations from risk aversion. A full characterisation of uncertainty in QSAR predictions by a probability distribution may therefore be advantageous, even under a cautious risk management (*Argument II*). In practice, the assessment problem consists of knowing when to apply safety factors and when to apply probability distributions to consider uncertainty in an input parameter.

Figure 2: The hazard assessment of BTAZs compounds under the different treatments of QSAR uncertainty



The number of BTAZs compounds classified as very toxic or not (including potentially) toxic under the different treatments of QSAR uncertainty, both in the input and in the output of the assessment. Uncertainty in QSAR predictions is considered in hazard assessment alternatives 2 to 4.*

Uncertainty analysis

The uncertainty in chemical safety assessment has four tiers of treatment (15): point estimates with conservative assumptions (tier 0); qualitative, which indicates ranges for the unquantifiable uncertainties (tier 1); deterministic (tier 2), which re-evaluates by considering worst-case assumptions; and probabilistic (tier 3), which quantifies uncertainty through probabilities. Risk assessment involves the specification of values for parameters, either for direct determination of the exposure or effect, or as input for mechanistic, empirical or distribution-based models. Uncertainty in a QSAR prediction is an example of parameter uncertainty in probabilistic risk assessment, which, according to the European Chemicals Agency (ECHA), is the uncertainty involved in the specification of numerical values.

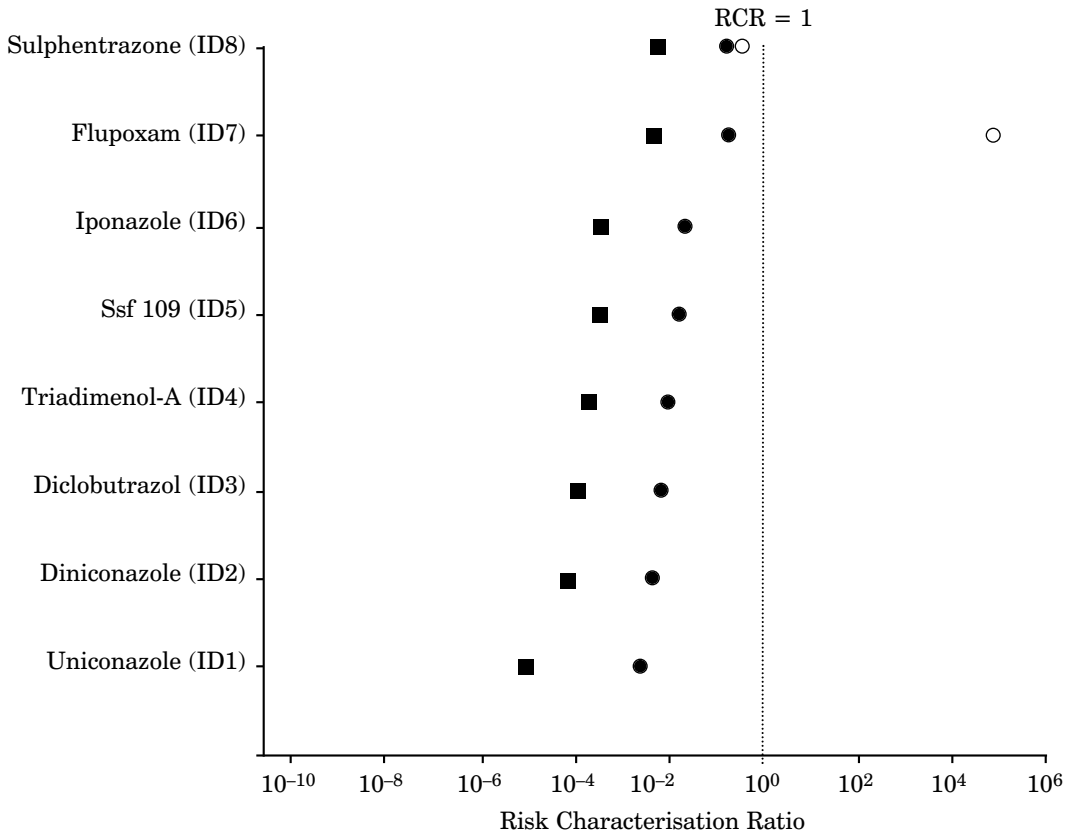
The purpose of an uncertainty analysis is to highlight the sources of uncertainty with the greatest impact on the analysis (i.e. sensitivity analysis), and to indicate to the regulator the degree of seriousness of the hazard under consideration (15). A technique that can be used for probabilistic analysis is Monte Carlo simulation. In such a simulation, uncertainty is propagated by drawing random samples from probability distributions that specify uncertainty in input parameters and variability. Therefore, in a tier 3 uncertainty analysis, the uncertainty in all input parameters should be assessed, including the uncertainty in input values predicted by QSARs (*Argument III*).

It is important to understand how to interpret the results from a sensitivity analysis. In general, a sensitivity analysis reveals the influence from one source of uncertainty, but sensitivity is also dependent on the other sources of uncertainty and their magnitude. For example, when there is only one uncertain input parameter, it will influence the overall uncertainty by 100%. The addition of more uncertain parameters reduces the relative contribution from each uncertain parameter. The influence of a QSAR prediction is not only related to the accuracy of the prediction itself, but also depends on how the uncertainty in the prediction propagates in the assessment model. The latter depends on the number of times the parameter is used and whether it reduces or increases the assessed risk. Even though a sensitivity analysis may show, for instance, that the influence of uncertainty in a QSAR prediction is small, it is not possible to know this beforehand, without doing the full uncertainty analysis. The case study on QSAR integrated risk assessment (Appendix 2) compares

the risk assessed by deterministic to probabilistic levels (Figure 3).

