CADASTER

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Grant agreement no.: 212668

Collaborative project

Sub-Priority ENV2007 3.3.1.1: In-silico techniques for hazard-, safety-, and environmental risk-assessment

Work package 4: Integration of QSARs within hazard and risk assessment

Task 4.1 QSAR models in a probabilistic risk assessment framework (Deliverable 4.1)

Due date of deliverable: 30 April 2012

Actual submission date: 30 April 2012

Start date of project: 1 January 2009 Duration: 4 years

Lead Contractor: National Institute of Public Health and the Environment (RIVM), Laboratory for Ecological Risk Assessment

Corresponding authors of document: Ullrika Sahlin[1], Tom Aldenberg[2], James Blevins[1], Laura Golsteijn[3], Mark AJ Huijbregts[3], M Sarfraz Iqbal[1], Willie Peijnenburg[4], Emiel Rorije[5] and Igor Tetko[6]

1. Faculty of Science and Engineering, School of Natural Sciences, Linnaeus University, SE- 391 82 Kalmar, Sweden
2. RIVM, IT department, PO Box 1, 3720 BA, Bilthoven, The Netherlands
3. Radboud University Nijmegen, Institute for Water and Wetland Research, Department of Environmental Science, PO Box 9010, 6500 GL, Nijmegen, The Netherlands
4. RIVM, Laboratory for Ecological Risk Assessment, PO Box 1, 3720 BA, Bilthoven, The Netherlands
5. RIVM, Expert Centre for Substances, PO Box 1, 3720 BA, Bilthoven, The Netherlands
6. Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany

Deliverable no. 4.1 – Application of QSAR models for probabilistic risk assessment (report and model)

Project co-funded by the EU Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	x
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Work package 4. Integration of QSARs with risk assessment

Work package leader: Ullrika Sahlin (Partner 5, Linneaus University, Sweden)

Summary

This report is a deliverable within the CADASTER project with an aim to show how to increase the use of non-testing information for regulatory decision whilst meeting the main challenge of quantifying and reducing uncertainty. The objective is to exemplify the integration of testing and non-testing information into assessment models for carrying out safety-, hazard- and risk assessments. The application of QSAR models for probabilistic risk assessment will be discussed in respect to the characterization of uncertainty in QSAR predictions, the propagation of uncertainty in the assessment and sensitivity analysis of individual QSAR predictions with regard to their contributions in the overall risk assessment framework.

Integration of QSARs in chemical safety assessment should acknowledge that treatment of uncertainty is context dependent, and uncertainty is to be interpreted in relation to the background information.

Two case-studies were used to demonstrate the computational framework for QSAR based risk assessment. The lesson was that the application of QSARs in probabilistic risk assessment leads to questions such as:

- Are there any QSAR data available to use as weight-of-evidence of a chemico-specific input parameter?
- Which algorithm for prediction and approach for predictive inference should be used?
- Is a QSAR prediction reliable enough to support the intended decision making?

The conclusions related to uncertainty in QSAR predictions for probabilistic risk assessment were that:

1) The integration of QSARs into probabilistic risk assessment is possible given proper assessments of predictive uncertainty and predictive reliability.

2) Probabilistic risk assessment is supported by QSAR predictions derived from Bayesian predictive inference. Predicting must be done with care, and the use of different bases for predictive inference is possible when a QSAR is treated as a scientific based hypothesis supported by empirical data.

3) The extent of extrapolation in a QSAR prediction influences predictive error and predictive reliability, and the domain of applicability is from an applied perspective context dependent and considered in the treatment of uncertainty.

4) Bayesian predictive inference provide a flexible philosophy for predictive inference of QSARs

This report foresee several aspects of the reporting and documentation of QSARs that need to be changed with respect to the information needs when QSARs are integrated into probabilistic risk assessment. Those will be further explored in the CADASTER deliverable “A guidance document on the use of QSARs in probabilistic risk assessment” (due in December 2012).

Task 4.1 QSAR models in a probabilistic risk assessment framework.

Within this task the following activities were planned to be carried out: 1) Characterization of variability and uncertainty in available experimental data collected by WP2, 2) Characterization of variability and uncertainty in QSARs identified in WP2 and developed in WP3, 3) Sensitivity analyses of individual models with regard to their contributions in the overall risk assessment framework, 4) QSAR modelling of variability, e.g. species sensitivity distributions (SSD), interfacing with WP3, 5) Probabilistic evaluations for a representative set of chemicals for the ecotoxicological endpoints, as specified in REACH. This will be done by implementing the relevant QSAR models – as specified in WP3 – in an EUSES spreadsheet platform.

Two deliverables are planned for task 4.1: D4.1. Application of QSAR models for probabilistic risk assessment (report and model) and D4.2. Guidance on using QSAR models for probabilistic risk assessment (report). This report is deliverable D4.1 and includes the report "Applications of QSARs for probabilistic risk assessment" and a model consisting of a computational platform for QSAR based risk assessment, which is described in the report and available from the corresponding author (Ullrika Sahlin) on request.

Activities performed

The following activities have been performed in task 4.1:

1) Characterizations of variability and uncertainty have been done on experimental data collected by WP2 whenever that has been needed for a specific purpose. The project has identified several QSAR data sets for which variability have shown to be large, that cannot be explained by errors. An approach to assess uncertainty in experimental data has been developed to open up for the comparison to uncertainty in QSAR predictions. A method to consider uncertainty in input data to SSD has been implemented, but the method is in development,

2) The characterization of variability and uncertainty in QSARs identified in WP2 and developed in WP3 has been approached by establishing the bases for predictive inference in general, and in particular for linear regressions. Methods to make assess applicability domain dependent predictive errors have been developed.

3) Sensitivity analyses of individual models with regard to their contributions in the overall risk assessment framework have been carried out on case-studies and will be done in a larger set if compounds representative for the CADASTER classes for general conclusions.

4) QSAR modelling of variability has been done by considering the uncertainty in QSAR predictions for modeling of SSD.

5) Probabilistic evaluations for a representative set of chemicals for the ecotoxicological endpoints, as specified in REACH have been carried out in a study where QSPR predictions to aid assessment of long range transport and overall persistence of BDEs using the Simplebox model. This computational platform has been developed further to perform QSAR and QSPR integrated assessments of Predicted Environmental Concentration, Predicted No Effect Concentration and Maximum Permissible Emission for triazoles, and Expected Risk for BDEs, which are included as case-studies in this deliverable.

Activities foreseen

The principles and bases to assess uncertainty in QSAR predictions is often described and communicated for classification models (Walker, 2003). Taking this into account, the work by WP4 task 4.1 have been focused towards developing and evaluating approaches to assess predictive uncertainty in QSAR (including QSPR) regressions, i.e. models that predict a continuous endpoint as opposed to a categorical.

The characterization of variability and uncertainty in available experimental data collected by WP2 will be continued on more situations, such as to support the validation of the predictive models built in WP3 based on the experimental data derived from WP2. The possibilities of considering variability in data will also be carried on further by identifying possible model algorithms for this such as weighted least square regression.

Uncertainty in input data to a Species Sensitivity Distribution is foreseen to have a large impact on the uncertainty in PNEC or fraction of species affected, as the influence on risk from QSARs predicting effects tend to be larger than the influence from QSPRs supporting the exposure assessment. A goal is therefore to have a model for SSD that can work for a mixture of experimental data and QSAR predictions.

The computational platform for QSAR integrated probabilistic risk assessment will be developed further to be used in the forthcoming studies in the project.

Methods of predictive inference to support the assessment predictive uncertainty and predictive reliability of QSARs developed in the project will be described, evaluated and (in some cases) implemented to the computational platform for future applications and will be reported in deliverable D4.2.

Application of QSAR models for probabilistic risk assessment (report and model)

Ullrika Sahlin[1*], Tom Aldenberg[2], James Blevins[1], Laura Golsteijn[3], Mark AJ Huijbregts[3], M Sarfraz Iqbal[1], Willie Peijnenburg[4], Emiel Rorije[5] and Igor Tetko[6]

1. Faculty of Science and Engineering, School of Natural Sciences, Linnaeus University, SE- 391 82 Kalmar, Sweden
2. RIVM, IT department, PO Box 1, 3720 BA, Bilthoven, The Netherlands
3. Radboud University Nijmegen, Institute for Water and Wetland Research, Department of Environmental Science, PO Box 9010, 6500 GL, Nijmegen, The Netherlands
4. RIVM, Laboratory for Ecological Risk Assessment, PO Box 1, 3720 BA, Bilthoven, The Netherlands
5. RIVM, Expert Centre for Substances, PO Box 1, 3720 BA, Bilthoven, The Netherlands
6. Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany

Acknowledgements

We are grateful to comments and suggestions provided by Paola Gramatica, Nina Jeliazkova, Simona Kovarich, Ester Papa and Tomas Öberg.

List of contents

Index of Tables

Index of Figures

Appendices

1. Introduction

1.1. QSARs in Chemical Safety Assessment

Chemical legislation allows the use of QSARs to support or replace experimental testing risk assessment.

The European legislation Registration, Evaluation and Authorization of Chemicals (REACH) demand that all relevant industrial chemicals be assessed before 2018 and the responsibility for this is on the industry itself (EU, 2006). REACH aims to achieve a proper balance between societal, economic and environmental objectives, and attempts to efficiently use the scarce and scattered information available on the majority of substances. Thereupon REACH aims to reduce animal testing by optimized use of in silico and in vitro information on related compounds. In order to achieve better and more efficient assessments, it is suggested to use all information in an integrated manner (Ahlers et al., 2008). The equivalence and adequacy of different types of information needs to be verified in weight-of-evidence approaches. The REACH regulation advocates the use of non-testing methods, but guidance is needed on how these methods should be used.

Quantitative Structure-Activity Relationships[1] (QSARs) is a non-testing method that predict chemicals activity or property based on analogy saying there is a correlation between a chemical's structure, its physical or chemical properties and a measured biological activity (Walker et al., 2003b; Eriksson et al., 2003). A QSAR consist of descriptors, endpoint being predicted and a derived relationship between descriptors and the endpoint. When the endpoint is a chemical property the model can be termed Quantitative Structure-Property Relationship (QSPR). For further introduction to QSAR we refer to existing literature (Walker et al., 2003b; Puzyn et al., 2010).

Experimental (standard) data have the highest priority when drawing conclusions on the regulatory endpoints. Non-standard information is particularly useful where it can help to avoid an assessment on the bases of invalid or missing experimental data (Ahlers et al., 2008). Potential uses of QSARs in chemical safety assessment are Classification and labeling, design of strategies for experimental testing, hazard assessment, screening, or replacement of experimental values (Eriksson et al., 2003; Cronin et al., 2003). A good introduction to the use of QSARs in risk assessment is the book edited by Walker (2003).

QSARs have been implemented into computer based tools to predict properties, fate, hazard, exposure, and risk. An early set of tools are found in the P2 framework of the USEPA Office of Pollution Prevention for organic compounds (Walker, 2003). These tools uses chemical structure (such as Simplified Molecular Input Line Entry System (SMILES) code or a chemical property such as log Kow) as input, and are provided by the USEPA or are commercially available. This framework is used for the screening level that is of most value when chemical specific data is lacking. QSARs are also included in the OECD toolbox. Since 2003, a large number of tools are freely available beyond the OECD toolbox (Worth, 2010). The extent to which such computational platforms provide uncertainty in QSAR predictions needed for probabilistic risk assessment is limited. An aim with WP4

[1] In this report QSAR is often used as a common name for QSAR and Quantitative Structure-Property Relationship (QSPR).

in the CADASTER project has therefore been to address issues on the quantification of uncertainty in QSAR predictions for the use in probabilistic risk assessment.

Environmental risk is, according to the ecotoxicolgoical risk paradigm, derived from the combination of a chemicals exposure and effect. The European Chemical Safety Assessment characterizes risk based on hazard and exposure assessments (Figure 1.1) (ECHA, 2008a).

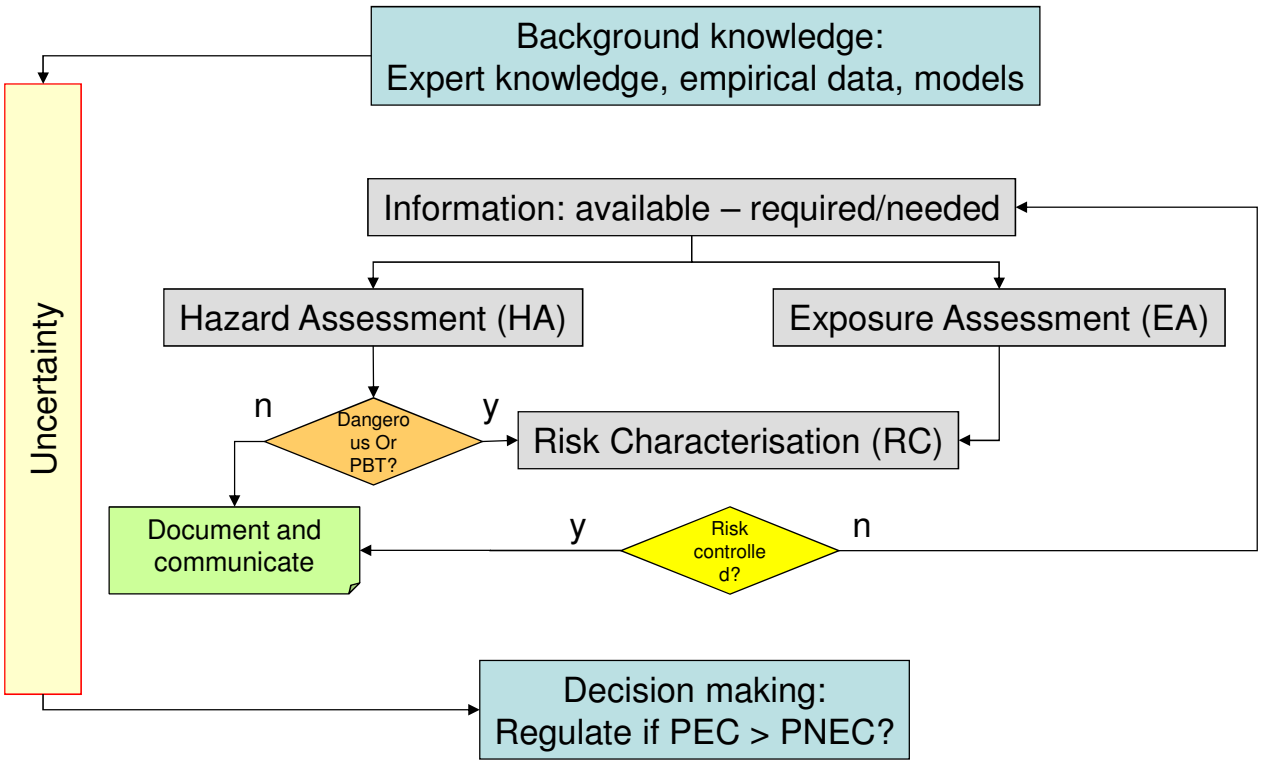


Figure 1.1. The risk assessment scheme modified from ECHA 2008. Guidance on information requirements and chemical safety assessment.

1.2. Probabilistic risk assessment and uncertainty

Risk assessment asks for a quantification of uncertainty, which is to be understood and interpreted in relation to the background knowledge. There are solved and unsolved issues on how to treat uncertainty when going from testing to non-testing information in probabilistic risk assessment.

Risk assessment is a tool to describe uncertainty in unknown quantities (Aven, 2010b). Uncertainty in exposure and effects of chemicals are therefore to be given a proper and transparent treatment (Verdonck et al., 2005; National Research Council, 2009). Environmental risk assessment usually distinguishes variability, i.e. natural variation that cannot be reduced by adding more information, from uncertainty. Epistemic uncertainty (i.e. knowledge based uncertainty) is different from variability (stochastic uncertainty or population level), the latter an inherent property of a quantity or system that cannot be reduced by making more observations or gaining more knowledge.

Probabilistic risk assessment quantifies, as the name suggest, uncertainty by probabilities. To complicate things, there are different kinds of probabilities. For example, probability can be seen as expressing a subjective belief or relative frequency that something occur. Under severe epistemic

uncertainty, probabilities have been criticized for being too precise, which has led to the use of alternative non-probabilistic treatments of uncertainty in risk assessment. If a risk assessors uncertainty due to lack of knowledge and systematic measurement errors (partial ignorance and epistemic uncertainty) is adequately quantified by probability, a major advantage is that its results in an interpretable decision support (Aven, 2010a). There is currently an ongoing discussion of the meaning of probability and the use of probabilistic, non-probabilistic or hybrid approaches, which may be of less interest, but nevertheless useful to be aware of, to researchers in empirical sciences. Risk assessment is a science-based approach, but nevertheless the characterization of uncertainty rest upon assumptions and decisions taken by the risk assessor. Therefore the uncertainty in a risk assessment is of the risk assessor conditioned on the background knowledge (Figure 1.2). In order to avoid the problem of having different kind of probabilities, it has been suggested to regard all probabilities as "subjective[2]" believes, even though there might be probabilities that have been assessed by methods that can be seen as "objective" (Aven, 2010a). The interpretation of a risk assessment product, and especially the probabilities describing uncertainty, is to be interpreted as the risk assessors belief conditioned on the available background knowledge (Aven, 2010b). Background knowledge constitute of empirical data, models (of the system that are to be managed or to relate empirical data to important system variables) and expert knowledge.