As pointed out earlier, uncertainty in QSAR predictions is of two kinds: qualitative, as in a statement of how reliable we think a prediction is; and quantitative, which is a probability distribution describing the error in a prediction (6). A practical question is how to treat predictive reliability in uncertainty analysis. Predictive reliability can be dealt with by 'flagging' (i.e. including it as a statement in the risk report, but using the QSAR prediction as it is), by incorporating other non-*in vivo* testing information (perhaps in combination with the QSAR prediction), or by allowing it to be reflected in the parameter uncertainty followed by sensitivity analysis. Ahlers *et al.* (2) suggest that when the amount of information gathered remains somewhat below the standard requirements, it might be preferable to increase the uncertainty factor instead of performing a missing test. If the higher safety factor subsequently results in no apparent risk, further testing can be avoided and animals can be saved. For example, when experimental EC50 values are available for *Daphnia* and algae, and a QSAR estimate is available for fish, if the PEC/PNEC ratio is much less than one, then a fish test may be redundant. However, a chronic fish test should be considered, if the PEC/PNEC ratio is close to, or exceeds, one (2). The same line of reasoning could be applied to QSAR predictions for which the predictive reliability is evaluated as low. Regulators are cautious over accepting predictions outside a model's applicability domain (AD), but the determination of where to make the distinction between what is in and out, and whether other information is needed, might be context-dependent. In the QSAR integrated risk assessment case study, uncertainty due to extrapolations close to

Figure 3: Risk characterisation ratio of eight BTAZs



$$\textit{Risk characterisation ratio} = \frac{\textit{Predicted Environmental Concentration}}{\textit{Predicted No-Effect Concentration}}$$

Risk Characterisation Ratio (RCR) under deterministic (■; most likely) and probabilistic (●; 95% percentile) risk assessments of eight BTAZs showing the influence by considering uncertainty in input parameters. A re-assessment by using enlarged uncertainty in unreliable QSAR predictions (○) was made for compounds 7 and 8. Compound ID are the same as in Table A2.1.

the border of the AD is dealt with by enlarging the parameter uncertainty by some arbitrary uncertainty factor. We refer to this as an extended uncertainty analysis. Sensitivity analysis can then be used to evaluate the impact of a QSAR prediction and thereby aid in the judgement of whether or not to use it. The increase in uncertainty when a compound is considered to be out of the QSAR's AD should reflect our lower confidence in the prediction compared to if it had been inside the AD. In the case study described in Appendix 2, we have enlarged the standard deviation in the predictive distribution by a factor of two. We show that taking predictive reliability into consideration may have an impact on the evaluated risk (Figure 3).

The sensitivity of the assessment of different input parameters, model assumptions or scenarios, can be evaluated in different ways. It can be an evaluation of the contribution of the magnitude of the uncertainty, such as the variance or the width of the predictive distribution, in an input parameter to the uncertainty in the assessment output. It can be evaluated on a decision, e.g. if reducing uncertainty would lead to another decision. For example, it can be demonstrated that increasing (or reducing) uncertainty alters the regulatory decision, e.g. by moving a critical value such as the 95th percentile of the PEC/PNEC ratio over a decision threshold. Alternatively, sensitivity analysis can be performed, to evaluate the net benefit of reducing uncertainty, which can be used to decide whether further testing is needed. Consider a toxicity classification based on non-*in vivo* testing information, e.g. QSARs (16). The chance of achieving a correct classification is increased by performing experimental *in vivo* tests, under the assumption that this would give more-precise estimates of toxicity. The benefit of testing, also known as the value of information, can be estimated as the difference in the increase in expected utility and the cost of testing. A positive value of information means that reducing uncertainty, and thereby the probabilities of committing type I and II errors, has a high influence on the decision.

Quality assurance in weight-of-evidence

In the arguments mentioned so far, we have focused on uncertainty in the input values of a hazard or risk assessment. In this case, QSAR predictions do not differ substantially from other kinds of predictions or other uncertain parameters. In our penultimate argument (*Argument IV*), we consider whether uncertainty in predictions might improve the usefulness of QSARs in WoE approaches under REACH. WoE approaches are used under REACH to fill gaps in missing information and to avoid unnecessary *in vivo* testing (17). A WoE approach considers the quality (strength and weaknesses) of different sources of information, and uses that to reach a consensus on the chemical properties of a substance. Thus, a WoE approach seeks to make use of all the available information for an endpoint, where the information from each source might be insufficient to support decision making. QSARs are frequently used as pieces of information on physicochemical properties and ecotoxicological effects (17).

A prediction based on a QSAR of established scientific validity, with a response that fits with the regulatory purpose, and for which the substance to be assessed lies inside the QSAR's AD, fulfils the information requirements under REACH, and can be used as replacement for missing *in vivo* data in hazard and risk assessment (R.6.1.10.1 in Reference 18). When the information is not enough to address all of these points, the QSAR prediction can be used as part of a WoE approach. For example, scientific validity is hard to vindicate for QSARs built on small QSAR data sets, when there has been no external validation, or without a mechanistic understanding of the structure–activity relationship.

The assessment of the uncertainty associated with a prediction is helpful in the evaluation of the quality of a QSAR prediction, which increases the usefulness of QSAR predictions in WoE, since more information is extracted from the background knowledge. The requirements relating to the quality of information depend on the importance of the decision taken, and on the consequences, if the decision is wrong due to inaccurate predictions (compare type I and II errors above). The magnitude of the error in a prediction provides useful information for someone who is planning to evaluate the quality of a WoE approach. After all, the WoE approach ought to use all the available information, including the assessed uncertainty associated with a prediction. A single value on an error may be of less value, as it has to be related to some reference or to its impact on the decision made, preferably as demonstrated through sensitivity analysis. The use of a QSAR prediction on the border of the model's AD can be encouraged, if it can be shown, by uncertainty and sensitivity analysis, that the uncertainty in the QSAR prediction does not significantly influence the final decision. Given a clear motivation as to why a QSAR prediction adequately describes the REACH endpoint of concern, further information on that particular endpoint may not be necessary (14, 18). For example, even though some of the BTAZs in the case study on QSAR integrated hazard assessment were assessed on the basis of QSAR predictions with relatively low confidence, the assessment can be judged as acceptable, since all the compounds were classified as very toxic and the lower confidence in individual QSAR predictions does cause type II errors (Figure A1.1).

Validation of *in vivo* experimental data

Our final argument concerns a situation in which a QSAR prediction is used to validate an experimental *in vivo*-derived value that is used as input in a hazard or risk assessment. The verification of such experimental input data is carried out by regulators to check assessments submitted to dossiers, or by industry to provide more support for their assessments. When uncertainty in a QSAR prediction is not considered, a validation is based on a difference between two point values — the QSAR prediction and the corresponding *in vivo*-derived estimate. Without any reference to what is a large or small error, we are left with a difficult judgement. The predictive distribution of a QSAR can be used as a probabilistic model to test the experimental value (*Argument V*). Then the assumption is that the experimental value is drawn from the same statistical population as the compounds in the QSAR training data set. When we believe the QSAR prediction to be true, an experimental value can be verified by testing whether it lies inside a credible interval based on the resulting predictive distribution for that compound. A 95% Bayesian confidence interval is the range where we expect 95% of the true values to be found. If the experimental value falls outside the confidence interval, the test suggests that the confidence in the experimental value is lower than 5%. In the third case study, as part of the CADASTER project, experimental values for BTAZs were found to be valid at a confidence level of 95% (Appendix 3).