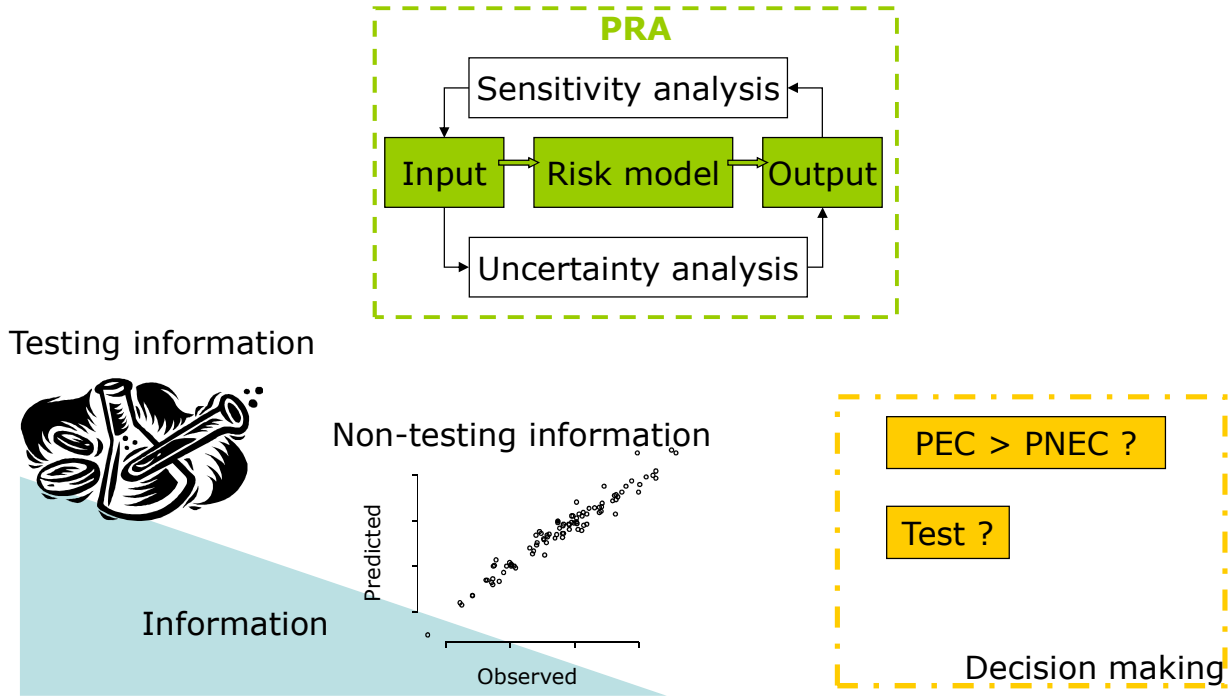


Figure 1.2. QSAR models in probabilistic risk assessment framework lead to interesting questions about the reliability of non-testing versus testing information, the treatment of uncertainty from predictive models in computer models, and the impact non-testing information and its uncertainty may have on decision making.

[2] Using the word subjective here is like throwing stones in a glass house. Nevertheless is it important to emphasis the subjective nature of risk assessment and of the treatment of uncertainty.

This report focuses on two kinds of information[3] in the background knowledge. First, there is empirical data derived from experimental testing of a chemicals specific activities or properties. This so called testing information is strong in the sense that we are confident in to support decision making. The use of QSARs means that experimental data is replaced by a predictive model. Such non-testing information is less strong than testing information, dependent on our confidence in the predictive model. This means that using QSARs in risk assessment not only introduced uncertainty related to a prediction, but also alters the strength of the background information, and it is not clear how to treat both these aspects in probabilistic risk assessment.

2. Applications of QSARs in probabilistic risk assessment

2.1. Uncertainty analysis

Treatments of uncertainty can be qualitatively (tier 1), deterministic (tier 2) using point estimates and worst case assumptions, and probabilistic (tier 3). Probabilistic risk assessment includes analysis of the uncertainty in a risk assessment, followed up by an analysis of the sensitivity of risk to different sources of uncertainty (Figure 2). The CADASTER project specifically aims towards probabilistic risk assessment.

Uncertainty in probabilistic chemical safety assessment can roughly be divided in to three categories: parameter uncertainty, model uncertainty and scenario uncertainty (ECHA, 2008b). Scenario uncertainty is “the uncertainty in specifying the scenario(s) which is consistent with the identified use(s) of the substance” and is of less relevance for the uncertainty related to a QSAR.

The ECHA guidelines (ECHA, 2008b) make the following description of model and parameter uncertainty:

> **Model uncertainty** is the uncertainty in the adequacy of the model used with the scope and purpose of the assessment. In risk assessment, mathematical and statistical models are often applied to represent an exposure or hazard process though a model is always a simplification of reality. Model uncertainty is principally based upon extrapolation (i.e. use of a model outside the domain for which it was developed), modeling errors (i.e. non-consideration of parameters in the model structure itself, assumption of well-mixed phases etc.) and dependency errors (i.e. lack of consideration of correlations between parameters).
>
> **Parameter uncertainty** is the uncertainty involved in the specification of numerical values. Risk assessment involves the specification of values for parameters, either for direct determination of the exposure/effect or as input for mechanistic, empirical or distribution based models which are used. The uncertainties surrounding these values are very common due to lack or insufficiency of data.
>
> Parameter uncertainties include:

[3] In this report we treat knowledge and information as exchangeable terms, aware that information can be restricted to a product when empirical data is confronted with a probabilistic model and can for example be measured by the likelihood.

- Measurement errors: e.g. influence of the methodology used, errors in the analytical method used to measure chemical concentration, technical inadvertence;

- Sample uncertainty: representativeness of the data set, e.g. a small sample may not give the entire range of values found in reality; the sample may be biased towards lower or higher values as a result of the selection criteria used to take the sample; averaging methodologies;

- Selection of the data used for assessing the risk: i.e. use of default data (e.g. TGD default data are frequently used for exposure assessment) or choice of the dose descriptor (i.e. uncertainty in choosing one data among others for risk assessment purpose);

- Extrapolation uncertainty: i.e. use of alternative methods (e.g. QSAR, in-vitro test, read-across for similar substances) or use of assessment factors (e.g. inter-species, intra-species, acute to chronic, route to route, lab to field extrapolation). [end of citation from the ECHA guidelines]

The terms parameter and model uncertainty are not frequently used in QSAR modeling. Instead a distinction lies between predictive uncertainty and predictive reliability (Sahlin et al., 2011) According to Sahlin et al. QSARs applied in risk assessment or decision making result in two kinds of model uncertainties in need of treatment; the reliability in using a model for prediction (also known as confidence in prediction) and the consideration of alternative QSAR models to predict the same endpoint (e.g. consensus modeling by model averaging). The treatment of the second kind of model uncertainty serves as input to the assessment of parameter uncertainty (Apostolakis, 1990).

Here follows a brief introduction to uncertainty in QSAR predictions. Uncertainty in predictions is part of the prediction and is a major concern, especially when predictions may influence human and animal lives as well as the safety of environmental systems. Predictive inference must therefore be as correct as possible (Appendix 1). For a more detailed guidance on treatment of parameter and model uncertainty related to QSARs when applied in probabilistic risk assessment we refer to the forthcoming CADASTER deliverable D4.2.

2.2. Predictive uncertainty

"Every model is associated with a certain degree of uncertainty for QSAR models dominated by input uncertainty arising from the quality of the experimental testing data and input variability arising from heterogeneity in the endpoint and descriptors, and structural uncertainty (also referred to as model uncertainty) arising from the fact that every model is a simplification of reality, or more extreme that every model is wrong, but some are less wrong than others. Despite these uncertainties is the product of the QSAR usually reported as a point estimate." [(Walker et al., 2003a)]

Applying QSARs in risk assessment raises the need to consider uncertainty in predictions and the accuracy of a QSAR prediction in relation to the intended use of the QSAR (ECETOC, 1998). While QSARs are based on data that are variable (e.g. due to measurement errors or variability), the product of the QSAR is reported as a point estimate (page 14 Walker, 2003). In a recent overview of current practice to characterize uncertainty in QSAR predictions Sahlin et al (2011) found that current QSAR practice include several approaches to assess parameter uncertainty (roughly divided into expert judgment, estimates based on re-sampling and assessments based on probabilistic modeling),

but that the integration of QSARs in risk assessment would benefit from probabilistic QSARs in which uncertainty is quantified by probabilities. The need of probabilistic models were pointed out by Walker et al (2003a) who suggested

> *"that errors needs to be evaluated when applying QSARs by providing confidence intervals that take into consideration the uncertainty associated with the estimate"*.

We see several underlying statements in this phrase. First, if a confidence interval can be calculated there must be an underlying probability distribution (parametric or non-parametric) and this is what shall be used to the describe the parameter uncertainty when the QSAR provides an input parameter to a probabilistic risk assessment. Further, if it is implicit that the confidence interval should cover the actual value with a certain degree of confidence, it presumes a Bayesian interpretation of uncertainty (see Appendix 1). There is no problem having a Bayesian approach to uncertainty, this is in fact the way that risk assessment usually deal with uncertainty (Aven 2010). The Bayesian approach is to regard a model as uncertain and let the combination of prior belief and the QSAR training data lead to a posterior belief in what QSAR models that most likely describe observed data (inspired by Obrezanova and Segall (2010). Taking the average over the posterior means gives a point prediction, while the full posterior distribution provides an estimate of the uncertainty in prediction (Obrezanova and Segall, 2010). A predictive distribution is the posterior of observables as opposed to model parameters.

A prediction can be a result of several QSAR models. QSARs can be built on different algorithms for supervised learning and divisions into training and validation data sets. Model Averaging is a technique for consensus modelling of an ensemble model developed on the same training data set, or validated on the same external test set. A test set is a set of chemicals, not present in the training set, that is used to validate (assess the predictive ability of) a QSAR. Model Averaging is a weighted average of predictions where each weight is assigned by some measure of performance based on the common data set using measures of divergence such as Kullback-Leiber divergence, Akaike weights or Bayes factor (Johnson and Omland, 2004). Model averaging is a way to deal with model uncertainty in the probabilistic risk assessment via the characterisation of the predictive distribution. Predictions from several QSARs (e.g. local and global models) can be combined based on expert judgment.

Here it is appropriate to add some comments on predictive error, which is measure describing the distance (error) between a point prediction and the actual value. Predictive error is not a fixed value, it changes from compound to compound. For example, predictive error ought to increase with the extent of extrapolation. This holds even for models where errors are assumed to be equally distributed. The extent of extrapolation is one factor that influences a models predictive reliability, i.e. the reliability in using this particular QSAR to predict a particular chemical compound.

2.3. Predictive reliability

The acceptability of QSAR prediction depend on the regulatory endpoint regarded (Ahlers et al., 2008; ECETOC, 2003). QSAR predictions are regarded as less suitable for activity or effect in chemical classes where absorption, distribution, metabolism and excretion (ADME) is important (ECETOC, 2003). Greater confidence is based on models off acute effects compared to chronic ones and on models of baseline toxicity compared to predictions based on specific models of action or chemical

classes showing more than baseline toxicity. Determining which QSAR models are suitable for regulatory purposes is not the focus in this report, and we refer to existing literature (Gramatica, 2010). We therefore discuss how to evaluate and consider reliability given an acceptable QSAR model.

Predictive reliability, or confidence in prediction, is a statement of the strength of non-testing information as part of the background knowledge. In relation to an experimental test, is a QSAR prediction information of a lower strength, since it is not a direct empirical observation of the activity or property. It is relevant to ask in what way the lower strength of non-testing information can be considered in the probabilistic risk assessment. Overconfidence in a QSAR to produce reliable predictions can be avoided if the assessor is aware of, if, how and why the QSAR was developed and validated. There is a need to understand the limitations of chemical structure representation, descriptors, statistics, data sets, endpoints, and variability of measured data. In order to maintain reliability it has been suggested to test the acceptability of QSARs by the so called OECD principle (OECD, 2007).

The chemical domain for which a QSAR has been built is an important factor to evaluate the reliability of a model, by looking to what respect a compound to be predicted falls inside the applicability domain (Clark and Waldman, 2012). The applicability domain is a region in chemical space determined by the training set and (but less clear) by the the model. There is a danger in treating a QSAR model as a black-box, de Roode et al (De Roode et al., 2006) showed that QSARs are not always in the models domain of applicability and the accuracy of prediction is low. Given that OECD principle of a defined domain of applicability is fulfilled, predictive reliability must be evaluated in every situation where a QSAR model is applied for prediction (Gramatica, 2010). Predictive reliability of a QSAR should be judged both globally (average) and locally (item-specific). Global measures such as (Schultz et al., 2004) confidence index based on crucial factors influencing the confidence of a computation model of toxicity used to compare models, do not say anything about how the confidence in predictions vary between items to be predicted, i.e. does not provide local and item-specific reliability. There are attempts to assess predictive reliability by sensitivity analysis and a shown correlation between a measure of the applicability domain and the assessed predictive reliability (Bosnic and Kononenko, 2009). Even though such correlations have been found, it is without any further elaboration difficult to integrate such qualitative statements in a probabilistic risk assessment.

Important questions are whether a compound lies inside the models domain of applicability, and if the associated uncertainty to a prediction following from predictive inference reflects our confidence in the prediction. Aspects of predictive reliability can be dealt with by flagging (i.e. put it down in the risk report but use the QSAR prediction as it is), go for other non-testing information (maybe in combination with the QSAR prediction), or let it be reflected in the parameter uncertainty followed up by sensitivity analysis. Uncertainty due to extrapolating outside the applicability domain can be dealt with by enlarging parameter uncertainty by some uncertainty factor. Ahlers et al (2008) 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 results in no apparent risk, further testing may be avoided and animals may be saved. For example (from Ahlers et al 2008) if EC50 values for daphnia and algae and a QSAR estimate for fish are available and the PEC/PNEC ratio is very low, a fish test may not be

necessary; whereas a chronic fish test should be considered directly when the PEC/PNEC ratio is high. Thus, sensitivity analysis is a helpful tool to evaluate whether a QSAR prediction can be used or not. The influence of QSAR prediction is not only related to the accuracy of the prediction itself, but depends on how the uncertainty in the prediction propagates in the assessment model, which depends on number of times the parameter is used and if it reduces or increases the assessed risk (page 154 Walker, 2003).

Predictive reliability can be assessed in several ways, roughly divided into measures of extrapolation and measures of performance. The former includes various metrics of the applicability domain (Netzeva et al., 2005). Performance measures includes non-probabilistic performance measures, such as standard deviation in ensemble predictions, uncertainty measures, such as locally assessed predictive errors, and probabilistic performance measures, such as local coverage (hit rates or empirical confidence levels). For example, the variation between ensembles of predictions is a measure of predictive reliability but not an estimate of predictive error per se. However, it can be correlated to predictive error, since items for which predictions differ between models most likely are given less reliable predictions.

Assessments of predictive uncertainty and predictive reliability have been carried out in WP4 in the CADASTER project. There are approaches that use non-parametric bootstrap based upon an assumption of a positive correlation between reliability and predictive error. A nice feature with bootstrapping is that the assessment of reliability dependent predictive errors does not need external data. Instead, predictive errors are calibrated using n-fold cross-validation, e.g. bagging approaches (Tetko et al., 2006; Tetko et al., 2008; Sushko, 2010; Sushko et al., 2010a; Sahlin, 2012), or locally assessed Predictive Error Sum of Squares (Sahlin et al., submitted). The assessment of predictive errors is done according to the concept of “distance to model”, which is a generalized idea of a similarity of a tested molecule to the training set molecules. Several distance to models were analyzed and benchmarked (Tetko et al., 2008; Sahlin et al., submitted). The concept has been further extended for classification models (Sushko et al., 2010a; Sushko et al., 2010b). A complete description of the analysis and discussion of the concept “distance to model” is found PhD thesis of I Sushko (2010). Parametric bootstrap includes Bayesian predictive inference (discussed in Appendix 1 and 2).

A decision maker may be interested in the consequences the model usage may have on the accuracy of the risk assessment. To this end, a useful measure of predictive reliability is the probability of a prediction being wrong (i.e. 1 – probability of being accurate). The probability of having an erroneous prediction for compounds at the border of the AD can assist in the consequences of being on the border. In this respect, many of the measures of reliability fail as they are not probabilistic. Conan statistics provide different measures related to the probability of making different kinds of errors and are applicable when the prediction is a classification. For classification models specificity and sensitivity and the uncertainty in the probability of being in one class or the other may vary over the applicability domain. For regression models, even though predictive reliability ought to decrease, and predictive error ought to increase, in regions of the applicability domain where the model is less defined, it does not necessarily imply that a prediction having a larger predictive error must be less reliable. We therefore recommend to avoid using the predictive error (as it is) to characterize the accuracy in using a model for prediction (predictive reliability). Of importance is if the prediction and its associated uncertainty cover the actual value (Sahlin, 2012). There is a relationship between the

domain of applicability and predictive error (Weaver and Gleeson, 2008), but the change in predictive error may not be large enough to reflect the reduced reliability in model predictions. Instead we argue for the use of confidence (empirical confidence, but see also tolerance intervals) that reveal how well we believe the predictive distribution is expected to cover the actual value. Tong et al. (2004) assessed coverage (i.e. accuracy estimated as the number of compounds that fell inside the corresponding prediction interval) for a given confidence level over different regions of the AD defined by extrapolation measured by the proportion of items in the training data set that are further away than the item to be predicted. Coverage was lowest for the most extreme region of the applicability domain. It is however difficult to get good measures of predictive performance in the most extreme regions as there is by definition few data points there.

The probability of committing different types of errors to guide decision making whether the risk assessment is reliable or if it is worthwhile to reduce the probability of being wrong in some of the input parameters. This is easiest to understand for a classification models (a test) and where the outcome of the test directly influences the decision (Jaworska et al., 2010). The decision to test or not is directly seen whether a test have an increase in the expected utility (or decrease in expected loss). However, it is difficult to derive how the probability of an input parameter of being wrong propagates through a risk assessment models, such as the Simplebox. The option is instead to study the influence uncertainty in an input parameter has on the overall uncertainty of the assessed output (Iqbal and Öberg, In review). For example, will a reduction of the uncertainty change the decision by moving a critical value such as the 95^{th} percentile of the PEC/PNEC ratio over a decision threshold?

When the extent of extrapolation is judged as unacceptably high the recommendation could be to

- Do not use a QSAR if the compound to be predicted is an unacceptable extrapolation.

Alternatively one could use the QSAR but

- Flag that the compound is extrapolated and in a sense judged as being outside the statistical applicability domain by reporting the extent of extrapolation from the QSAR training data set.
- Add extrapolation uncertainty to the predictive uncertainty derived by predictive inference.
- Combine QSAR prediction with other non-testing methods.

Read Across is a frequently used non-testing method in the OECD toolbox. Both QSAR and Read Across predict by analogy, where the major difference being that QSAR are based on a learning algorithm and asses the uncertainty or accuracy in predictions based on empirical data, whereas uncertainty or accuracy in read across is if at all derived by expert judgment. Alternatively a Read Across is an extreme kind of local QSAR.

In all these cases is it important to communicate the lowered reliability in the prediction in the risk assessment report (“AD statement” in Figure 2.1). A recommendation is to follow up the use of a parameter with associated low reliability with a sensitivity analysis of its impact on the regulatory decision.

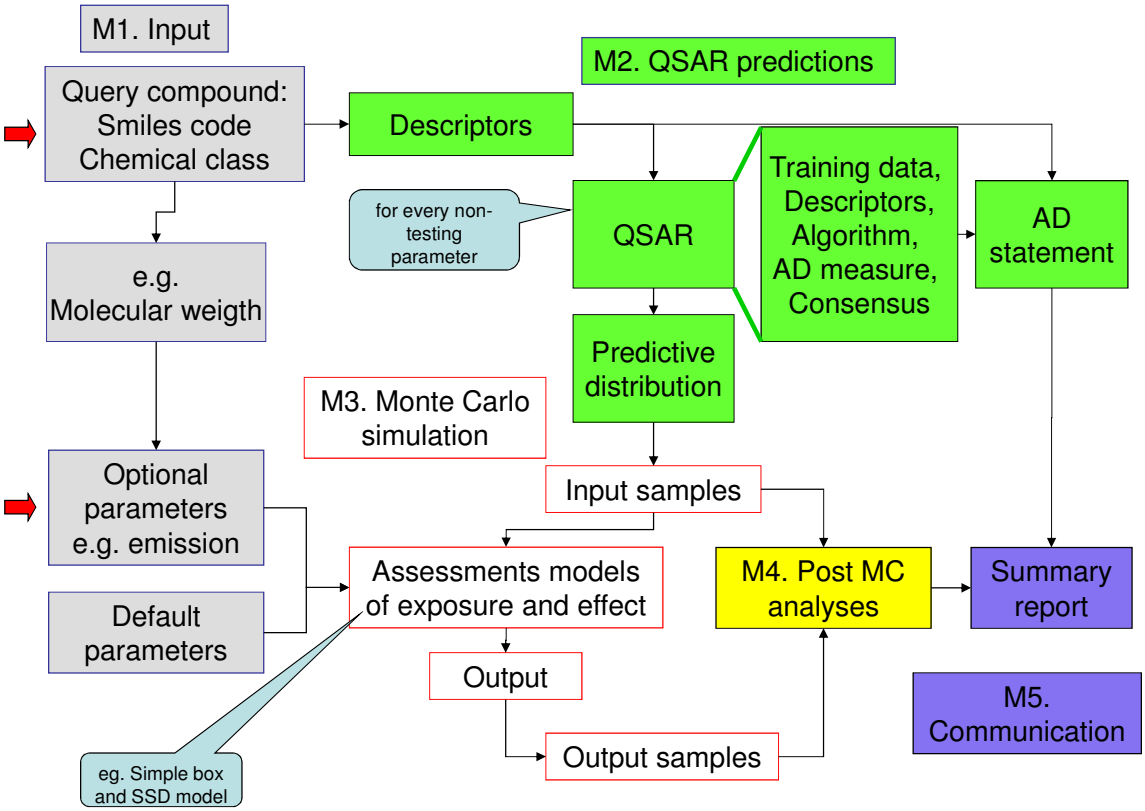


Figure 2.1. Workflow for QSAR-based risk assessment.

2.4. The integration of QSARs into probabilistic risk assessment illustrated by two case studies.

For proper risk assessment, a systematic procedure through estimation of exposure and effects is required (Van de Meent, 1998). Within the CADASTER project, we studied four groups of chemicals for which the risk assessment is hampered by the fact that chemical monitoring data as well as toxicity measurement data are rarely available. Quantitative structure-property relationships (QSPRs) have been developed that can be used to model a chemical's fate in the environment when measured data for chemical properties are lacking. Similarly, quantitative structure-activity relationships (QSARs) have been developed to predict a chemical's effects. The use of QSARs and QSPRs makes a full risk assessment possible, provided that their predictions qualify as replacement of experimental data (or empirical observations). However, with respect to the reliability and uncertainty of the input data, careful interpretation of the outcome is required. Therefore, the use of QS(A)PRs in risk assessment can be justified if their uncertainties were treated properly and do not have substantial impact on the regulatory decision following risk assessment. We illustrated the integration of QS(A)PRs in probabilistic risk assessment by two case studies, one on Triazoles and one on PolyBrominated Diphenyl Ethers (PBDEs). Within these case studies, we determined the influence of the use of QS(A)PRs on the uncertainty in the outcome of a risk assessment.

The work was implemented into a computational platform for QSAR-based probabilistic risk assessment, consisting of the 5 modules:

- Module 1. Input
- Module 2. QSAR and QSPR predictions
- Module 3. Monte Carlo simulation and assessment
- Module 4. Post Monte Carlo analysis
- Module 5. Communication

In the first module the compound to undergo risk assessment is described and chemico-specific parameters specified (Figure 2.1). Further, molecular descriptors for the QSAR in the second module are calculated. The input module also offers the possibility of specifying other parameters in the risk assessment models, which otherwise are given default values. This means that uncertainty is not specified in other parameter than the chemical- specific. Based on the predictive distribution and evaluated predictive reliability, probability distributions for the input parameters are defined and sampled, and propagated into the exposure and effect assessments using Monte Carlo simulation. Based on the output from exposure and effect assessment regulatory endpoints such as Risk Characterisation Ratio (PEC/PNEC) or Expected Risk (Appendix 3) is calculated (Figure 2.2). The result of the uncertainty and sensitivity analysis, such as an evaluation of predictive reliability for each QSAR model and its importance on the regulatory endpoint, are communicated.

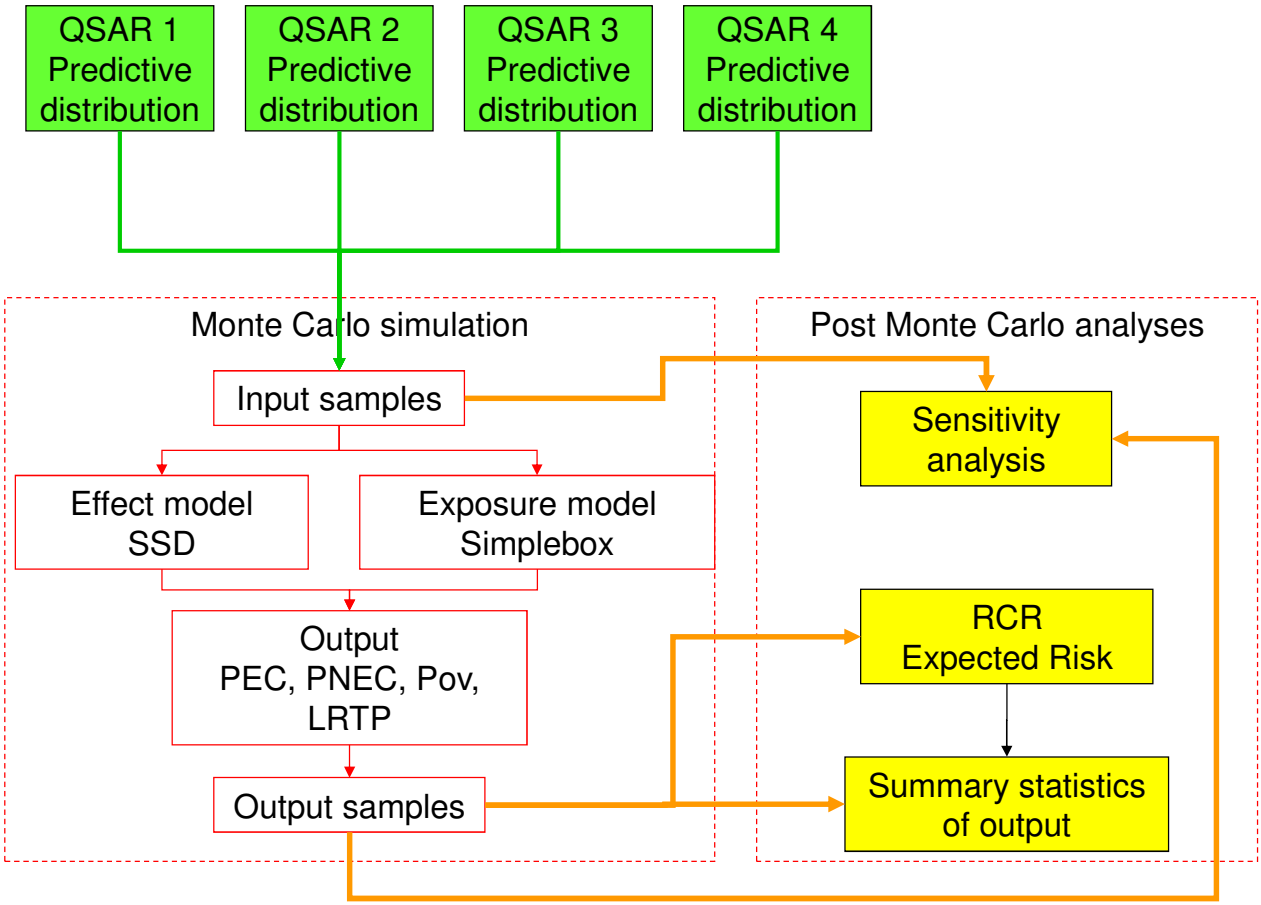


Figure 2.2. Detailed workflow of the computational platform of QSAR-based risk assessment.

3. CASE-STUDY 1: Uncertainties in a Triazole risk assessment based on QS(A)PRs

3.1. Summary

The risk assessment of the fungicides triazoles is hampered by a lack of monitoring and toxicity data. The goal of this study was to determine the influence of the use of quantitative structure-activity (property) relationships (QS(A)PRs) on the uncertainty in the risk assessment for a selection of five triazoles. Soil sorption partition coefficients, solubility, melting point, vapor pressure and hydroxyl radical reaction in air were predicted with QS(A)PRs; biodegradation rates with combined use of semi-quantitative ratings and experimental half-lives; and no effect concentrations with QSARs. All were implemented in the multimedia fate model Simplebox. Parameter uncertainty was treated as a probability distribution, and assessed using statistical methods propagated by Monte Carlo Analyses. The maximum permissible emissions (MPE) to agricultural soil were highest for Bromuconazole and Difenoconazole, i.e. 2.09 ·106 and 2.26 ·106 kg/day, respectively (90%-confidence intervals (CIs) of four orders of magnitude). For Tebuconazole, Triazemate, and Metconazole we found MPEs between 5.15 ·104 and 8.00 ·104 kg/day (90%-CIs of three to five orders of magnitude). Uncertainty in the MPE to agricultural soil was mainly determined by uncertainty in the soil sorption partition coefficient (10.8 – 58.3%), the biodegradation in water (30.1 – 82.7%), and the toxicity (up to 10.6%).

3.2. Introduction

Triazoles are chemical compounds that are globally used for fungi control. Their importance for pest management has increased, among other reasons because of their broad spectrum of activity (Klix et al., 2007). Belonging to the group of demethylation inhibitors (DMIs), triazoles act specifically on the biosynthesis of ergosterol (Maštovská, 2005). Each triazole compound may act slightly different on the ergosterol production pathway, but in the end they all cause abnormal fungal growth and death. The application of triazoles to plants and crops can lead to contamination of the aquatic environment, i.e. ground and surface waters (Li and Randak, 2009). Although triazoles were designed to interfere with the ergosterol biosynthesis in target fungi, they can also display different modes of action in non-target organisms. Hassold and Backhaus (2009) showed that DMI fungicides from four different chemical classes, including triazoles, exhibit baseline toxicity as well as specific toxicity in *Daphnia Magna*. Furthermore, Ankley et al. (2005) showed multiple modes of actions of the DMI fungicides prochloraz and fenarimol in Fathead Minnow. Hassold and Backhaus (2009) emphasize the risk for aquatic invertebrates due to the high toxicity and ubiquitous use and resulting occurrence in the aquatic environment of demethylation inhibitors.

For proper risk assessment, a systematic procedure through estimation of exposure and effects is required (Van de Meent, 1998). Important steps are the determination of the Predicted Environmental Concentration (PEC) and the Predicted No Effect Concentration (PNEC). However, the risk assessment of triazoles is hampered by the fact that chemical monitoring data as well as toxicity measurement data are rarely available. Therefore, they are one of the four classes of chemicals studied in the European Union-Framework Project-7 CADASTER (CAse studies on the Development and Application of in Silico Techniques for Environmental hazard and Risk assessment) project (EU FP7 CADASTER, 2009), in which the authors are involved. Within this project, quantitative structure-property relationships (QSPRs) have been developed that can be used to model a chemical's fate in

the environment when measured data for chemical properties are lacking. Similarly, quantitative structure-activity relationships (QSARs) have been developed to predict a chemical's effects.

The use of QSARs and QSPRs makes a full risk assessment possible, provided that their predictions qualify as replacement of experimental data or empirical observations. However, with respect to the reliability and uncertainty of the input data, careful interpretation of the outcome is required. Therefore, the use of QS(A)PRs in risk assessment can be justified if their uncertainties were treated properly and do not have substantial impact on the regulatory decision following risk assessment. The goal of this study was to determine the influence of the use of QS(A)PRs on the uncertainty in the outcome of a risk assessment for triazoles. We implemented QSPRs in the multimedia fate model Simplebox (Den Hollander et al., 2004) to predict the aquatic concentration of triazoles after a single unit emission, and used QSARs to model the no effect concentrations. In the end, the uncertainty in the outcome was quantified, and a sensitivity analysis was performed to determine the relative contribution of the different predictive models to the overall uncertainty.

3.3. Method

3.3.1. Risk assessment based on QS(A)PRs

The environmental fate of the triazole fungicides is determined by different chemical properties and processes. We used QSPRs to predict chemical properties, which enabled the environmental fate modeling of the triazoles. Soil sorption partition coefficients (K_{oc}) were predicted with a multiple linear regression developed for a set of heterogeneous, organic, non-ionic compounds by Gramatica et al. (2007). Aqueous solubility, melting point, and vapor pressure were predicted with the multiple linear regressions of Bhhatarai et al. (2011). The rate constants for hydroxyl radical reaction in air were predicted with the multiple linear regression of Roy et al.(Roy et al., 2011). All QSPR models fulfill the fundamental principles laid down by the OECD (OECD, 2007). The descriptors used in the QSPR models can be found in Table 3.1. For more information about the descriptors used in the multiple linear regressions we refer to the references mentioned.

Biodegradation rates are a function of both the chemical properties and the surrounding environment. They are very uncertain and have not been measured for most chemicals. The time required for biodegradation in the aquatic environment was predicted with combined use of the Biowin3 semi-quantitative ratings from Episuite™ (Boethling et al., 1994) and the experimental half-lives determined by Aronson et al. (2006). The half-lives for soil and sediment were assumed to be two and nine times as long as in water (US EPA, 2002).

The predicted environmental concentrations of triazoles in fresh water were modeled with the Simplebox model (Den Hollander et al., 2004). This is a fugacity model in which the environment is modeled as a set of homogenous compartments; one compartment for each environmental medium in which the chemical is assumed to be evenly distributed. Results from Simplebox are commonly used in EU risk assessments for new and existing chemicals (European Commission, 2003b). We modeled dissolved freshwater concentrations on the regional scale after a single unit emission to agricultural soil.

Simplebox was also used to predict the triazoles' potential for long-range transport (LRTP) through the environment. It was defined as the fraction transferred from the regional scale to the continental and Northern hemispheric scale:

$$LRTP_x = M_{r,x} / M_{tot,x} \quad (1)$$

In this equation $LRTP_x$ is the dimensionless long-range transport potential of chemical x, $M_{r,x}$ is the steady state mass of chemical x on the regional-scale, $M_{tot,x}$ is the total steady state mass of chemical x present in the environment.

Long-range transport potential and degradation half-lives together determine the chemical's overall persistence. A commonly used numerical indicator for the overall persistence of a chemical is its overall residence time in the environment (Klasmeier et al., 2006). This can be calculated by:

$$P_{ov,x} = M_{tot,x} / E_x \quad (2)$$

where $P_{ov,x}$ is the overall residence time of chemical x in the environment (days), $M_{tot,x}$ is the total steady state mass of chemical x present in the environment (kg), and E_x is the emission rate of chemical x (kg/day).

According to the European Commission (2003b), an aquatic effect assessment should be composed of at least one short term LC50 or EC50 for each trophic level, i.e. a base set of algae, Daphnia and fish. In this study, we used QSARs based on dragon descriptors for the derivation of toxic concentrations. Three multiple linear regressions were available for triazoles, namely for the LC50 of *Onchorynchus Mykiss*, for the EC50 of *Daphnia Magna*, and for the EC50 of *Pseudokirchneriella Subcapitata*. The descriptors used in the QSAR models can be found in Table 3.1. With little effect data, the PNEC is determined by using fixed assessment factors that are calculated by means of a statistical extrapolation model with an arbitrary cut-off value set at a protection level of 95 percent of the species (Bro-Rasmussen, 1988; European Commission, 2003b). This should account for, among other things, the different sensitivities of other untested species. Here, the PNEC was predicted as:

$$PNEC_x = \frac{\min(LC50_{s1,x}, EC50_{s2,x}, EC50_{s3,x})}{1000} \quad (3)$$

where the PNEC of chemical x (g/L) is the minimum of the toxicity measures available for chemical x in species 1 to 3 (*Onchorynchus Mykiss, Daphnia Magna, and Pseudokirchneriella Subcapitata*, respectively), divided by the assessment factor 1000.

The risk assessment was performed on the basis of the maximum permissible emission for chemical x (MPE_x in kg/day), i.e. the maximum emission to agricultural soil without effects in 95 percent of the aquatic species. It was calculated as the ratio of the PNEC and the aquatic PEC multiplied by the emission mass (E_x in kg/day):

$$MPE_x = \frac{PNEC_x}{PEC_x} \cdot E_x \quad (4)$$

A risk assessment based on QS(A)PRs was performed for a selection of five triazoles: Tebuconazole (CAS 107534-96-3), Triazamate (CAS 112143-82-5), Bromuconazole (CAS 116255-48-2), Difenoconazole (CAS 119446-68-3), and Metconazole (CAS 125116-23-6). All were known to be commonly used.