Conclusions

QSAR predictions are subject to added uncertainty, as compared to estimates based on experimental *in vivo*-derived data. Note that we use the word 'added', since there are other relevant sources of uncertainty in experimental *in vivo* data as well. The discussion on the need to consider uncertainty in QSAR predictions has been based on five arguments: a) rational decision-making; b) conservative risk management; c) uncertainty analysis; d) quality evaluation of a QSAR prediction in a WoE approach; and e) validation of experimental (i.e. *in vivo*-derived) data used in a risk assessment.

It is relatively easy to think of reasons not to consider, and thereby not to assess, uncertainty in QSAR predictions. For instance, it may increase the already high workload of risk assessors and regulators, and may involve more training. The treatment of uncertainty is context-dependent and is framed by attitudes to risk and caution. Thus 'cookbook' recipes for its integration are difficult to provide. The uncertainty associated with QSAR predictions is not available in the majority of databases and tools for QSAR-use that are available today. The CADASTER web tool (www.cadaster.eu) has been developed to support qualitative and quantitative characterisations of uncertainty in QSAR predictions, in terms of measures related to AD and assessments of predictive errors. The use of web tools is, in general, limited when it comes to providing the full extent of uncertainty. After all, the final characterisation of uncertainty in QSAR predictions that are used in risk assessment involves subjective judgement and is context-specific.

This paper is an attempt to counter the absence of uncertainty consideration in the implementation of QSARs in probabilistic risk assessment. Here we suggest that allowing for uncertainty in QSAR predictions may influence the outcome of rational decision-making, if it may alter the expected utility. In the case of a risk-averse decision maker, more-uncertain outcomes are less favoured, and taking uncertainty in QSAR predictions into account can alter the order of prioritisation based on hazard or risk assessment. Cautious risk management calls for uncertainty in QSAR predictions to be considered, for instance, by the worst plausible limits as the full probability distribution in an uncertainty analysis. Uncertainty analysis (15) offers a means of considering and propagating uncertainty that is quantified by probability distributions related to the magnitude of the error in a prediction. Uncertainty in QSAR predictions can prevent risk assessors from being too conservative by informing or replacing the use of safety factors associated with non-testing information. The evaluation of the quality of a QSAR prediction in a WoE approach is better informed when the associated uncertainty is presented, and is an efficient way of enhancing the use of the available knowledge. Finally, uncertainty in a QSAR prediction, as quantified by a probability distribution, may be used as a probability model to test the plausibility of an experimentally based estimate used in a risk assessment.

The case studies provided in this paper demonstrate the integration of QSARs in probabilistic risk assessment. Uncertainty associated with a QSAR prediction can be considered at different points in the production of hazard and risk assessments (Figure 1). QSAR uncertainty tells us something about the input in an assessment, while decisions are made on the output of an assessment. The process where uncertainty is considered can be related to the decision maker's attitude toward uncertainty or related to the way uncertainty is characterised in an input parameter. It is less straightforward to allow risk assessment to consider the qualitative characteristic of uncertainty that occurs when the suitability of QSARs for predicting a compound is questionable, e.g. when the compound is on the border of the model's AD. Here, we briefly point out a possible way of considering

qualitative uncertainty through sensitivity analysis. Further insight on how this can be achieved in practice could be gained from examples showing the impact from uncertainty in QSAR predictions on decision-making in chemical safety assessment.

Acknowledgements

The authors are grateful to Mike Comber for useful comments. This study was funded by the European Seventh Framework Programme through the CADASTER project FP7-ENV-2007-1-212668.

References

1. European Parliament (2006). *Regulation (EC) No 1907/2006* of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending *Directive 1999/45/EC* and repealing *Council Regulation (EEC) No 793/93* and *Commission Regulation (EC) No 1488/94* as well as *Council Directive 76/769/EEC* and *Commission Directives 91/155/EEC*, *93/67/EEC*, *93/105/EC* and *2000/21/EC*. *Official Journal of the European Union* **L396**, 30.10.06, 1–849.
2. Ahlers, J., Stock, F., & Werschkun, B. (2008). Integrated testing and intelligent assessment — new challenges under REACH. *Environmental Science & Pollution Research* **15**, 565–572.
3. Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H., Watts, C.D. & Worth, A.P. (2003). Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environmental Health Perspectives* **111**, 1376–1390.
4. Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T., McDowell, R.M. & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives* **111**, 1361–1375.
5. Walker, J.D., Carlsen, L. & Jaworska, J. (2003). Improving opportunities for regulatory acceptance of QSARs: The importance of model domain, uncertainty, validity and predictability. *QSAR & Combinatorial Science* **22**, 346–350.
6. Sahlin, U. (2013). Uncertainty in QSAR predictions. *ATLA* **41**, 111–125.
7. Aven, T. (2010). Some reflections on uncertainty analysis and management. *Reliability Engineering & System Safety* **95**, 195–201.
8. Jager, T., Vermeire, T.G., Rikken, M.G.J. & Van Der Poel, P. (2001). Opportunities for a probabilistic risk assessment of chemicals in the European Union. *Chemosphere* **43**, 257–264.
9. Sahlin, U., Filipsson, M. & Oberg, T. (2011). A risk assessment perspective of current practice in characterizing uncertainties in QSAR regression predictions. *Molecular Informatics* **30**, 551–564.
10. Tebby, C. & Mombelli, E. (2012). A kernel-based method for assessing uncertainty on individual QSAR predictions. *Molecular Informatics* **31**, 741–751.
11. Viscusi, W.K., Hamilton, J.T. & Dockins, P.C. (1997). Conservative *versus* mean risk assessments: Implications for superfund policies. *Journal of Environmental Economics & Management* **34**, 187–206.
12. Hansson, S.O. (1999). Adjusting scientific practices to the precautionary principle. *Human & Ecological Risk Assessment* **5**, 909–921.
13. Edler, L. (2009). Quantification of uncertainty within and between species and the role of uncertainty factors. *Human & Experimental Toxicology* **28**, 115–117.
14. ECHA (2008). *Guidance on Information Requirements and Chemical Safety Assessment*. Helsinki, Finland: European Chemicals Agency. Available at: http://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment (Accessed 17.01.13).
15. ECHA (2012). *Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.19: Uncertainty Analysis*, 36pp. Helsinki, Finland: European Chemicals Agency.
16. Jaworska, J., Gabbert, S. & Aldenberg, T. (2010). Towards optimization of chemical testing under REACH: A Bayesian network approach to Integrated Testing Strategies. *Regulatory Toxicology & Pharmacology* **57**, 157–167.
17. ECHA (2010). *Practical Guide 2: How to Report Weight of Evidence,* 21pp. Helsinki, Finland: European Chemicals Agency. Available at: http://echa.europa.eu/documents/10162/13655/pg_report_weight_of_evidence_en.pdf (Accessed 17.01.13).
18. ECHA (2009). *Practical Guide 5: How to Report (Q)SARs,* 11pp. Helsinki, Finland: European Chemicals Agency. Available at: http://echa.europa.eu/documents/10162/13655/pg_report_qsars_en.pdf (Accessed 17.01.13).