Table 3.1: QSPRs used for the estimation of the physico-chemical properties at 25°C, and QSARs used for the estimation of the no-effect concentration

Parameter	Unit	Multiple linear regression	Reference
soil sorption partition coefficient (K_{oc})	L/kg	Log K_{oc} = – 1.92 + 2.07 VED1 – nHAcc – 0.31 MAXDP – 0.39 CIC0	Gramatica et al. 2007
aqueous solubility (WS)	mg/L	Log WS = 13.80 – 2.41·CICO – 0.44·AMW + 1.65·MATS7e	Bhhatarai et al. 2011
melting point (MP)	°C	MP = 1098.25 – 162.83·R2e + 53.22·GGI4 + 26.82·F03[N-N] – 1693.0·χ1A	Bhhatarai et al. 2011
vapor pressure (VP)	mmHg	Log VP = 17.30 – 15.67·BELp2 + 0.44·RBN + 1.38·B09[N-Cl]	Bhhatarai et al. 2011
rate constants for hydroxyl radical reaction (k_{OH})	cm^3s^{-1}/mol	Log $1/k_{OH}$ = 4.07 – 0.72 HOMO + 0.37 nX + 0.16nCbH – 0.34IDE	Roy et al. 2011
LC50 Onchorynchus Mykiss	Mol/L	Log 1/LC50 = a + b_1 CIC1 + b_2Mp + b_3H-052 – b_4TPSA(Tot)	unpublished equation[4]
EC50 Daphnia Magna	Mol/L	Log 1/EC50 = a - b_1TPSA(NO) + b_2Aeigm + b_3nCar + b_4nHDon + b_5H-052	unpublished equation
EC50 Pseudokirchneriella Subcapitata	Mol/L	Log 1/EC50 = a + b_1AEigZ + b_2T(N..S) + b_3Seigv	unpublished equation

CIC0	Complementary Information Content index: neighborhood symmetry of 0-order (i.e. the degree of the diversity of the elements in the molecule)
AMW	Average Molecular Weight
MATS7e	Moran autocorrelation of lag 7 weighted by Sanderson electronegativity (i.e. the charge distribution)
R2e	R autocorrelation of lag 2 weighted by Sanderson electronegativity (i.e. the geometry topology and atomic weight assembly)
GGI4	topological charge index of order 4 (i.e. charge transfer between atom pairs)
F03[N-N]	Frequency of N - N at topological distance 3
χ1A	connectivity index of order 1 (Randic connectivity index) (which describes molecular branching and complexity)
BELp2	Lowest eigenvalue n. 2 of the Burden matrix weighted by atomic polarizabilities
RBN	number of rotatable bonds
B09[N-Cl]	Presence/absence of N - Cl at topological distance 9
VED1	eigenvector coefficient sum from distance matrix
nHAcc	number of acceptor atoms for H-bonds (N,O,F)
MAXDP	maximal electropological positive variation
HOMO	highest occupied molecular orbital
nX	number of unsubstituted sp^2-carbon in any ring, mainly aromatics
nCbH	number of unsubstituted sp^2-carbon in benzene-type rings
IDE	mean information content on the distance equality (topological descriptor similar to CIC0)
CIC1	Complementary Information Content index: neighborhood symmetry of 1-order (i.e. the degree of the diversity of the elements in the molecule)

[4] a, b_0,…,b_5 are intercept and coefficients in a linear regression.

Mp	mean atomic polarizability (scaled on Carbon atom)
H-052	H attached to C0(sp3) with 1X attached to next C
TPSA(Tot)	topological polar surface area using N,O,S,P polar contributions
TPSA(NO)	topological polar surface area using N,O polar contributions
Aeigm	Absolute eigenvalue sum from mass weighted distance matrix
nCar	number of aromatic sp^2-carbon
nHDon	number of donor atoms for H-bonds (N and O)
AEigZ	Absolute eigenvalue sum from Z weighted distance matrix (Barysz matrix)
T(N..S)	sum of topological distances between N..S
Seigv	Eigenvalue sum from van der Waals weighted distance matrix

3.3.2. Quantification of uncertainties

The integration of model predictions into probabilistic risk assessment requires validation of the QS(A)PR for its ability to make reliable predictions, and quantification of the associated uncertainty (ECHA 2008; Sahlin et al., 2011). Parameter uncertainty is treated as a probability distribution describing the range of possible values (or classes if categorical), and can be assessed using statistical methods based on empirical data or expert judgment. With respect to the use of QSA(P)Rs in risk assessment, model uncertainty means the reliability of a QSA(P)R in predicting a property of a specific compound (Sahlin et al., 2011). The reliability is among other things restrained by the applicability domain (AD) of the QS(A)PRs, which is a research area in development (Nikolova and Jaworska, 2004). We discussed the consequences of the AD for our risk assessment in the discussion section.

We restricted the uncertainty analysis to chemical-specific input parameters, thereby excluding uncertainty in e.g. landscape parameters. The experimental data underlying the multiple regressions were used to assess statistical uncertainty in the QSPR and QSAR predictions. The uncertainty in a prediction Y based on the descriptors W using a QS(A)PR as a linear regression fitted by ordinary least squares was assigned according to the approach of statistical inference discussed in Appendix 1 and 2. In this case, the predictive distribution is fully specified by the predictive mean PRED(Y), the predictive error SEP(Y), the number of data points in the training data set (n) and the number of descriptors in the linear regression model (p), as:

$$Y \sim PRED(Y) + t_{n-p-1} \cdot SEP(Y) \quad (5)$$

where t_{n-p-1} stands for the *t*-distribution with n–p–1 degree of freedom. The predictive error is estimated as

$$\left[SEP(Y)\right]^2 = \sigma^2 (1 + W^T (X^T X)^{-1} W) \quad (6)$$

where σ^2 is the variance in model errors and $(X^TX)^{-1}$ is the information matrix (e.g. page 46 and onwards in Box and Tiao, 1992)

Because of a lack of more precise information, the uncertainty in biodegradation half-lives was not treated with statistical methods. Instead we made an expert judgment and assigned a log-normal distribution (Slob, 1994), of which the geometric mean and geometric standard deviation were based on the work of Aronson et al. (2006). This is an arbitrary but plausible choice since biodegradation shows natural variability.

Finally, the uncertainties in the predictive modeling output were determined in Monte Carlo Analyses using the spreadsheet-based application Chrystal Ball (Oracle©, Release 11.1.2.0.00, March 2010) in MS Excel with 10,000 iterations per run.

A sensitivity analysis was performed to determine the relative contribution of the uncertainty per input parameter to the uncertainty in the aquatic PEC, in the PNEC, and in the MPE for an emission to agricultural soil. Chrystal Ball was used to calculate the Spearman's rank correlation coefficients between each input parameter and the outcome variable, as a measure of statistical dependence between the two. By squaring the rank correlation coefficients and normalizing them to 100 percent, the contribution to variance was calculated. This way, the relative contributions were obtained for the impact a QS(A)PR has on the uncertainty in the outcome variable, via both its uncertainty and its model sensitivity.

3.4. Results

3.4.1. Probabilistic risk assessment

Figure 3.1 shows the median potential for long range transport of the five triazoles assessed in this study, which ranged from 2.64 $\cdot 10^{-5}$ for Difenoconazole to 3.27 $\cdot 10^{-3}$ for Tebuconazole, with 90% confidence intervals (90%-CIs) of up to six orders of magnitude. Looking at persistency, Triazemate differs from the other four triazoles. Its median value for overall persistency is 2.54 $\cdot 10^{1}$ days with a 90%-CI of almost three orders of magnitude, whereas the other chemicals have a median overall persistency of 1.46 $\cdot 10^{2}$ to 1.52 $\cdot 10^{2}$ days with accompanying 90%-CIs ranging two orders of magnitude. The differences between the five triazoles for the aquatic PEC show the same pattern as LRTP. The median aquatic PEC value after an emission of 1 kg/day to agricultural soil was the lowest for Difenoconazole (1.24 $\cdot 10^{-13}$ g/L), and the highest for Tebuconazole (1.41 $\cdot 10^{-11}$ g/L). The 90%-CIs ranged up to five orders of magnitude. Bromuconazole was the least toxic triazole in this study. We found median PNEC values ranging from 3.13 $\cdot 10^{-7}$ g/L to 2.27 $\cdot 10^{-6}$ g/L with 90%-CIs ranging one to two orders of magnitude. Consequently, the typical maximum permissible emissions to agricultural soil were highest for Bromuconazole and Difenoconazole, to be exact 2.09 $\cdot 10^{6}$ and 2.26 $\cdot 10^{6}$ kg/day, respectively, with 90%-CIs of four orders of magnitude. For Tebuconazole, Triazemate, and Metconazole we found lower typical MPEs, that is between 5.15 $\cdot 10^{4}$ and 8.00 $\cdot 10^{4}$ kg/day, with 90%-CIs ranging three to five orders of magnitude. **Table S1** (Appendix 4) shows the predictions of the input parameters with their predictive error (or geometric mean and standard deviation for the half-lives in water) for all triazoles in this study and the assigned distribution.

3.4.2. Sensitivity analysis

In a sensitivity analysis, the relative contribution of the uncertainty per input parameter to the variance of the outcome variable for an emission to agricultural soil was quantified. **Table 3.2** shows that the uncertainty in the aquatic PEC and MPE for agricultural soil was mainly determined by uncertainty in the soil sorption partition coefficient, and in the biodegradation rate in water. However, uncertainty in the toxicity to different species was also relevant. The contribution to variance was <0.05 percent for water solubility, melting point, vapor pressure, and hydroxyl radical reaction in air, which were therefore excluded from the table. The five triazoles in this study showed differences in the importance of the input parameters. The relative contributions to the variance of the MPE for agricultural soil ranged from 10.8 to 58.3 percent for the K_{oc}, from 30.1 to 82.7 percent for the $k_{biodeg,water}$, from 1.5 to 6.4 percent for the LC50 of *Onchorynchus Mykiss*, from <0.05 to 0.6

percent for the EC50 of *Daphnia Magna*, and from 1.9 to 10.6 percent for the EC50 of *Pseudokirchneriella Subcapitata*.

Table 3.2: Relative contribution to the variance of the aquatic PEC, of the PNEC, and of the maximum permissible emission to agricultural soil.

Deterministic parameter		Tebuconazole	Triazamate	Bromuconazole	Difenoconazole	Metconazole
Relative contribution to the uncertainty in the aquatic PEC (%)						
Physicochemical properties	K_{oc}	47.7	11.5	61.3	65.8	49.5
Biodegradation in water	$k_{biodeg,water}$	52.2	88.4	38.6	34.0	50.4
Relative contribution to the uncertainty in the PNEC (%)						
LC50 O. Mykiss		27.7	75.3	68.5	29.1	16.5
EC50 D. Magna		3.2	-	1.4	2.7	4.0
EC50 P.Subcapitata		69.1	24.6	30.1	72.2	79.7
Relative contribution to the uncertainty in the maximum permissible emission to agricultural soil (%)						
Physicochemical properties	K_{oc}	42.3	10.8	55.4	58.3	43.4
Biodegradation in water	$k_{biodeg,water}$	45.7	82.7	34.6	30.1	43.8
LC50 O. Mykiss		2.6	4.5	6.4	2.7	1.5
EC50 D. Magna		0.5	-	0.2	0.4	0.6
EC50 P.Subcapitata		8.8	1.9	3.3	8.4	10.6

3.4. Discussion

3.4.1. Method

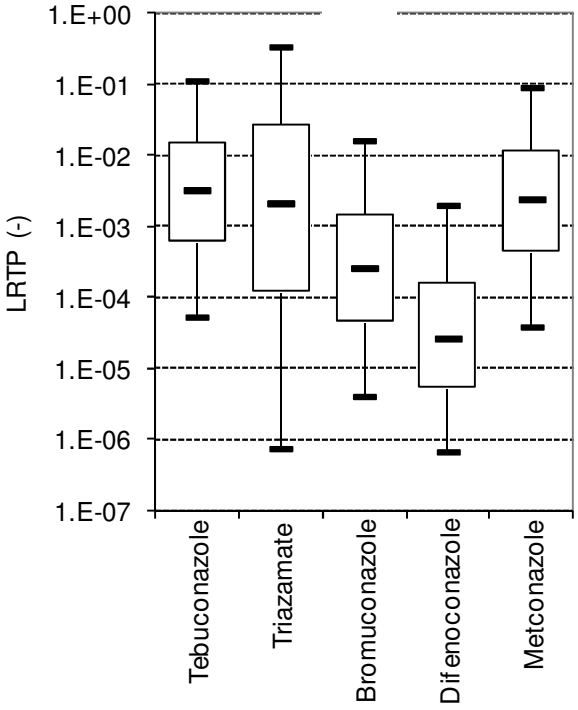
The environmental fate of triazoles in air was diminished by photolytic degradation via indirect photolysis, i.e. through a reaction with photo-oxidizing OH-radicals (European Commission, 2003a). In soil, indirect photolysis was not taken into account, since Kim et al. reported it is minimal (2002). In water, photolytic degradation involves both indirect photolysis and direct reactions according to some authors (Vialaton et al., 2001; Vialaton and Richard, 2002). Wallace et al. (2010) found as well that indirect photolysis via a reaction with photo-oxidizing nitrate-radicals significantly enhances the degradation of propiconazole in water. However, the relevance of direct photolysis, by UV irradiation, is not conclusive. Wallace et al. (2010) stated that propiconazole is stable to direct photolysis. Abramovitch et al. (2001) and Da Silva et al. (2001) explained that direct photolysis in water is not expected because triazoles do not absorb irradiation with a wavelength of $\lambda > 200$ nm. In addition, Breedveld et al. (2002) reported that direct photolysis in water requires high radiation doses. The European Commission (2003a) concluded that direct photolysis is not significant. Hydrolysis in water was not included in the model calculations, since the European Food Safety Authority (2010) reported it is negligible for 1,2,4-triazole.

The applicability domains of the QS(A)PRs restrain their reliability, meaning that only the predictions that fall within the AD can be considered reliable. An informative review about the validation of QSARs was written by Gramatica (2007). She states that when the leverage value of a compound is lower than the critical value (which is depending on number of model variables and the number of

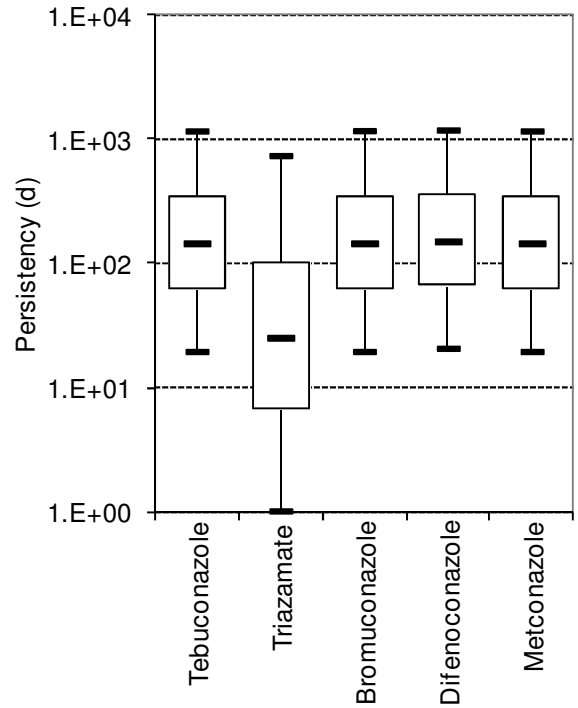
the objects used to calculate the model), the probability of accordance between predicted and actual values is as high as that for the training set chemicals. **Table S2** (Appendix 4) shows that most QS(A)PR predictions were within the AD, except for the water solubility prediction for Triazemate, and the hydroxyl radical reaction rate in air for Difenoconazole. For Bromuconazole, the k_{OH} prediction was just outside the border of the AD. Whether the triazoles of this study are within the applicability domain of the QS(A)PR models does not directly influence the predictions themselves, but a prediction that is outside the model's AD has a higher uncertainty than what was calculated in this study. As stated by Nikolova and Jaworska (2004), it is a warning for model applicability, but not a final decision on prediction quality. In principle, there are two options. One could still judge that the QS(A)PR model gives a reliable outcome. This requires a decision on how to treat the extra uncertainty that is caused by being outside the AD, which is a research area that is still in development. A (hypothesized) mechanistic understanding of the modeled property could be a start to decide the best treatment for the extra uncertainty (Nikolova and Jaworska, 2004). Furthermore, one could judge that the outcome of the QS(A)PR model is not reliable and cannot be used. In that case, either a better QS(A)PR or experimental data are required. Despite of one water solubility prediction and two k_{OH} predictions outside the AD, we think this risk assessment performed in this study is reliable. After all, the results of the sensitivity analysis showed the uncertainty in these parameters has negligible influence on the uncertainty of the outcome variables.

The PNEC calculations were based on a fixed assessment factor of 1000, because only a small dataset of acute toxicity predictions was available. The assessment factor should be applied on the lowest L(E)C50 value. It accounts for the intra- and inter-species variations; intra- and inter-laboratory variation of toxicity data; short-term to long-term toxicity extrapolation; and laboratory data to field impact extrapolation (e.g. multi-substance effects)(European Commission, 2003b). Since the assessment factor accounts for the uncertainty inherent in acute toxicity data (i.e. intra- and inter-species variations), and we also applied a Monte Carlo simulation to include the uncertainty in the QSAR model predictions, the calculated PNEC may be in this case be interpreted as a worst-case value. However, in case of one short-term L(E)C50 from each of three trophic levels of the base-set, variation from a factor of 1000 should not be regarded as normal and should be fully supported by accompanying evidence (European Commission, 2003b).

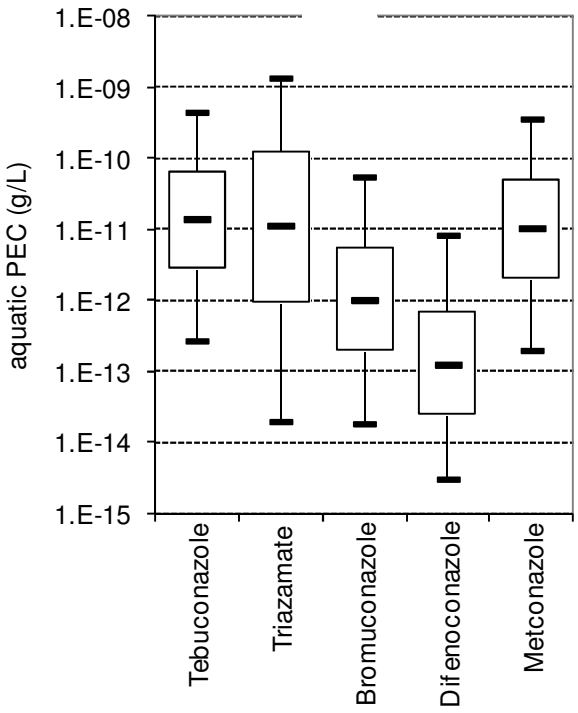
a)


1.E+00
1.E-01
1.E-02
1.E-03
1.E-04
1.E-05
1.E-06
1.E-07
LRTP (-)
Tebuconazole
Triazamate
Bromuconazole
Difenoconazole
Metconazole

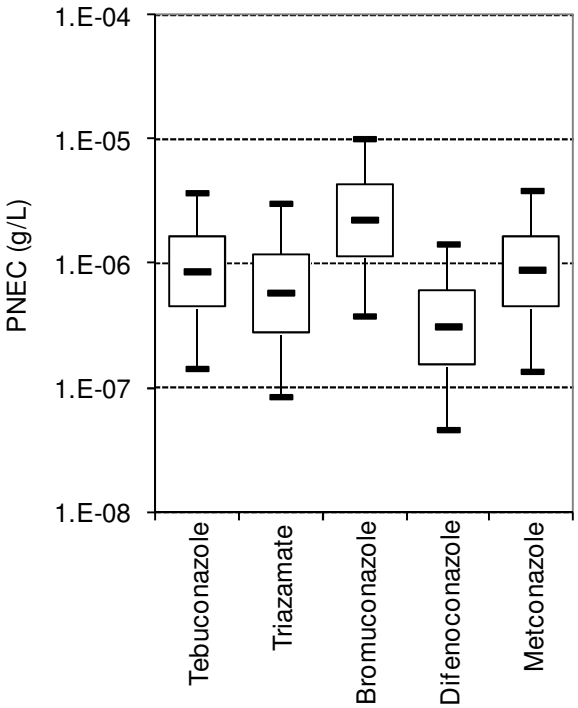
b)


1.E+04
1.E+03
1.E+02
1.E+01
1.E+00
Persistency (d)
Tebuconazole
Triazamate
Bromuconazole
Difenoconazole
Metconazole

c)


aquatic PEC (g/L)
1.E-08
1.E-09
1.E-10
1.E-11
1.E-12
1.E-13
1.E-14
1.E-15
Tebuconazole
Triazamate
Bromuconazole
Difenoconazole
Metconazole

d)


PNEC (g/L)
1.E-04
1.E-05
1.E-06
1.E-07
1.E-08
Tebuconazole
Triazamate
Bromuconazole
Difenoconazole
Metconazole

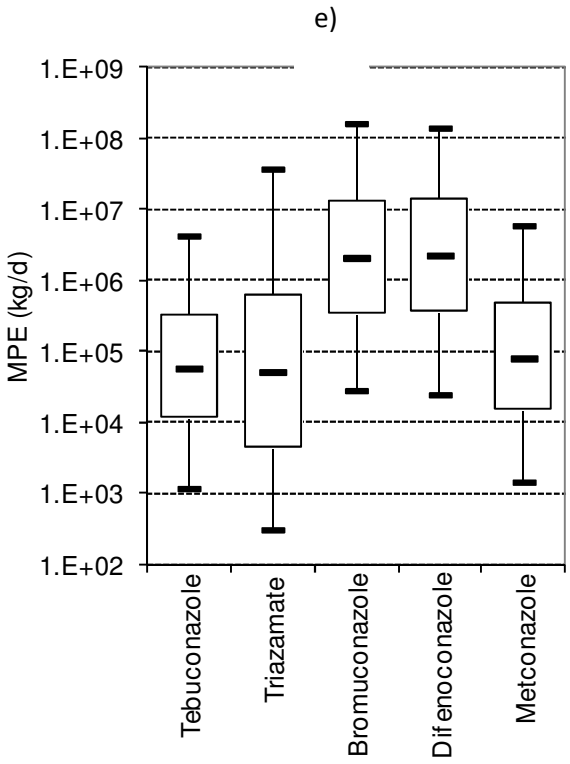


Figure 3.1. Box plots of the (a) dimensionless potential for long-range transport (LRTP), (b) Persistency in days, (c) aquatic PEC in g/L, (d) PNEC in g/L, and (e) maximum permissible emission to agricultural soil (MPE in kg/day), for Tebuconazole, Triazamate, Bromuconazole, Difenoconazole, and Metconazole. The columns represent the 25th and 75th percentile, and the whiskers the 5th and 95th percentiles. In the columns, the median value is marked.

3.4.2. Interpretation of results

The uncertainties in the soil sorption partition coefficient, biodegradation rate in water, and toxic concentrations contributed to the uncertainty in the MPE for an emission to agricultural soil. The predicted K_{oc} of Triazemate is the lowest in this study ($5.87 \cdot 10^{1}$) with a relative contribution to variance of the MPE of 11 percent, whereas the other four triazoles have a higer K_{oc} (>$5.70 \cdot 10^{2}$) and a higher relative contribution to variance of the MPE (>43 percent). The K_{oc} is an important property of triazoles, because the high sorption to soil organic matter is probably responsible for the limited movement and leaching from the soil (Kim et al., 2002). Moreover, sorption to soil organic matter could also explain the moderate soil longevity, since it makes the chemical less available for micro-organisms to degrade. Nevertheless, soil microorganisms degrade triazole fungicides in soil, as reported for ipconazole by Eizuka et al. (2003). In accordance with that, the half-live time in water was the smallest for Triazemate, with accompanying high biodegradation rates in soil and sediment. However, the uncertainty was large. This is also reflected in a relative contribution to variance of the MPE of >82 percent for $k_{biodeg,water}$. As a consequence, the persistency, the aquatic PEC, and the MPE have the largest 90%-CI for Triazemate.

The predicted PNEC was based on three multiple linear regressions, i.e. for the LC50 of *Onchorynchus Mykiss*, for the EC50 of *Daphnia Magna*, and for the EC50 of *Pseudokirchneriella Subcapitata*.

Typically, *P. Subcapitata* was the most sensitive species for three out of the five triazoles, and *O. Mykiss* for the other two triazoles. However, the uncertainty distribution of the PNEC is not the equivalent of the uncertainty distribution of the most sensitive species. For the most sensitive species, the contribution to variance in the PNEC ranged from 68.5 percent (Bromuconazole in *O. Mykiss*) to 79.7 (Metconazole in *P. Subcapitata*). Furthermore, in four out of five triazoles, the EC50 of *D. Magna* also had a minor influence on the variance in the PNEC and MPE for agricultural soil. These findings emphasize the importance of including species of different trophic levels, rather than choosing one sensitive species.

3.4.3. Conclusion

We studied the influence of the use of QS(A)PRs on the uncertainty in the outcome of a risk assessment for triazoles, and determine the relative contribution of the different predictive models to the overall uncertainty. The typical maximum permissible emissions to agricultural soil were highest for Bromuconazole and Difenoconazole, i.e. 2.09 ·106 and 2.26 ·106 kg/day, respectively, with 90%-CIs of four orders of magnitude. For Tebuconazole, Triazemate, and Metconazole we found lower typical MPEs, that is between 5.15 $\cdot 10^4$ and 8.00 $\cdot 10^4$ kg/day, with 90%-CIs ranging three to five orders of magnitude. We found that the uncertainty of the maximum permissible emission to agricultural soil was mainly determined by uncertainty in the QSPR soil sorption partition coefficient, in the QSPR for biodegradation in water, and in the QSAR for toxicity to different species. In this case three predictions were outside the applicability domain of the QSPR. Nevertheless, we think the risk assessment performed in this study is reliable, because the results of the sensitivity analysis showed the uncertainty in these parameters has negligible influence on the uncertainty of the MPE.

Supplementary data – the supplementary data provides the predictions of the input parameters with their predictive error (or geometric mean and standard deviation for the half-lives in water) for all triazoles in this study and the assigned distribution. It also compares QS(A)PR predictions with their applicability domain.

4. CASE-STUDY 2: Non-testing versus testing based risk assessments on three PBDEs

4.1. Summary

Exposure (PEC) and effect (PNEC) were assessed based on QSPR and QSAR predictions for three PBDEs. Whenever available, parameters for the exposure assessment and species effect concentrations were replaced by experimentally tested information. Unfortunately, did none of the three chosen PBDEs have testing versus non-testing based assessments of exposure and effect, only one at a time. QSPR-based exposure assessments were more uncertain, compared to when experimental values were used, which resulted in more conservative risk estimates. QSAR-based effect assessment were less uncertain since it were based on three instead of just two species, and therefore less penalization for uncertainty were added (as safety factors). This resulted in a less conservative risk estimate. Our conclusions is that non-testing versus testing based risk assessments are different, but this difference depend to a large extent to how uncertainty is dealt with. As long as the predictions are precise (i.e. with a good coverage of the experimental values) non-testing information are useful complement to reduce uncertainty in existing testing information of effects. This case-study did not consider uncertainty in tested physic-chemical properties, not because it does not exist, but because that uncertainty is not included in QSAR data. The uncertainty in QSAR predictions were derived from predictive inference based on the information in the underlying QSAR data. Besides predictive uncertainty, the reliability in predictions were evaluated for the QSPR models based on what was known about the so called applicability domain.

4.2. Introduction

Replacing testing information with non-testing information to support decision making must be done with care. Requirement of strong reliability in risk assessment depends on the consequences of the decision made. For example, there is a difference when QSAR[5] predictions are used to design experiments or to generate hypothesis about possible mechanics, compared to when QSAR predictions are used as weight-of-evidence replacing experimental testing information in risk assessment or waiving. REACH complies with the 3R philosophy for animal welfare that is to Replace, Refine and Reduce experimental testing on animals. The aim of these R's can be enhanced by studies aimed to address the reliability in replacing testing with non-testing information in risk assessment. Replacing testing with non-testing information may introduce an error in assessed risk, which in turn may lead to less safe, or unnecessary strict, regulatory decisions (the former worse than the latter). Decisions may be improved by considering the uncertainty in non-testing information. That is why one of the goals of the project CADASTER was to characterize the uncertainty in QSAR predictions, with the purpose to address the question that by considering uncertainty in the predictions may increase the reliability in non-testing information in risk assessment.

The objective of this case-study is to demonstrate the application of QSARs and QSPRs in probabilistic risk assessment, and the evaluation of reliability in using such non-testing information instead of testing information in regulatory decisions. The use of non-testing information is limited to methods for which uncertainty have been quantified. That is why uncertainty analysis was restricted to chemical specific parameters for which the majority had been predicted by QSPRs and QSARs.

---

[5] QSAR is often used as a general term including QSPR

Uncertainty in assessments for decision making is formed by subjective beliefs relating to the background knowledge, and it means that uncertainty is not quantities that exist independent of the method for measurement. Instead, uncertainty is given a treatment, which means to identify, quantify and respond to when making decisions.

A scientific approach to evaluate the reliability in non-testing information requires an established principle for predictions and thereby a sound basis for quantifying uncertainty in QSAR predictions. Predicting is hard and in this case-study we assume some ideal conditions such as that the conditions for predictive inference are fulfilled for every model. Reliability in non-testing information is assessed in retrospect by studying the consequences in real applications where both kinds of information are present.

Here we demonstrate this on probabilistic risk assessment using the framework for QSAR/QSPR based probabilistic risk assessment developed in CADASTER on three polybrominated diphenyl ethers (PBDEs), which belong to of the CADASTER chemical classes. The three selected PBDEs were BDE-03(4-monoBDE), BDE-28(2,4,4'-TriBDE) and BDE-47(2,2',4,4'-TetraBDE) and available experimental data were sought for as many of the QSAR /QSPR predicted input parameters as possible. PBDEs belong to an emerging class of organic pollutants widely used, especially in the past, as flame retardants in a variety of consumer products. PBDEs potentially include 209 congeners divided into 10 congeneric groups (mono- to decabromodiphenyl ethers).

4.3. Exposure assessment

Environmental fate of the three PBDEs was calculated using the multimedia fate model SimpleBox (Den Hollander et al., 2004) for a unit emission to air at the regional scale. Non-testing information was provided by QSPRs of chemical-specific properties at 25°C, of which some have been developed or specified in Work Package 3 in CADASTER. QSPR predictions were used to specify Simplebox input parameters which are water solubility (S, mg/L) (Papa et al., 2009), melting point (Tm, °C) (Papa et al., 2009), vapor pressure (Vp, Pa) (Papa et al., 2009), organic carbon - water partition coefficient (Koc, L/kg) (Gramatica et al., 2007) and hydroxyl radical reaction rate (k_{OH} , cm3/s.molecule) (Roy et al., 2011). The QSPRs for K_{oc} and k_{OH} are given in Table 3.1,, whereas QSPRs for Tm, S and Vp are reported in the Table 4.1.

Table 4.1. Selection of QSPRs for physico-chemical properties of PBDEs at 25°C from(Papa, Kovarich et al. 2009)

Parameters	units	R^2 %	Model description
Melting point (T_m)	°C	84.37	T_m = 1968.06 – 6227.09 X2A
Water solubility (S)	mol/L	91.80	log 1/S = 6.09 – 1.18 Mor23m
Vapor pressure (V_p)	Pa	98.71	log 1/V_p = 0.115 + 0.213 T(O…Br)

X2A = average connectivity index chi-2
Mor23m = Morse signal no 23 weighted by atomic masses
T(O…Br) = sum of topological distances between oxygen and bromine atoms

Biodegradation in water was predicted by the ultimate biodegradation estimation model, BIOWIN3, included in the estimation software EPI Suite™ (Boethling et al., 1994)The BIOWIN3 classifies a

compound into a biodegradation category based on molecular fragments. However, Simplebox uses biodegradation half-live in surface waters (τ_{wat}, days) as input parameter and not a category. Therefore, values for half-live was assigned based experimental data on half-lives collected and described for each of the eight BIOWIN3 categories (Aronson et al., 2006) and used the half-life for each PBDE falling in the relevant category. Half-lives in sediments and soils were assessed from the half-lives in water by assuming it to be two times higher in soils and nine times higher in sediments, respectively as it is commonly implemented in EPI Suite™. Models to assess fate of PBDEs have for long only considered the OH reaction rate constants in the gas phase (Wania and Dugani, 2003; Gouin and Harner, 2003), but the photolysis in the atmosphere have been suggested to be a critical parameter in the assessment of PBDEs (Raff and Hites, 2007; Schenker et al., 2008; Eriksson et al., 2004). Photolysis of PBDEs in air has been seen in laboratory experiments only, but since field studies are difficult, there are no experimental data available so far for photolysis that correspond to field conditions. Under these laboratory conditions, a QSAR for the photolytic half-life in air ($\tau_{photo,}$ 1/s) were fitted by linear regression on the adsorption spectra and quantum yield measurements in the atmosphere of two compounds Di-BDE3 and Tri-BDE-7, resulting in a negative slope with increasing homologues of PBDEs (Raff and Hites, 2007). This regression is based on a simplification and have been applied in risk assessment where 209 PBDE congeners were grouped in homologues that made it easier to handle the photo-degradation in the fate assessment (Schenker et al., 2008). The predictions of the regression model were observed within the estimates of photolytic rates for PBDEs.

4.3.1. Reliability in QSPR predictions

The models in Table 4.1 fulfill the OECD principles (OECD, 2007). The third OECD principle states that a model should have a well-defined domain of applicability (AD). Here we ask what it means when applying the model in risk assessment. Following the practice suggested by several authors (Papa et al., 2009; Eriksson et al., 2003), reliability in using these five QSPR models to predict the three compounds under consideration were judged by the leverage approach (Table 4.2). This means that leverage is calculated from model descriptors for the training data set and the selected BDE in question as the sum of the diagonal of the hat matrix. A leverage value can geometrically be seen as a distance in space spanned by the descriptors, and is a measure of the extent of extrapolation.

In order to judge whether an item is close enough, it has been suggested to compare leverage value to a cutoff $c \cdot p/n$ where p is the number of model descriptors (including the intercept), n is the number of points in the training data set, and c is a constant. In Table 4.2 the cutoff was defined by setting c = 3, as suggested by Gramatica (2010). A well-established software for chemoinformatics uses c = 3 as a default, but say that any value between 1 and 10 are possible (Martens and Næs, 1989). However, a clear cut value on c may cause problems in practical applications, especially since compounds to predict quite often are distant to the model, and in a sense close to being extrapolated.

Based on c = 3, the three PBDEs fell inside the AD for the models predicting S, Tm, Vp and Koc. The QSPR of K_{OH} had been trained on volatile organic compounds, and it has been shown that most PBDEs are not found inside the AD (Gramatica et al., 2004). However, the predictions showed a good agreement with the predictions obtained from EPI Suite verified by Roy et al (2011) showing the difference in the predictions for PBDEs was within 0.8 log unit. The predictions from EPI Suite and

the QSPR model for k_{OH} were similar but the crucial information on AD for chemicals was an advantageous aspect. Given that the PBDEs were outside the AD for k_{OH}, the use of these predictions also needs to be evaluated in light of the sensitivity of the assessed risk to the input parameter k_{OH}.

Table 4.2. Assessment of Applicability Domain of PBDEs using Leverage approach. Bold values means that a compound is outside the AD for a value on c = 3.

Parameters	S	Tm	Vp	K_{oc}	k_{OH}
Cut off	0.500	0.240	0.180	0.023	0.032
BDE-03	0.229	0.181	0.110	0.003	**0.040**
BDE- 28	0.191	0.108	0.031	0.006	**0.045**
BDE- 47	0.115	0.071	0.031	0.007	**0.057**

Concerning the applicability domain of BIOWIN in EPI Suite, there is no well-defined criterion about the reliability of the predictions. It is only mentioned there about the model domain that the user may wish to consider the possibility that biodegradability estimates are less accurate for the compounds outside the molecular weight (MW) range of the training set compounds.

4.3.2. Uncertainty in QSPR predictions

A statistical and quantitative approach to assess the error in a prediction involves predictive inference. The Bayesian framework for predictive inference is despite the problems of predicting in general (see Appendix 1) pointed out as the most robust and reliable framework that quantify the uncertainty in a prediction by a probability distribution. Bayesian inference provides an output with an interpretation that is in agreement with the interpretation of risk assessor, but most importantly decision makers. However, Bayesian inference is not the dominating statistical principle in QSAR modeling, and it does not have to be. Bayesian inference is useful when applying QSARs in decision making, such as to support uncertainty analysis in chemical safety assessment. The step from a documented QSAR model to an applied situation using predictive inference has been identified in WP4 CADASTER as crucial to integrate QSARs in risk assessment and will be addressed in the upcoming deliverable D4.2. It can be shown that when uncertainty in a QSPR prediction from a regression fitted by OLS is assigned a t-distribution defined by the predicted point estimate and predictive error (see Eq 6), well approximates corresponding predictive distribution from Bayesian inference under certain conditions (Appendix 2). Predictive means and predictive errors for the physic-chemical properties and corresponding degrees of freedom were here used to define the uncertainty in those input parameters predicted by OLS regression (Table 4.3).

Table 4.3. Quantification of uncertainty in physico-chemical properties, atmospheric degradation rates and biodegradation half-lives at 25°C for the selected PBDEs.