Appendix 1

Case study 1: The impact of uncertainty in QSAR integrated hazard assessment

A QSAR integrated hazard assessment was built for benzotriazoles (BTAZs) to illustrate the QSAR integrated chemical safety assessment performed in the CADASTER project. A compound was classified as potentially toxic or not by comparing a derived hazardous concentration in an aquatic environment to a predefined threshold. An assessment was based on QSAR predictions of aquatic toxicity in three species: an alga, *Daphnia* and a fish. The QSARs are described in (1). First we identified the lowest effect concentration where 50% of the individuals of the most sensitive population (among those tested) are affected, i.e. the EC50. Following the recommendations of REACH (2), we classified a compound as 'very toxic' to the aquatic environment if the EC50 value of the most sensitive (evaluated) species, min{EC50}, was less than 1mg/L.

QSAR predictions were derived for 386 BTAZs based on consensus modelling. With the aim of illustrating the effect of considering uncertainty in QSAR predictions, we based hazard assessments on QSAR predictions, either without uncertainty (i.e. point predictions) or with uncertainty (i.e. by a probability distribution for the error). The underlying QSARs predicted point predictions only. Uncertainty in QSAR predictions was characterised as a normal distribution (symmetric bell-shaped curve) with the point prediction as its mean and the predictive error as its standard deviation. Predictive error was assigned the root mean squared error (RMSE) value derived for the training QSAR data sets. The RMSE was chosen as a good approximation of the average predictive error for compounds that are in the model's applicability domain (AD). This approach to assess the predictive distribution in a QSAR prediction can be categorised as expert judgement informed by statistical measures. Predictions for all the compounds were derived from the three QSARs, even though some of them fell out of one or more AD. Compounds with the least reliable predictions were identified by the maximum absolute difference (MAD) between individual predictions in the consensus modelling. Compounds for which at least one MAD score among the three QSAR predictions were larger than 0.9, were judged to have been given less reliable assessment.

The hazard assessments based on QSAR predictions with uncertainty were performed by Monte Carlo simulation, where random samples were derived from the corresponding predictive distributions and, in each iteration, the min{EC50} value was stored. The resulting uncertainty in aquatic toxicity (i.e. min{EC50} values) is described by a probability distribution. From this probability distribution, we calculated the expected value, the median and the 5th percentile. A best guess or most likely value of aquatic toxicity can be provided by the median and the expected value. The median does not consider extreme values, while the expected value weights all possible values with their likelihood of occurring. Differences in expected and median values are found for skewed distributions with the presence of either high or low extreme events. When the interspecies variability in sensitivity is relatively larger than the uncertainty in individual QSAR predictions, the uncertainty in the min{EC50} value will be dominated by the uncertainty in the most sensitive species. In the QSAR models provided here, these predictions have a symmetric distribution. When interspecies variability is small in comparison to QSAR uncertainty, the minimum out of three values should result in a skewed distribution. In this case study, the differences between median and expected values were negligible, meaning that the uncertainty in the classification variable was rather symmetric.

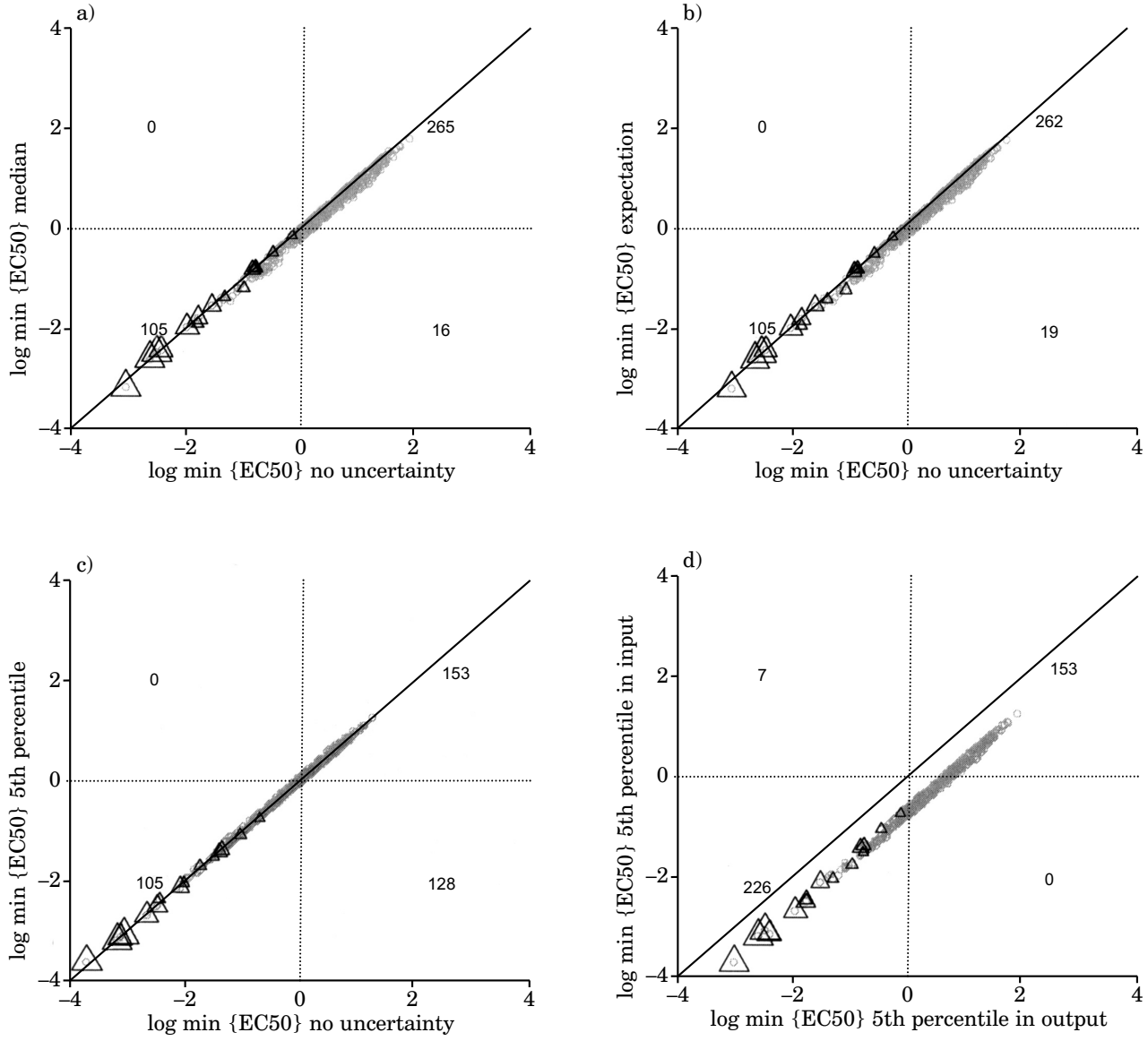
Thus, uncertainty in the output of the hazard assessment was considered in four ways: first, without considering QSAR uncertainty which gave only one value on the min{EC50}, then as the expected, the median and the 5th percentile of the distribution describing uncertainty in the assessment quantity min{EC50} derived from considering uncertainty in QSAR predictions.