PBDE	T_m [a] (°C)		log S [a] (mol/L)		log V_p [a] (Pa)		log K_{oc} [b] (L/kg)		log k_{OH} [c] (cm^3s^{-1} per Molecule)		Biodegradation half-lives in water [d] (τ_w, days)		Photolytic degradation rate [e] (k_{photo},1/s)
n / p [f]	25/1		12/ 1		34/1		643/4		460/4		Qualitative model	Lognormal distribution Based on experimental data	2/1
	PRED	SEP	PRED	SEP	PRED	SEP	PRED	SEP	PRED	SEP	BIOWIN3 Category [g]	(M, CV)	
BDE-03	43.88	21.44	–6.73	0.27	–1.18	0.17	3.55	0.56	–11.42	0.44	Weeks-months	(20, 7.45)	2.09E-06
BDE-28	68.79	20.77	–6.97	0.26	–2.88	0.16	4.10	0.56	–11.97	0.44	Months	(85,1.96)	1.34E-05
BDE-47	87.47	20.42	–7.35	0.25	–3.52	0.16	4.34	0.56	–12.26	0.45	Recalcitrant	(88,1.91)	3.38E-05

[a] Papa et al. (Papa et al., 2009), [b] Gramatica et al. (Gramatica et al., 2007), [c] Roy et al. (Roy et al., 2011)
[d] (Aronson et al., 2006)
[e] (Raff and Hites 2007)
[f] n number of compounds in training data, p number of descriptors for which degrees of freedom in a t-distribution is n-p-1
[g] Division of recalcitrant category with respect to BIOWIN output ; M = median; CV = coefficient of variance
PRED: predictive mean, SEP: predictive error, M: median, CV: coefficient of variation

Default values were given to the all parameters in the Simplebox model relevant for exposure in aquatic compartment, except for half-life for biodegradation and photolytic rate. These two input parameters were specified by expert judgment supported by QSPR predictions and experimental data. The uncertainty in half-life for biodegradation was assigned by a lognormal distribution based on analysis of experimental data by Aronson et al. (Aronson et al., 2006) that were used to revise the categories of biodegradability in the predictions of biodegradation BIOWIN3 (in EPISUITETM) (Table 4.3). A correlation of 1.0 was assumed among the half-lives in water, soil and sediment in order to quantify the uncertainty in soil and sediment by multiplying with factor 2 and 9 respectively.

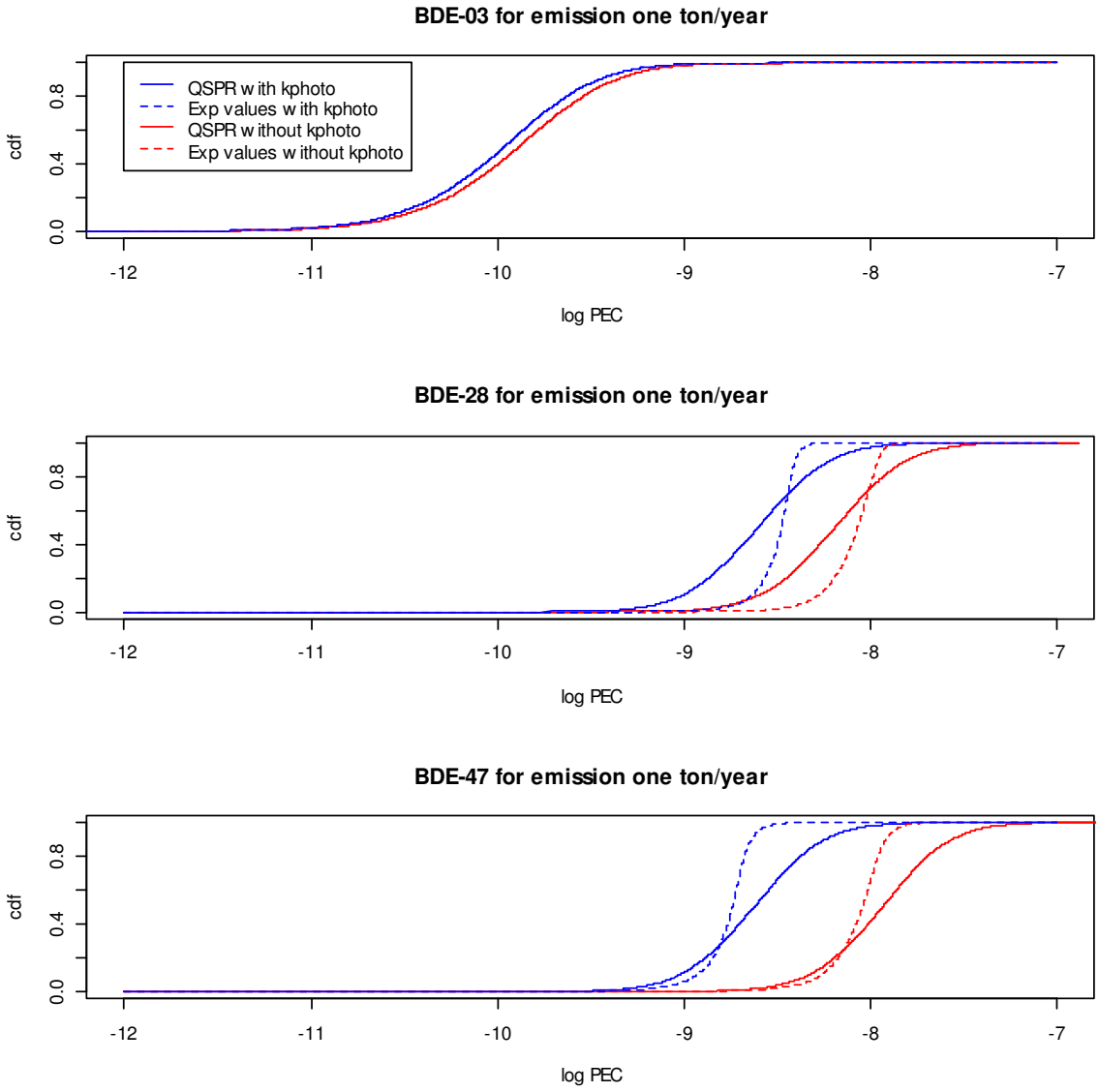


Figure 4.1. Comparison of log PEC (mg/l) for non-testing (QSPR predictions) and available testing information (experimental data instead of QSPR predictions when possible) with and without photolysis in fresh water (based on a unit emission in ton/year).

Photolysis was assigned a single value, which was regarded as an upper bound. This was motivated by seeing that the prediction by Raff and Hites was based on a model fitted to upper limits of photolysis estimates and that half-life decrease with the number of bromine atoms increase. No half-life of PBDE was used as a conservative lower bound. In this way the uncertainty in half-lives was quantified as an interval stating that the actual value can fall anywhere inside the interval with no specifications of one value being more likely than another. In practice two different risk assessments were done, one with photolysis at a large value and one without photolysis. Predicted Environmental Concentration (PEC) was obtained as a pair of cumulative probability distributions (CDFs), and it was verified by further simulations that these CDFs form a probability box (p-box) bounding all possible probability distributions generated by values on photolysis within this interval.

4.3.3. Sensitivity analysis of exposure

Sensitivity analysis was performed with the purpose to compare non-testing based risk assessment with experimental based risk assessment. This sensitivity analysis was performed on the two scenarios for photolysis which were the upper and lower bounds of the resulting p-box. Input parameters for which such information was available were described by predictive distributions from both QSPR and QSAR, and experimental data from direct tests on the compound in question. There were physico-chemical experimental data available for BDE-28 (Tm 64.25; log S -6.76; log Vp -2.8) and BDE-47 (Tm 82.58; log S -7.51; log Vp -3.5) since these two compounds had been present in the training or validation data set used to develop some of the QSPR models. Since experimental data available on melting point, solubility and vapor pressure were well covered by the corresponding predictive distributions, the reason to why uncertainty distributions based on QSPR predictions were wider in comparison to testing based distributions (Table S3 in Appendix 4), was that uncertainty was not considered in testing information. Uncertainty in experimental data may be large. Uncertainty in experimental data has been raised as an important aspect in QSAR modeling that potentially may improve models predictivity and reliability.

4.4. Effect assessment

A model to handle uncertainty in input data for SSD is under development. Meanwhile we fitted SSDs to point predictions. For BDEs there were three QSAR models available to predict acute fish, daphnid and algae LC50s and three QSARs predicting fish daphnid and algae chronic effects values (ChV, i.e. the geometric mean between the NOEC and LOEC) available from ECOSAR (Table S4 in Appendix 4). None of the three BDEs were included in the training data set for ECOSAR, and testing information to assess the effect in aquatic environment was searched for in other sources. Experimental acute values were found for BDE-03 on the species Daphnid and Fish, BDE-28 effect data was found for Nitocra spinipes, and BDE-47 effect data was found for Fundulus heterocliticus (Table S4 in Appendix 4). PNEC values were calculated by adding uncertainty factors when needed (Table S4 in Appendix 4).

The accepted procedure to arrive at a PNEC based on acute toxicity values is to take the lowest of the three species groups and divide the LC50 by 1000. This assessment factor is made up of a general acute-to-chronic ratio of 100 plus a safety factor of 10 to account for the fact that only 3 species are sampled.

Generation of a PNEC from chronic toxicity values can be performed by calculation of an SSD (based on the chronic NOEC values) and derivation of the HC5 from that SSD. When the SSD is reliable (adequate species variation, nr. of experimental data, distribution of the data close to log normal) this HC5 is used directly in the risk assessment as the PNEC. So the PNEC concentration is expected to protect 95% of all species in the environment, but at PNEC concentrations, i.e. at RCR 1 chronic (non-lethal) effects can still be expected for 5% of the species. The HC5 (the 5th percentile of a normal SSD) is estimated using the mean, SD and number of measurements for the distribution, in original log concentration units:

$$HC5 = Mean - k_s \cdot SD \quad (7)$$

where k_s, the extrapolation constant is 1.938 for n=3 and 2.339 for n=2.

When the SSD is less then optimal a safety factor ranging from 1 to 5 is applied to the HC5 to arrive at a PNEC for risk assessment. In our example the SSDs are based on only 2 or 3 values making the estimation of the distribution more uncertain and a maximum safety factor of 5 will be required.

4.4.1 Sensitivity analysis on effect

BDE-003

The comparison of the PNEC distribution based on experimental acute values (n=2), QSAR predicted acute values (n=3) and predicted chronic values (n=3) is shown in the following table and figure for BDE-003:

BDE-003 PNEC distributions

	ACR[a]	SF[b]	log PNEC[c]	Mean	SD	n
Acute experiment based	100	5	-4.54	-2.54	0.86	2
Acute QSAR based	100	5	-3.20	-2.97	0.12	3
Chronic QSAR based	1	5	-2.64	-1.73	0.47	3

[a] Acute-to-Chronic ratio. 100 is the standard used in EU risk assessment

[b] Safety Factor. Safety factory ranges from 1-5 see text. Maximum safety factor is used as all SSDs are based on too few measurements.

The PNEC distribution based on the 2 acute experimental values has a relatively large standard deviation, and the extrapolation constant used to estimate the 5th percentile gives a (5th percentile) PNEC of -4.54 = 3E-05 mg/l = 0.03 microg/l. The uncertainty in this estimate is large, indicated by the shallow slope of the PNEC distribution in the figure above.

The PNEC distribution based on the acute toxicity QSAR estimates has a relatively small standard deviation, hence the 5th percentile of the PNEC is close to the mean. However, the acute to chronic ratio of 100 and the safety factor of 5 (for the SSD approach) still give a conservative estimate of the 5th percentile PNEC estimate of -3.20 = 6.3E-04 mg/l = 0.63 microgr/l. Estimating the PNEC using the lowest of the three acute values and dividing by 1000 would still give a slightly more conservative estimate of the PNEC of 0.43 microgr/l.

The PNEC distribution based on the chronic toxicity QSAR estimates is more uncertain (higher standard deviation of the SSD curve) than the curve based on the acute QSAR estimates, but as the conservative ACR of 100 is not required, it still gives the least conservative, most realistic worst case 5th percentile of the log PNEC of -2.64 = 2.3E-03 mg/l = 2.3 microgr/l

As a calculation of an SSD (and the distribution of the PNEC based on this SSD) is not possible based using only one value (no SD can be calculated) the comparison for BDE-028 and BDE-047 will only give PNEC distributions for acute and chronic based QSARs, and a comparison with the experimental toxicity value can only be made absolute and not probabilistic.

BDE-028

BDE-028 PNEC distributions

	ACR	SF	log PNEC	Mean	SD	n
Acute QSAR based	100	5	-3.94	-3.48	0.24	3
Chronic QSAR based	1	5	-3.38	-2.19	0.61	3

[a] Acute-to-Chronic ratio. 100 is the standard used in EU risk assessment

[b] Safety Factor. Safety factory ranges from 1-5 see text. Maximum safety factor is used as all SSDs are based on too few measurements.

BDE-047

BDE-047 PNEC distributions

	ACR	SF	log PNEC	Mean	SD	n
Acute QSAR based	100	5	-4.34	-4.29	0.03	3
Chronic QSAR based	1	5	-4.24	-2.76	0.76	3

[a] Acute-to-Chronic ratio. 100 is the standard used in EU risk assessment

[b] Safety Factor. Safety factory ranges from 1-5 see text. Maximum safety factor is used as all SSDs are based on too few measurements.

When comparing the PNEC distributions for the three BDE's using the SSDs based on the chronic toxicity estimates a logical trend can be observed where the PNEC becomes lower for the higher brominated diphenyl ethers, but also the uncertainty in the PNEC distribution becomes larger as the SD on which the distribution is based also increases with higher bromination of the diphenyl ethers. This is clear in the following figure 4.2 where the three SSD based PNEC distributions for the three BDE's are given on top of each other.

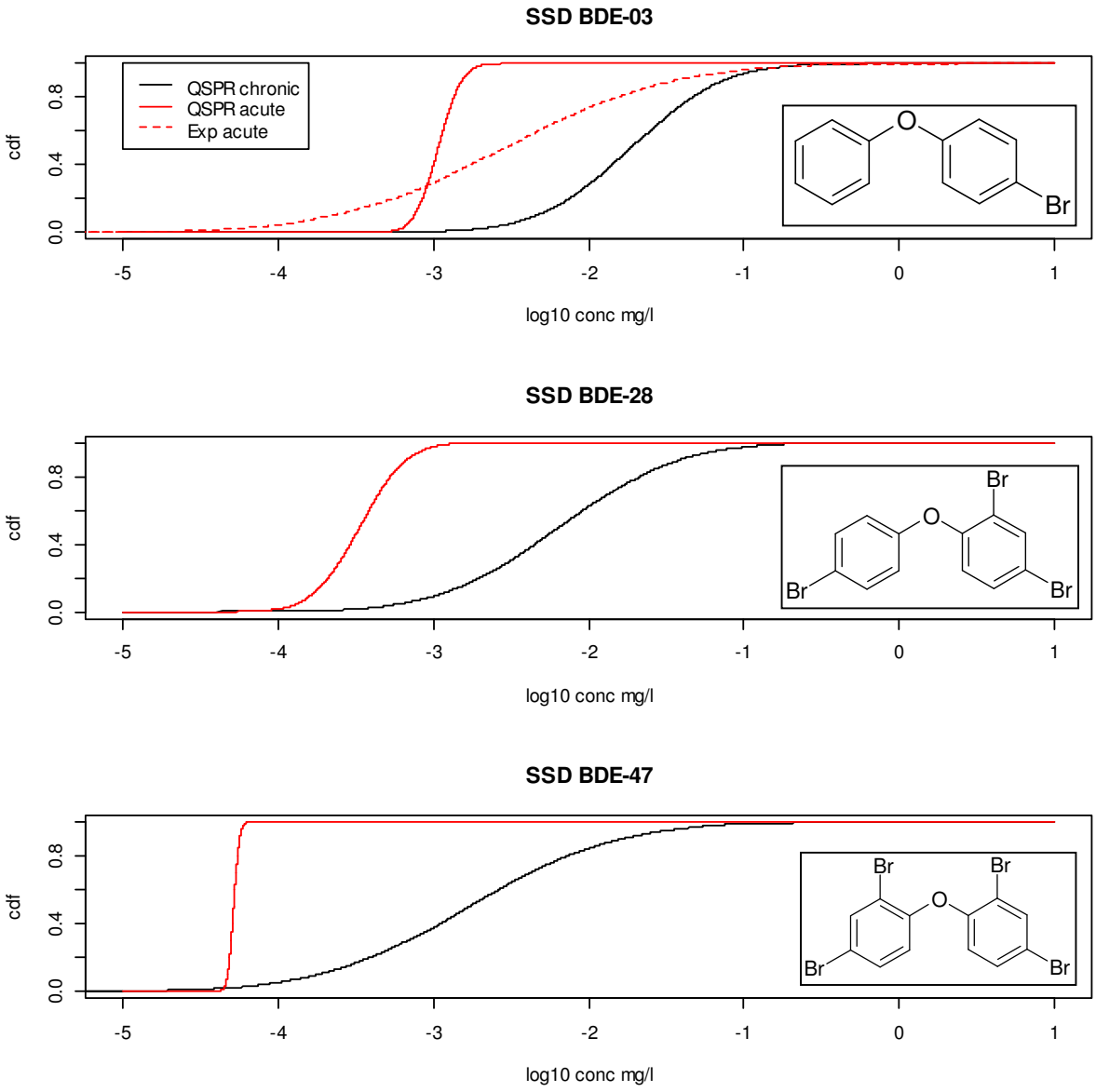


Figure 4.2. Species Sensitivity Distribution based on QSAR predictions of acute and chronic effects, and experimental values on acute effect for BDE-03.

4.5. Risk assessment

Environmental risks are typically estimated in a deterministic way using single point estimates for both exposure (PEC) and effects (PNEC). For uncertainty analysis these point estimates (PEC and PNEC) are replaced by probability distributions. Uncertainty in PEC is derived by Monte Carlo simulation of the Simplebox model based on the specified uncertainty in input parameters. PNEC are recommended to be derived as a Species Sensitivity Distribution (SSD) in order to capture the variability between species in the ecological system. The probabilistic PEC and PNEC (here derived for an aquatic environment) are

now combined into a quantitative risk measure. The deterministic measure Risk Characterization Ratio (RCR) is hampered by not being a measure of risk - it is sensitive to scaling and is therefore not comparable between substances. A probabilistic measure of risk is the probability of an undesired effect P(PEC > PNEC). This probability can alternatively be expressed as the Expected Risk (Figure 4.3.), which is the expected fraction of species affected for an uncertain exposure (Appendix TOM).