A list of 385 BTAZs were classified as potentially toxic or not, based on the four different ways to consider uncertainty. We calculated the number of compounds for which the consideration of uncertainty resulted in different classifications (Figure A1.1). In particular, we were interested in the number of compounds for which the toxicity classification changed from not toxic to potentially toxic, when considering uncertainty in QSAR predictions, or when taking a more risk-averse attitude (Figure A1.2).

Consideration of QSAR uncertainty resulted in more-cautious classifications (Figure A1.2) and an avoidance of making errors of type II. Of 386 compounds, 19 were classified as toxic after QSAR uncertainty in input had been taken into account. Adding risk-averse behaviour, an additional 115 compounds were classified as potentially toxic.

We found that the use of conservative values for QSAR predictions (5th percentile) as input to the

Figure A1.1: A comparison of the aquatic toxicity of BTAZs derived by QSAR integrated hazard assessment under different types of uncertainty



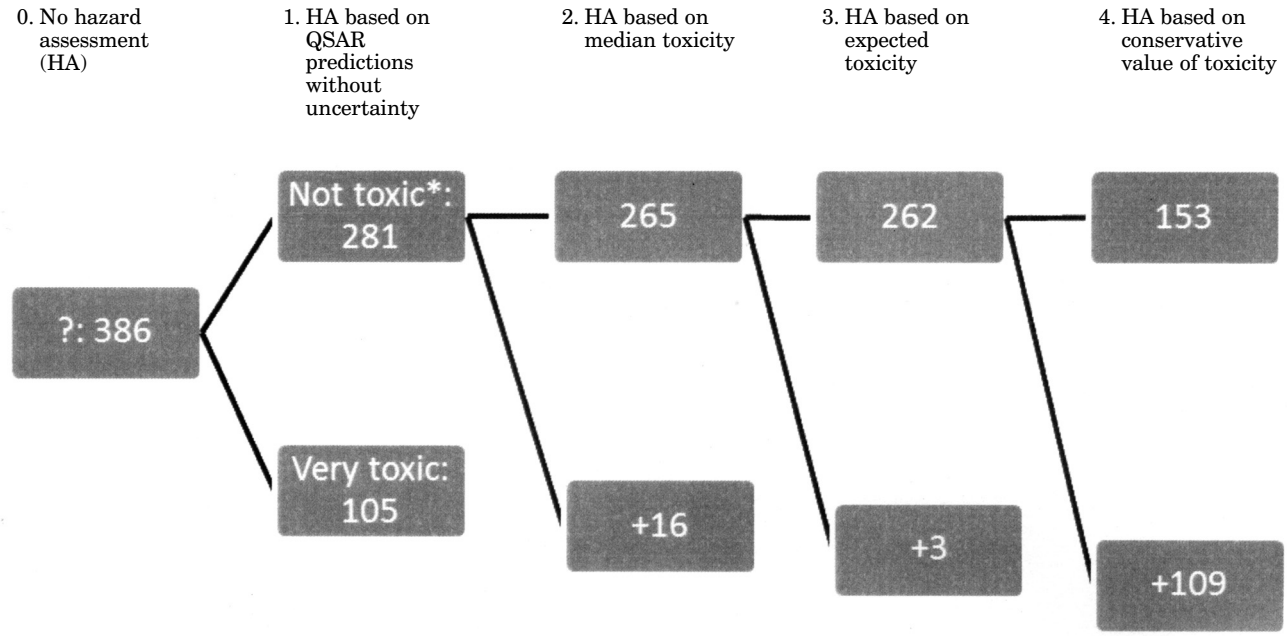
..... = *classification cut-off;* ——— = *1:1 line;* ○ = *compound;* △ = *low predictive reliability.*

A comparison of the aquatic toxicity of 385 BTAZs derived by QSAR integrated hazard assessment under different types of uncertainty in QSAR predictions and attitudes to uncertainty in the assessment output. Compounds with less reliable predictions were identified by the maximum absolute difference between individual predictions in QSAR consensus modelling.

hazard assessment resulted in an increased probability of making type II errors compared to classifications based on the 5th percentile of the output (Figure A1.1 d).

To conclude, the impact on decision-making of considering uncertainty in QSAR predictions was a reduced probability of making type II errors. This effect was found when the full predictive distribution was used as the input. Reducing the information on predictive uncertainty to a conservative value in the input into the hazard assessment did not automatically lead to more-conservative classifications. Seven compounds were classified as non-toxic under the conservative assessment, which were classified as potentially toxic based on hazard assessment with probabilistic uncertainty analysis and risk-averse behaviour. The use of conservative values to specify input increases the probability of

Figure A1.2: The classification of BTAZs under the different treatments of QSAR uncertainty both in the input and in the output of the assessment



The number of BTAZs compounds classified as very toxic or not (including potentially) toxic under the different treatments of QSAR uncertainty both in the input and in the output of the assessment. Uncertainty in QSAR predictions is considered in hazard assessment alternatives 2 to 4.*

committing errors of type II, hinders the decision-maker in being risk-neutral, and forces the decision-maker to be risk-averse to an unknown degree.

References

1. Cassini, S., Kovarich, S., Papa, E., Roy, P.P., Rahmberg, M., Nilsson, S., Sahlin, U., Jeliazkova, N., Kochev, N., Pukalov, O., Tetko, I., Brandmaier, S., Durjava, M.K., Kolar, B. Peijnenburg, W. & Gramatica, P. (2013). Evaluation of CADASTER QSAR models for aquatic toxicity of (benzo)triazoles and prioritisation by consensus. *ATLA* **41**, 49–64.
2. European Commission (1991). *Directive 67/548/EEC* (Annex VI). General Classification and Labelling Requirements for Dangerous Substances and Preparations. *Official Journal of the European Union* **L180**, 09.07.1991, 1–79.

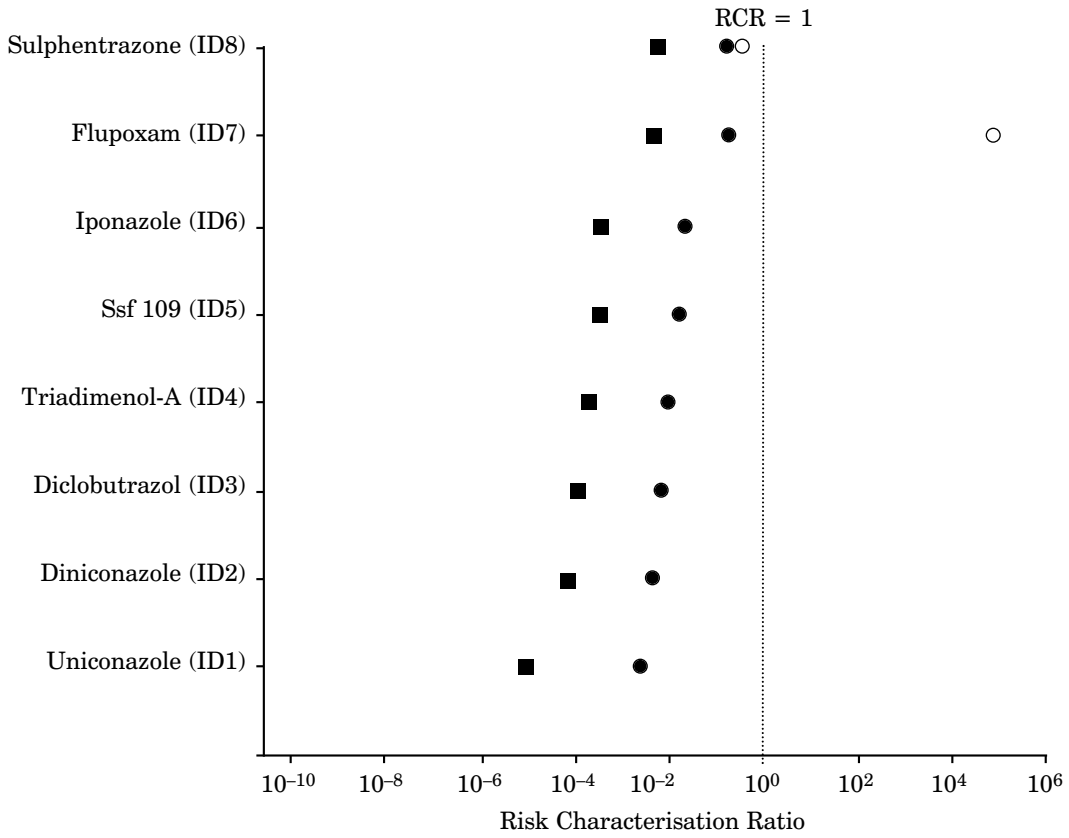
Appendix 2

Case study 2: Uncertainty analysis in QSAR integrated risk assessment

A non-safe emission occurs when the concentration in the environment, the Environmental Concentration (EC), exceeds the No-Effect Concentration (NEC). Fate and effect assessments were made to find the Predicted EC (PEC) and Predicted NEC (PNEC) of eight BTAZs (Table A2.1). Uncertainty in the value of these quantities was expressed as probability distributions around the ratio between PNEC and PEC. When there is strong evidence that EC will be larger than NEC, the decision is to apply risk management to reduce emissions, or to refine the assessment to improve the evidence. We refer to these actions as 'to regulate'. The case study used several QSARs to assess PEC and PNEC for risk assessment. The purpose of the case study was to demonstrate the integration of QSARs into probabilistic risk assessment, and was therefore based on a hypothetical common emission scenario for all evaluated compounds and did not consider all relevant non-QSAR sources of uncertainty.

The case study was based on the QSAR integrated risk assessment described in (1). Fate assessments (PEC) were made by using SimpleBox

Figure A2.1: Risk characterisation ratio of eight BTAZs



$$\textit{Risk characterisation ratio} = \frac{\textit{Predicted Environmental Concentration}}{\textit{Predicted No-Effect Concentration}}$$

Risk Characterisation Ratio (RCR) under deterministic (■; most likely) and probabilistic (●; 95th percentile) risk assessments of eight BTAZs showing the influence by considering uncertainty in input parameters. A re-assessment by using enlarged uncertainty in unreliable QSAR predictions (○) was made for compounds 7 and 8. Compound ID numbers are the same as in Table A2.1.

Table A2.1: Uncertainty analysis of risk characterisation ratio (RCR) given a common emission scenario based on assessments of eight BTAZs

Name CAS No.	Compound ID1 Uniconazole 083657-17-4	Compound ID2 Diniconazole 083657-24-3	Compound ID3 Diclobutrazol 075736-33-3	Compound ID4 Triadimenol-A 089482-17-7	Compound ID5 Ssf 109 129586-32-9	Compound ID6 Iponazole 125225-28-7	Compound ID7 Flupoxam 119126-15-7	Compound ID8 Sulphentrazone 122836-35-5
Original uncertainty analysis:								
Mean	–6.67	–4.17	–3.98	–3.69	–3.49	–3.52	–2.55	–2.17
5%	–0.86	–0.61	–0.44	–0.32	–0.10	0.03	0.97	0.88
25%	–2.24	–1.63	–1.45	–1.21	–1.01	–0.98	0.15	0.12
50%	–3.32	–2.43	–2.24	–1.97	–1.77	–1.77	–0.61	–0.51
75%	–4.47	–3.26	–3.07	–2.76	–2.58	–2.59	–1.45	–1.24
95%	–6.23	–4.41	–4.23	–3.92	–3.73	–3.76	–2.69	–41.00
Extended uncertainty analysis:								
Mean	–6.67	–4.17	–3.98	–3.69	–3.49	–3.52	–4.64	–2.17
5%	–0.86	–0.61	–0.44	–0.32	–0.10	0.03	8.66	0.88
25%	–2.24	–1.63	–1.45	–1.21	–1.01	–0.98	2.39	0.12
50%	–3.32	–2.43	–2.24	–1.97	–1.77	–1.77	0.51	–0.51
75%	–4.47	–3.26	–3.07	–2.76	–2.58	–2.59	–2.66	–1.24
95%	–6.23	–4.41	–4.23	–3.92	–3.73	–3.76	–4.97	–2.41