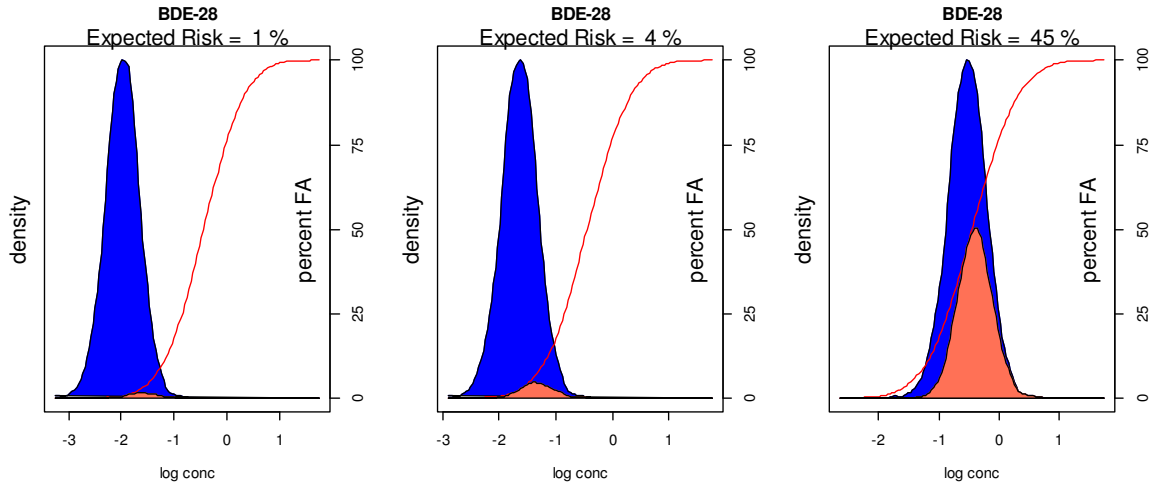


Figure 4.3. Expected risk is a measure to what extent the PEC and PNEC distributions overlap, has a clear interpreation in terms of the expected fraction of species affected, and is invariant to scale which facilitates comparison between different risk assessments.

4.5.1. Sensitivity analysis on Expected Risk

Risk depends on the rate of emission of the chemical into different compartments. Expected Risk was calculated on a range of alternative emission scenarios in air on a regional scale. Assuming that emission scenarios resulting in a risk less than 5% are regarded as safe, non-testing-based ER were compared to testing-based ER by searching for discrepancies in the following regulatory decision for different emission scenarios. Here the results based on PEC assessed without considering photolytic rate into air was used.

As seen in Figure 4.4 and 4.5 there are emission rates for which there is a discrepancy between non-testing and testing based probabilistic risk assessments. For example, an emission of 6 tons BDE-03 into air per year would be regarded as safe if based on QSAR predictions of acute effect, while regulatory actions would be necessary if the decision had been based on experimental tests of acute effects (Figure 4.4). In this case, the uncertainty in testing-based SSD is wider because of 2 instead of 3 species is used. QSAR predictions could have narrowed the SSD, producing a less conservative risk estimate.

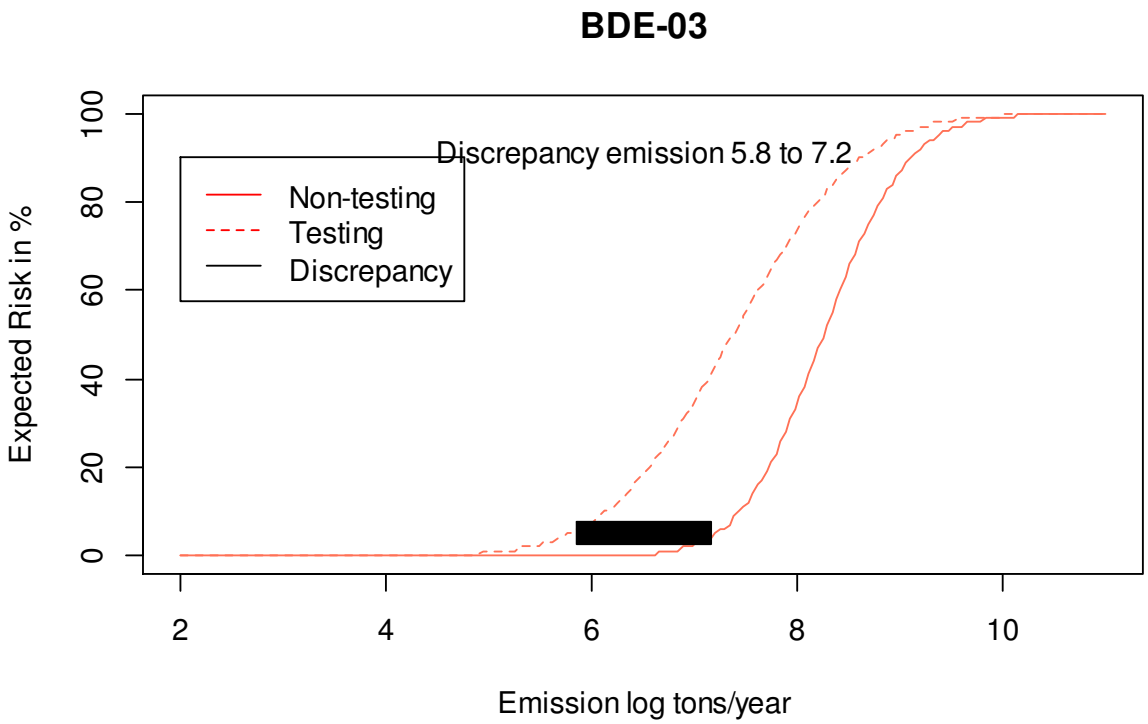


Figure 4.4. Discrepancy in regulatory decision when Expected Risk is derived from testing versus non-testing information of acute effects on PNEC for BDE-03.

An emission of BDE-28 of 6 tons would be regulated when based on QSPR predicted physico-chemical properties, while regarded as safe when based on testing information when available (Figure 4.5). Thus, the wider uncertainty in QSPR-based exposure level compared to exposure where some predictions had been replaced by experimental values (Figure 4.1), result in a more conservative risk estimate.

4.6. Conclusions

This case-study exemplifies the regulatory consequences of using non-testing information in the absence of testing information, but can also be seen as the consequences of combining non-testing information with weak testing information as a weight-of-evidence approach. A small discrepancy between non-testing and testing based risk assessment may not only be an effect of accurate QSAR predictions. For example, when the influence of a single parameter is small in comparison to other parameters in the assessment, whether testing or non-testing information is used does not make a large difference on the risk. Identifying which parameters with a large influence on risk or its uncertainty can be approached by different kind of sensitivity analysis, such as the one in the case-study of Triazoles. QSAR uncertainty needs to be put in perspective to other uncertainties. The exposure assessment of the three PBDEs shows that the influence of QSPR predicted parameters are small in comparison to whether or not photolytic rate in air should be considered. In order to generalize the impact of non-testing information provided by QSARs in chemical risk assessment the approach described here will be done on a larger set

of chemicals carefully selected to represent chemical space by experimental design (Appendix 4). General conclusions on the reliability of non-testing information in risk assessment will be difficult to make, since the importance of different sources of information depends on each other, the context for the assessment and the decisions made.

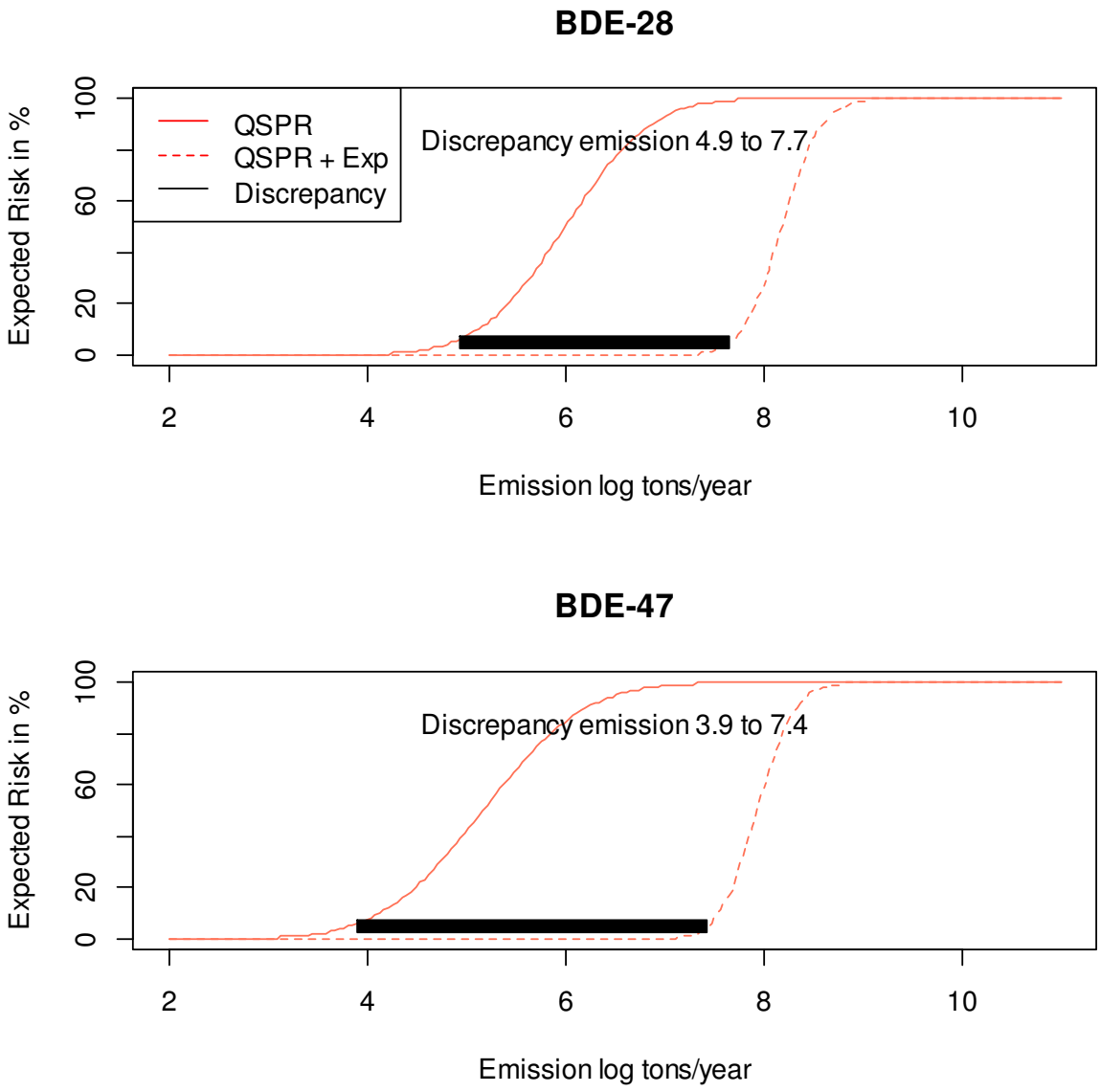


Figure 4.5. Discrepancy in regulatory decision when Expected Risk is derived from testing versus non-testing information on PEC for BDE-28 and BDE-47.

5. Conclusions and future outlook

Risk assessment is a tool to describe uncertainty in unknown quantities such as risk following the release of a chemical substance into the environment. Focus in this report has therefore been on the characterization and propagation of uncertainties relevant for the integration of QSARs into probabilistic risk assessment. A general prerequisite is that **treatment of uncertainty is context dependent**, and should be **interpreted in relation to the background information**. Integration of QSARs in chemical safety assessment is an example where it is obvious that aspects of uncertainty are linked to the background knowledge, since we have the option to enlarge background knowledge by further testing.

The two case-studies demonstrate the computational framework for QSAR based risk assessment. The application of QSARs in probabilistic risk assessment requires the answers to the questions:

- Are there any QSAR data available to use as weight-of-evidence of a chemico-specific input parameter?
- Which algorithm for prediction and approach for predictive inference should be used?
- Is a QSAR prediction reliable enough to support the intended decision making?

Predictive uncertainty assessed by predictive inference is precise in the sense that it, given necessary assumptions for predictive inference are fulfilled, covers the actual value with a certain degree of confidence. Precise predictions are valuable in design of testing strategies. Uncertainty of predictions assessed by predictive inference can be seen as enough when a prediction is judged to have high reliability. When predictive reliability is evaluated as low, such as when the extent of extrapolation in a prediction is evaluated as unacceptably high, the uncertainty in a QSAR prediction are to be assigned by experts, apart from predictive inference, also based on experience in experimental data. The use of extrapolation factors (uncertainty factors) can be used to widen the uncertainty reflecting the low reliability. The reliability in the overall risk assessment is verified by the analysis of the sensitivity the uncertain parameter may have on the resulting decision. If it has an influence on the decision further testing is needed.

The conclusion related to uncertainty in QSAR predictions for probabilistic risk assessment can be summarized as follows:

1) The integration of QSARs into probabilistic risk assessment is possible given proper assessments of **predictive uncertainty and predictive reliability**.

Predictive uncertainty and reliability are identified to inform the characterization of parameter and model uncertainty, two kinds of uncertainty to be identified in probabilistic risk assessment.

2) Probabilistic risk assessment is supported by QSAR predictions derived from Bayesian predictive inference. **Predicting must be done with care**, and the use of different bases for predictive inference is possible when a **QSAR is treated as a scientific based hypothesis supported by empirical data**.

In other words, the suggestion is to treat QSAR data that has been validated for its predictive performance as the QSAR to base predictions on. It is up to the assessor to choose an appropriate method for predictive inference, given that performance measures of predictivity do not deviate to

much from the measures in the peer-reviewed validation. This will result in a more flexible use of QSARs with a possibility of updating as long as new QSAR data becomes available.

3) The **extent of extrapolation in a QSAR prediction** influences predictive error and predictive reliability, and the domain of applicability is from an applied perspective context dependent and considered in the treatment of uncertainty.

A separation between predictive uncertainty and predictive reliability makes it possible to both apply models and discuss their reliability in a constructive way.

This report focuses on predictions of chemical specific properties and activities to replace testing information in chemical regulation. Statistical inference may have different purposes, and when QSARs are applied to support decisions making based on unobserved quantities such as in risk assessment, drug development, or experimental design, the statistical problem is to make predictions, referred to as predictive inference. Predictive science involves the use of a belief system about observables in science, and a philosophy of scientific methodology that implements that belief system. Predicting should be done with care (Appendix 1) and there are (solvable) practical problems when the purpose of statistical inference changes from inference on models to inference on predictions (Appendix 2).

The report discusses three kinds of **philosophies for predictive inference** of relevance for the application of QSARs in probabilistic risk assessment. **Sampling Theory** estimate predictive uncertainty based on a representative sample. Such (frequentist) inference rests upon assumptions of independent and, for example, identically distributed observations, in combination with a probabilistic assumption of uncertainty. Under violence of any of these assumptions, appliers of frequentist inference run into problems.

The Bayesian paradigm for inference *assign*, instead of *assume*, a probabilistic model for observations, and assign models for uncertainty in parameters (so called priors). **Bayesian inference** uses Bayes rule to update expert knowledge with information in empirical observations. The result is a well-defined probabilistic model of uncertainty. In cases of doubts, the caveat is the necessity to choose priors and probabilistic models (likelihoods). For example, there is no need to check an assumption of normality of errors (as in the frequentist case), as this is assigned through expert judgment.

The third alternative is to assign a probability distribution for predictive uncertainty based on **expert judgment** only. This can for example, be based on experience of experimental testing, or based on combinations of different sources of information.

Sampling Theory and solid expert judgment can be seen as extremes kinds of Bayesian inference; the first as Bayesian inference with non-informative priors (expressed simplistically, but it is more difficult than that); the second as Bayesian inferences with only priors. Therefore the recommendation we give is to use Bayesian inference as the statistical philosophy for predictive inference when QSARs are applied in probabilistic risk assessment.

QSAR models applied in this report were all **regressions** on a continuous non-bounded response variable. Predictive inference and QSAR modeling on categorical, discrete, or bounded response variables, or for

other reasons when a non-symmetric assumption of predictive uncertainty does not hold, is a challenge for future QSAR modeling. Another challenge is how to consider uncertainty in experimental QSAR data.

This report foresee several aspects of the reporting and documentation of QSARs that need to be changed with respect to the information needs when QSARs are integrated into probabilistic risk assessment. This will be further explored in the CADASTER deliverable “A guidance document on the use of QSARs in probabilistic risk assessment” (due in December 2012).

6. References

Abramovitch RA, Beckert JM, Gibson HH, et al. (2001) The 1,2,4-Triazolyl Cation: Thermolytic and Photolytic Studies. *The Journal of Organic Chemistry* 66: 1242-1251.

Ahlers J, Stock F and Werschkun B. (2008) Integrated testing and intelligent assessment-new challenges under REACH. *Environmental Science and Pollution Research* 15: 565-572.

Ankley GT, Jensen KM, Durhan EJ, et al. (2005) Effects of two fungicides with multiple modes of action on reproductive endocrine function in the fathead minnow (Pimephales promelas). *Toxicological Sciences* 86: 300-308.

Apostolakis G. (1990) THE CONCEPT OF PROBABILITY IN SAFETY ASSESSMENTS OF TECHNOLOGICAL SYSTEMS. *Science* 250: 1359-1364.

Aronson D, Boethling R, Howard P, et al. (2006) Estimating biodegradation half-lives for use in chemical screening. *Chemosphere* 63: 1953-1960.

Aven T. (2010a) On the Need for Restricting the Probabilistic Analysis in Risk Assessments to Variability. *Risk Analysis* 30: 354-360.

Aven T. (2010b) Some reflections on uncertainty analysis and management. *Reliability Engineering & System Safety* 95: 195-201.

Bhhatarai B and Gramatica P. (2011) Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. *Water Research* 45: 1463-1471.

Boethling RS, Howard PH, Meylan W, et al. (1994) Group-contribution method for predicting probability and rate of aerobic biodegradation. *Environmental Science & Technology* 28: 459-465.

Bosnic Z and Kononenko I. (2009) An overview of advances in reliability estimation of individual predictions in machine learning. *Intelligent Data Analysis* 13: 385-401.

Box GEP and Tiao GC. (1992) *Bayesian inference in statistical analysis,* New York: Wiley.

Breedveld GD, Roseth R, Hem L, et al. (2002) Triazoles in the terrestrial environment. Oslo, Norway: Norwegian Geotechnical Institute.

Bro-Rasmussen F. (1988) Hazard and risk assessment and the acceptability of chemicals in the environment. In: Richardson ML (ed) *Risk assessment of chemicals in the environment.* Cambridge, UK: Royal Society of Chemistry, 437-450.

Clark RD and Waldman M. (2012) Lions and tigers and bears, oh my! Three barriers to progress in computer-aided molecular design. *J Comput Aided Mol Des* 26: 29-34.

Cronin MTD, Walker JD, Jaworska JS, et al. (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.

Da Silva JP, Da Silva AM, Khmelinskii IV, et al. (2001) Photophysics and photochemistry of azole fungicides: triadimefon and triadimenol. *Journal of Photochemistry and Photobiology A: Chemistry* 142: 31-37.