Uncertainty analysis of Risk Characterisation Ratio (RCR) given a common emission scenario based on assessments of eight BTAZs. The extended analysis is made to evaluate the influence of non-reliable QSAR predictions, which are found for compounds 7 and 8.

under a unit emission rate. Effect assessments derived PNEC as the minimum out of QSAR predicted EC50 values on three aquatic species divided by an assessment factor of 1000. The level of emission was here adjusted to 500kg/day, in order to obtain compounds classified as safe and as risky.

The assessment was done on two levels of complexity in the consideration of uncertainty. The output from a deterministic risk assessment, where uncertainty is not considered, results in the most likely values of PEC and PNEC. Here, all eight compounds had a risk characterisation ratio (RCR) below one, which would indicate that they are safe (Figure A2.1). In an attempt to avoid making erroneous risk classifications, probabilistic assessments were performed to consider uncertainty (tier 3). The probabilistic evaluation of risk showed that the compounds were still safe, since the probability of PEC exceeding PNEC was more than 5% (Figure A2.1). Considering the sources of uncertainty does, in general, lead to safer decisions, and in that respect, uncertainty from QSAR predictions is no exception.

Two of the compounds (ID 2 and 7) were judged as being on the borderline of at least one QSAR model used for input to the assessment (see Peijnenburg *et al.* [1]). In order to evaluate the influence of these lower confidence QSAR predictions, we made a reassessment of risk where the corresponding predictive distributions had been enlarged by an arbitrary factor (the standard deviation in the predictive distribution had been multiplied by 10). This resulted in increased risk (Figure A2.1). The sensitivity analysis showed that the risk classification of compound ID 7 was sensitive to the lower confidence of the QSAR predictions, as the 95th percentile of the RCR changed from being less than, to larger than, one. In this situation, the risk assessment for compound ID 7 might need to include other sources of background information to achieve the same quality as the others, and even then we might still be unsure whether or not the assessment was acceptable.

Reference

1. Peijnenburg, W., Kos Durjava, M., Gramatica, P., Papa, E., Tetko, I. & Sahlin, U. (2013). *CADASTER Deliverable 4.6 Synthesis of Research Findings and Recommendations for Prioritization*, 168pp. Available at: http://www.cadaster.eu/sites/cadaster.eu/files/data/deliverable/public/Deliverable4.6.pdf (Accessed 04.02.13).

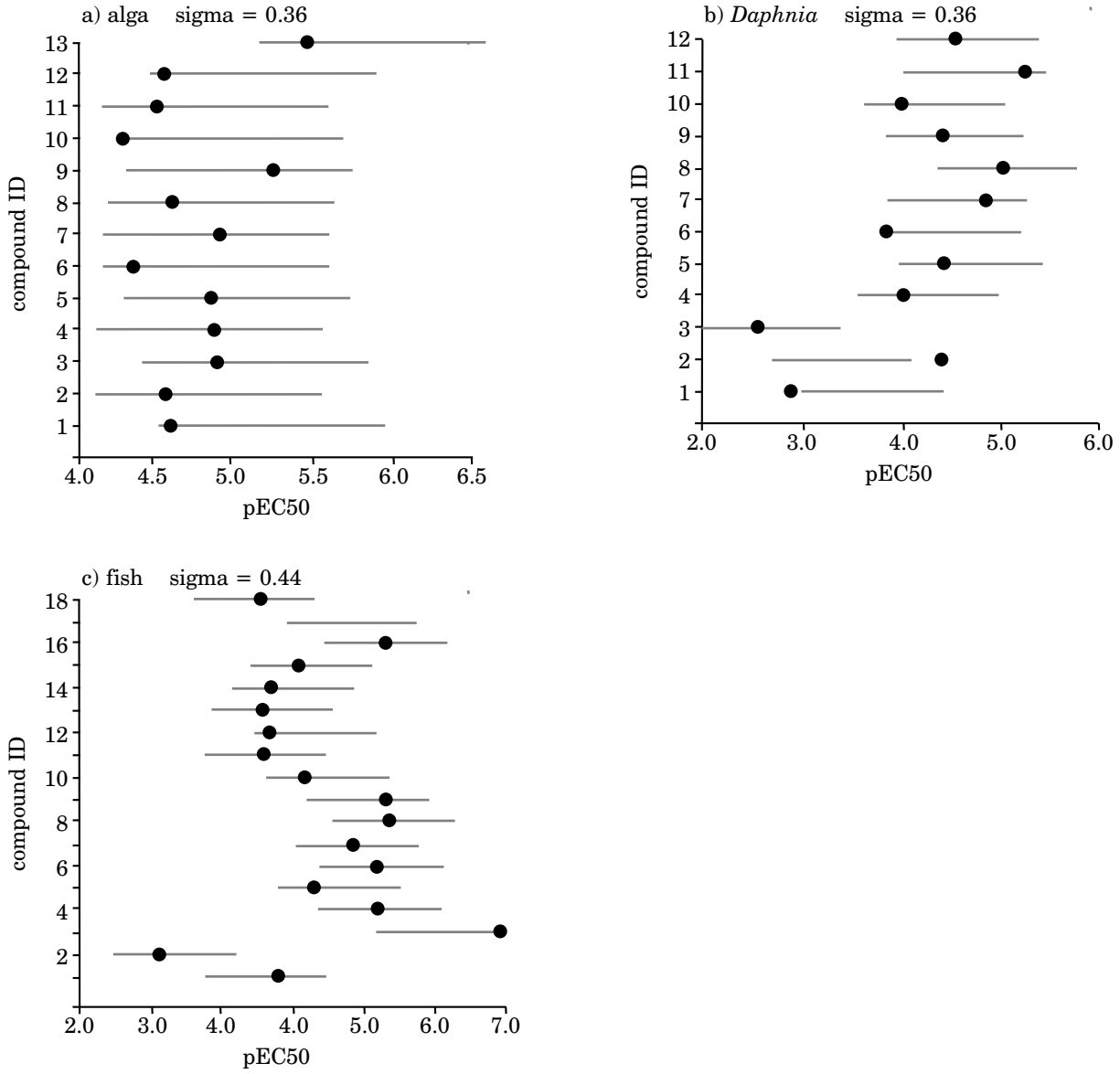
Appendix 3

Case study 3: Validation of individual *in vivo* experimental data by using QSARs

In this case study, we describe the validation of an estimate of an endpoint obtained by experimental (i.e. *in vivo*-derived) testing based on an existing QSAR for the same endpoint. The verification of *in vivo*-derived data by using QSARs is carried out to support the use of such *in vivo* data in a risk assessment or, from the regulators perspective, to check assessments submitted to dossiers.

What we seek is a test for judging whether or not an *in vivo* experimental value is valid. If we are to

Figure A3.1: New experimental values derived from QSAR predictions with uncertainty



——— = *confidence interval;* ● = *experimental value.*

New experimental values compared to 95 % confidence intervals derived from QSAR predictions with uncertainty to open up for validation. The plots show values on aquatic toxicity (pEC50) for three species, an alga, Daphnia *and a fish. The compound ID numbers are the same as in Table A3.1.*

Table A3.1: Threshold values on tail probabilities

ID	CAS No.	α´
Alga		
1	024017-47-8	0.108
2	043121-43-3	0.511
3	060207-90-1	0.557
4	066246-88-6	0.903
5	075736-33-3	0.687
6	076738-62-0	0.188
7	079983-71-4	0.950
8	083657-17-4	0.450
9	083657-24-3	0.579
10	088671-89-0	0.080
11	094361-06-5	0.337
12	106325-08-0	0.115
13	119446-68-3	0.280
Daphnia		
1	000095-14-7	0.040
2	001455-77-2	0.010
3	036791-04-5	0.807
4	043121-43-3	0.544
5	075736-33-3	0.500
6	076738-62-0	0.092
7	079983-71-4	0.457
8	085509-19-9	0.953
9	088671-89-0	0.740
10	094361-06-5	0.407
11	103112-35-2	0.194
12	131983-72-7	0.742

ID	ID in Ref 1	CAS No.	α´
Fish			
1	3	000095-14-7	0.714
2	10	000288-88-0	0.646
3	95	024017-47-8	0.060
4	158	055179-31-2	0.944
5	240	083657-17-4	0.467
6	305	106325-08-0	0.877
7	306	107534-96-3	0.921
8	310	114369-43-6	0.876
9	325	136426-54-5	0.574
10	329	139528-85-1	0.482
11	354	145701-23-1	0.927
12	361	219714-96-2	0.170
13	362	317815-83-1	0.770
14	364	422556-08-9	0.517
15	369	xxx006	0.655
16	371	xxx008	0.990
17	376	xxx013	0.048
18	377	xxx014	0.857

Threshold values on tail probabilities α (denoted by α´) for which the compound fails the test defined by the predictive distribution from the QSAR and confidence level (1 – α).

use statistical inference, we need to specify a probabilistic model for the difference between the experimental value and the true value. This probabilistic model can specify that the difference is a symmetric probability distribution (e.g. normal) with zero mean and variance σ^2.

A 95% (Bayesian) confidence interval is the range within which we expect 95% of the differences to occur. This is the same as saying that there is a 5% chance that the experimental value will fall outside the confidence interval based on a probability distribution with the true value as mean and variance σ^2. The confidence level is usually written as $1-\alpha$, where α is the probability of judging the experimental value as not valid, when in fact it is. The value of α is usually 5%.

A QSAR can be used to provide the probabilistic model for this test. When we believe the QSAR to be true, an experimental value can be verified by testing whether it lies inside a confidence interval based on the resulting predictive distribution assessed for that compound. If the experimental value falls inside the interval with a confidence level of $1-\alpha$, we are $1-\alpha$ confident in the accuracy of the experimental value. Alternatively, one could derive the smallest confidence level for which the corresponding interval covers the experimental value, $1-\alpha'$. If α' is smaller than our chosen significance level α, then the experimental value is poorly supported by the QSAR.

When uncertainty in a QSAR prediction is not considered, the validation is based on a difference between two point values — a QSAR prediction and the corresponding experimental estimate. There is the need for a reference to be able to judge what is large or small. By taking uncertainty in the QSAR prediction into account, it becomes possible to use the predictive distribution as a probabilistic model to test the experimental value. The test is then based on the model (null hypothesis) that the newly tested compound is exchangeable with (or drawn from the same statistical population as) the compounds in the QSAR training data set. The predictive distribution describes our belief in the magnitude of the error in the prediction.

As part of the CADASTER project, this method was applied to the experimental values obtained from the toxicity testing of benzotriazoles performed on three aquatic species. Probabilistic models for the tests were specified by predictive distributions derived from the QSARs obtained by consensus modelling (1). The predictive distributions were derived by a simple rule of thumb that does not distinguish between possible differences in the error of individual predictions. In this case study, we assigned a normal distribution with the QSAR point prediction as the mean and the MSE as its variance, σ^2. This is probably an underestimate of the predictive variance but, since all except one of the compounds fall within the 95th confidence interval (Figure A3.1), we do not think it is reasonable to enlarge it in this par-

ticular case. One of the compounds had an experimental estimate for which the threshold value on the confidence level was smaller than desired (Table A3.1). Given that the estimates for all the other compounds were acceptable, the tests had been performed under similar conditions, and it was close to the boundary of the confidence interval, there is no reason to discard that particular experimental value as not valid.

Reference

1. Cassini, S., Kovarich, S., Papa, E., Roy, P.P., Rahmberg, M., Nilsson, S., Sahlin, U., Jeliazkova, N., Kochev, N., Pukalov, O., Tetko, I., Brandmaier, S., Durjava, M.K., Kolar, B., Peijnenburg, W. & Gramatica, P. (2013). Evaluation of CADASTER QSAR models for aquatic toxicity of (benzo)triazoles and prioritisation by consensus prediction. *ATLA* **41**, 49–64.

Appendix 4

Table A4.1: The decision matrix for conservative estimates of PNEC

Decision	Outcome	Conservative assessment P(PNEC < NEC) high
Regulate: P (PEC > PNEC) > 0.05	EC > NEC, correct EC < NEC, type I error	P (correct) increase P (type I error) increase
No concern: P (PEC < PNEC) > 0.05	EC > NEC, type II error EC < NEC, correct	P (type II error) decrease P (correct) decrease

To 'err on the side of safety' illustrated by the effect of using conservative estimates of PNEC (assuming PEC to be equal to EC for simplicity) on the probability of different outcomes.