De Roode D, Hoekzema C, de Vries-Buitenweg S, et al. (2006) QSARs in ecotoxicological risk assessment. *Regulatory Toxicology and Pharmacology* 45: 24-35.

Den Hollander HA, Van Eijkeren JCH and Van de Meent D. (2004) SimpleBox 3.0: Multimedia mass balance model for evaluating the fate of chemicals in the environment. Bilthoven, The Netherlands: National Institute for Public Health and the Environment (RIVM).

ECETOC. (1998) QSARs in the assessment of the envirionmental fate and effect of chemicals.

ECETOC. (2003) (Q)SARs: Evaluation of the commercially available software for human health and envirionmental endpoints with respect to chemical management applications.

ECHA. (2008a) Part E. Guidance on information requirements and chemical saftey assessment.

ECHA. (2008b) R.19 Guidance on information requirements and chemical saftey assessment.

Eizuka T, Ito A and Chida T. (2003) Degradation of ipconazole by microorganisms isolated from paddy soil. *Journal of Pesticide Science* 28: 200-207.

Eriksson J, Green N, Marsh G, et al. (2004) Photochemical decomposition of 15 polybrominated diphenyl ether congeners in methanol/water. *Environmental Science & Technology* 38: 3119-3125.

Eriksson L, Jaworska J, Worth AP, et al. (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives* 111: 1361-1375.

EU. (2006) Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH). . *Commission of the European Communities, Brussels,* .

European Chemicals Agency (ECHA). (2008) Guidance on information requirements and chemical safety assessment. Chapter R.6: QSARs and grouping of chemicals.

European Commission. (2003a) Review report for the active substance propiconazole. European Commission Health and Consumers Directorates-General.

European Commission. (2003b) Technical Guidance Document on Risk Assessment in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances, Commission Regulation (EC) No

1488/94 on Risk Assessment for existing substances, and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. Ispra (VA), Italy: European Chemicals Bureau, Joint Research Centre.

European Food Safety Authority (EFSA). (2010) Conclusion on the peer review of the pesticide risk assessment of the active substance cyproconazole. *EFSA Journal* 8.

Gouin T and Harner T. (2003) Modelling the environmental fate of the polybrominated diphenyl ethers. *Environment International* 29: 717-724.

Gramatica P. (2007) Principles of QSAR models validation: internal and external. *Qsar & Combinatorial Science* 26: 694-701.

Gramatica P. (2010) Chemometric methods and theoretical molecular descriptors in predictive QSAR modeling of the environmental behavior of organic pollutants. In: Puzyn T, Leszczynski J and Cronin MTD (eds) *Challenges and advances in computational chemistry and physics Volume 8: Recent advances in QSAR studies : methods and applications.* Dordrecht ; New York: Springer, xiv, 423 p.

Gramatica P, Giani E and Papa E. (2007) Statistical external validation and consensus modeling: A QSPR case study for K-oc prediction. *Journal of Molecular Graphics & Modelling* 25: 755-766.

Gramatica P, Pilutti P and Papa E. (2004) Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. *J Chem Inf Comput Sci* 44: 1794-1802.

Hassold E and Backhaus T. (2009) Chronic toxicity of five structurally diverse demethylase-inhibiting fungicides to the crustacean daphnia magna: a comparative assessment. *Environmental Toxicology and Chemistry* 28: 1218-1226.

Iqbal MS and Öberg T. (In review) Description and propagation of uncertainty in input parameters in environmental fate models.

Jaworska J, Gabbert S and Aldenberg T. (2010) Towards optimization of chemical testing under REACH: A Bayesian network approach to Integrated Testing Strategies. *Regulatory Toxicology and Pharmacology* 57: 157-167.

Johnson JB and Omland KS. (2004) Model selection in ecology and evolution. *Trends in Ecology & Evolution* 19: 101-108.

Kim IS, Beaudette LA, Shim JH, et al. (2002) Environmental fate of the triazole fungicide propiconazole in a rice-paddy-soil lysimeter. *Plant and Soil* 239: 321-331.

Klasmeier J, Matthies M, Macleod M, et al. (2006) Application of multimedia models for screening assessment of long-range transport potential and overall persistence. *Environmental Science & Technology* 40: 53-60.

Klix MB, Verreet J-A and Beyer M. (2007) Comparison of the declining triazole sensitivity of Gibberella zeae and increased sensitivity achieved by advances in triazole fungicide development. *Crop Protection* 26: 683-690.

Li ZH and Randak T. (2009) Residual pharmaceutically active compounds (PhAcs) in aquatic environment - status, toxicity and kinetics: a review. *Veterinarni Medicina* 54: 295-314.

Martens H and Næs T. (1989) *Multivariate calibration,* Chichester England ; New York: Wiley.

Maštovská K. (2005) Role of Pesticides in Produce Production, Preservation, Quality, and Safety. In: Ukuku D, Imam S and Lamikanra O (eds) *Produce Degradation.* CRC Press, 341-378.

National Research Council. (2009) *Science and decisions : advancing risk assessment,* Washington, D.C.: National Academies Press.

Netzeva TI, Worth AP, Aldenberg T, et al. (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships - The report and recommendations of ECVAM Workshop 52. *Atla-Alternatives to Laboratory Animals* 33: 155-173.

Nikolova N and Jaworska J. (2004) Approaches to measure chemical similarity - A review. *Qsar & Combinatorial Science* 22: 1006-1026.

Obrezanova O and Segall MD. (2010) Gaussian Processes for Classification: QSAR Modeling of ADMET and Target Activity. *Journal of Chemical Information and Modeling* 50: 1053-1061.

OECD. (2007) Report on the regulatory uses and applications in OECD member countries of quantitative structure-activity relationship (QSAR) models in the assessment of new and existing chemicals. In: Development OfEC-oa (ed).

Papa E, Kovarich S and Gramatica P. (2009) Development, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers. *Qsar & Combinatorial Science* 28: 790-796.

Puzyn T, Leszczynski J and Cronin MTD. (2010) *Recent advances in QSAR studies : methods and applications,* Dordrecht ; New York: Springer.

Raff JD and Hites RA. (2007) Deposition versus photochemical removal of PBDEs from lake superior air. *Environmental Science & Technology* 41: 6725-6731.

Roy PP, Kovarich S and Gramatica P. (2011) QSAR model reproducibility and applicability: a case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-)triazoles. *J Comput Chem* 32: 2386-2396.

Sahlin U. (2012) Reliability in Predictive Models under Alternative Treatments of Predictive Uncertainty – QSPRs in Chemical Safety Assessments. *ESREL 2012*.

Sahlin U, Filipsson M and Oberg T. (2011) A Risk Assessment Perspective of Current Practice in Characterizing Uncertainties in QSAR Regression Predictions. *Molecular Informatics* 30: 551-564.
Sahlin U, Jeliazkova N and Öberg T. (submitted) Applicability domain dependent Predictive Errors Sum of Squares to assess predictive uncertainty in QSAR regressions.
Schenker U, Soltermann F, Scheringer M, et al. (2008) Modeling the Environmental Fate of Polybrominated Diphenyl Ethers (PBDEs): The Importance of Photolysis for the Formation of Lighter PBDEs. *Environmental Science & Technology* 42: 9244-9249.
Schultz TW, Netzeva TI and Cronin MTD. (2004) Evaluation of QSARs for ecotoxicity: A method for assigning quality and confidence. *Sar and Qsar in Environmental Research* 15: 385-397.
Slob W. (1994) Uncertainty Analysis in Multiplicative Models. *Risk Analysis* 14: 571-576.
Sushko I. (2010) Applicability domain of QSAR models. Technical University of Munich.
Sushko I, Novotarskyi S, Korner R, et al. (2010a) Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *Journal of Chemical Information and Modeling* 50: 2094-2111.
Sushko I, Novotarskyi S, Korner R, et al. (2010b) Applicability domain for in silico models to achieve accuracy of experimental measurements. *Journal of Chemometrics* 24: 202-208.
Tetko IV, Bruneau P, Mewes HW, et al. (2006) Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* 11: 700-707.
Tetko IV, Sushko I, Pandey AK, et al. (2008) Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *Journal of Chemical Information and Modeling* 48: 1733-1746.
Tong WD, Xie W, Hong HX, et al. (2004) Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environmental Health Perspectives* 112: 1249-1254.
US EPA. (2002) *Persistent, Bioaccumulative, and Toxic Profiles Estimated for Organic Chemicals (PBT Profiler).* Available at: http://www.pbtprofiler.net/.
Walker JD. (2003) *Quantitative structure-activity relationships for pollution prevention, toxicity screening, risk assessment, and Web applications,* Pensacola, FL: SETAC Press.
Walker JD, Carlsen L and Jaworska J. (2003a) Improving opportunities for regulatory acceptance of QSARs: The importance of model domain, uncertainty, validity and predictability. *Qsar & Combinatorial Science* 22: 346-350.
Walker JD, Jaworska J, Comber MHI, et al. (2003b) Guidelines for developing and using quantitative structure-activity relationships. *Environmental Toxicology and Chemistry* 22: 1653-1665.
Wallace DF, Hand LH and Oliver RG. (2010) The role of indirect photolysis in limiting the persistence of crop protection products in surface waters. *Environmental Toxicology and Chemistry* 29: 575-581.
Van de Meent D. (1998) Environmental chemistry: instrument in ecological risk assessment. In: De Kruijff HAM, De Zwart D, Ray PK, et al. (eds) *Manual on Aquatic Ecotoxicology.* Dordrecht, The Netherlands: Kluwer Academic Press, p 31-35.
Wania F and Dugani CB. (2003) Assessing the long-range transport potential of polybrominated diphenyl ethers: A comparison of four multimedia models. *Environmental Toxicology and Chemistry* 22: 1252-1261.
Weaver S and Gleeson NP. (2008) The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics & Modelling* 26: 1315-1326.
Verdonck FAM, Van Sprang PA and Vanrolleghem PA. (2005) Uncertainty and precaution in European environmental risk assessment of chemicals. *Water Science and Technology* 52: 227-234.
Vialaton D, Pilichowski JF, Baglio D, et al. (2001) Phototransformation of propiconazole in aqueous media. *Journal of Agricultural and Food Chemistry* 49: 5377-5382.
Vialaton D and Richard C. (2002) Phototransformation of aromatic pollutants in solar light: Photolysis versus photosensitized reactions under natural water conditions. *Aquatic Sciences* 64: 207-215.
Worth AP. (2010) The role of QSAR methodology in the regulatory assessment of chemicals In: Puzyn T, Leszczynski J and Cronin MTD (eds) *Challenges and advances in computational chemistry and physics Volume 8: Recent advances in QSAR studies : methods and applications.* Dordrecht ; New York: Springer, xiv, 423 p.

Appendix 1. Statistical Concerns about QSAR Predictions

Main author: James E. Blevins

Abstract:
In implementing the REACH legislation, regulators have had difficulty using the *estimated end-points*. Such "estimated (future) end-points" are called "*predicted responses*" in statistics.

Uncertainty in the predicted responses has been quantified with probability distributions in the enclosed case-studies by other members of the Cadaster project. Such probability distributions may be expected to provide more insight into QSAR decision-problems. The mean (expected) predicted-response of the probability-based models should provide comparable performance as have the previous parameter-estimates.

We may wish that the probability distributions improve decision-making by allowing predictions of extreme performances: For example, we may wish that our QSAR predict the probability of having a response quite different than the average, predicted by the model, for a given chemical. However, such risk-management applications require expert knowledge that exceeds the capabilities of contemporary statistical models. The responsible use of QSAR models for risk-management requires the active participation of subject-matter experts.

Section A: Introduction

In implementing the REACH legislation, regulators have expressed frustration with their attempts to use *estimated end-points*. Such "estimated (future) end-points" are called "*predicted responses*" in statistics. Uncertainty in the predicted responses has been quantified with probability distributions in the enclosed case-studies by other members of the Cadaster project. We explain the statistical issues involved in providing such probability distributions.

Statistical issues in QSAR have been discussed in many publications, some elementary,[1] and others advanced;[2][3] other advanced discussions are referenced throughout the other parts of this document. Our reader should understand elementary statistics,[4] especially the basic ideas of regression (Freedman);[5] the reader would benefit from having read an elementary discussion of the philosophy of scientific statistics (Howson and Urbach)[6] or having studied calculus-based statistics (DeGroot and Schervish).[7]

---

[1] Statistics in Preclinical Pharmaceutical Research and Development, Bert Gunter and Dan Holderm, *Journal of the American Statistical Association,* Vol. 95, No. 451 (Sep., 2000), pp. 998-1001

[2] Beata Walczak, Micha Daszykowski, and Ivana Stanimirova, Robust methods in QSAR, *Recent Advances in QSAR Studies* (Tomasz Puzyn, Jerzy Leszczynski, and Mark T. Cronin, eds.), Challenges and Advances in Computational Chemistry and Physics, vol. 8, Springer Netherlands, 2010, pp. 177{208.

[3] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, and B. Walczak, Robust statistics in data analysis a review: Basic concepts, *Chemometrics and Intelligent Laboratory Systems* 85 (2007), no. 2, 203-219.

[4] David A. Freedman, Robert Pursani, and Roger Purvis. 2007. *Statistics*, Fourth Edition. Norton.
David S. Moore and George P. McCabe. 2006. *Introduction to the Practice of Statistics*, Fifth Edition. Freeman.

[5] David A. **Freedman**. 2009. *Statistical Models*, Second Edition. Cambridge UP.

[6] Colin **Howson** and Peter **Urbach** (2005). *Scientific Reasoning: the Bayesian Approach* (3rd ed.). Open Court Publishing Company. ISBN 978-0-8126-9578-6. (Topics are discussed with greater depth in the 2nd ed.)

[7] Morris H. **DeGroot** and Mark J. **Shervish**. *Probability and Statistics*, Third Ed. Addison-Wesley 9780201524888

Section B: Decision makers have requested probability distributions.

As an *example of a decision problem*, we describe the *design of experiments*. The issues raised by experimental design are present in other decision problems, and therefore their consideration can be informed by this discussion.

One goal of the REACH legislation is to minimize the cost of experiments, including accounting costs and *complementary costs* (e.g. imputed costs for exposure to harmful chemicals and for delayed access to safe chemicals). Of particular concern is to reduce inefficiencies with animal subjects, thereby reducing the accounting costs of experiments and improving the ethical treatment of animals.[8] Costs are reduced with the help of the statistical theory of experiments. An optimum design maximizes the information obtained from an experiment (subject to constraints for budgets and for technology).[9] Non-optimal experiments use more resources without any gains in statistical information.

Experiments are designed by scientists and statisticians who use their beliefs about potential outcomes as a function of the experimental conditions;[10] by definition, a "belief" is a truth-claim on

---

[8] Extrapolation from animal experiments to human populations is difficult. First, different species and different strains of species differ in their responses. For example, carcinogenicity results for rats and mice agreed 75% or less.
Tony Lin, Lois Swirsky Gold, David Freedman, Carcinogenicity Tests and Interspecies Concordance, *Statistical Science.* Volume 10, Number 4 (1995), 337-353.

Rats and mice diverged about 33 million years ago, about 63 million years after their common ancestor diverged from human ancestors.
Masatoshi Nei, Ping Xu, Galina Glazko. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms *PNAS 2001 98 (5) 2497-2502; published ahead of print February 20, 2001, doi:10.1073/pnas.051611498*

The heterogeneity of responses in animal experiments and the great difficulty of extrapolating results on high-dose short-term exposure of animals to humans are discussed in this paper (and the following discussion).
Freedman, David A. and Zeisel, H. From Mouse-to-Man: The Quantitative Assessment of Cancer Risks. *Statist. Sci.* Volume 3, Number 1 (1988), 3-28.

[9] Peirce, Charles Sanders (July-August 1967). "Note on the Theory of the Economy of Research". *Operations Research* 15 (4): pp. 643-648. doi:10.1287 [Reprint of an article from 1876]
Atkinson, A. C.; **Donev**, A. N.; **Tobias**, R. D. (2007). *Optimum experimental designs, with SAS*. Oxford University Press. pp. 511+xvi. ISBN 978-0-19-929660-6.

[10] Klaus Hinkelmann and Oscar Kempthorne. 2005. *Design and Analysis of Experiments, Volume 2: Advanced Experimental Design*, Wiley. ISBN 978-0-471-55177-5, page xxii.

which the thinker is prepared to act.[11] In the past, the design of experiments has informally relied on experts' beliefs; increasingly expert knowledge is being formalized in terms of probability distributions. For such experiments, a probability distribution on the parameter-values has been used to design experiments for decades.[12]

In experimental designs, then, decision-makers benefit from the provision of probability distributions. A method for producing such probability distributions is described next.

Section C: From estimated parameters to probability distributions

The probability distributions that have been provided to decision-makers usually have been rigidly specified. In most cases, one statistical-model which has been specified by constants ("estimated parameters"). In QSAR and in implementing REACH, practical users of chemometric models have trouble using existing parameter-estimates. Their frustration has motivated the Cadaster project.

Each predictive model provides the probability of a future event given the observed data, that is, **the conditional probability of the future-event given the data**. By manipulating the definition of conditional probability, Laplace and Bayes proved that this conditional probability can be **computed as the product of the data's *likelihood* (the conditional probability of the data given a value of the parameter) and a *"prior" probability***. (Howson and Urbach; DeGroot and Schervish). Statistical reasoning using such conditional probability became known as "Bayesian statistics" in the twentieth century. The Bayesian formalism expresses the probability model used for predictions---the posterior distribution---as the product of the likelihood and the prior, each of which can be studied, subjected to criticism, and improved. A failure of the predictive model necessarily can be located in a failure of the prior or likelihood or in a failure of both (Box, Tiao).[13]

Subsection C.1: Cautions about the use of parametric models for non-randomized data

We next discuss the problems with using parametric probability models for data, rather than the objective randomization procedures favored in experiments and in sample surveys.

---

[11] Peirce, Charles Sanders (1878 January), "How to Make Our Ideas Clear", *Popular Science Monthly*, v. 12, pp. 286–302.
Ramsey, Frank Plumpton (1931) "Truth and Probability", Chapter VII in *The Foundations of Mathematics and other Logical Essays*, Reprinted 2001, Routledge. ISBN 0415225469,

[12] Atkinson, A. C.; Donev, A. N.; Tobias, R. D. (2007). *Optimum experimental designs, with SAS*.
Fedorov, V. 1973. *Optimal design of experiments.* Translated from the Russian by W. J. Studden. Academic Press.

[13] Box, G.E.P. and Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis*, Wiley, ISBN 0-471-57428-7

Admittedly, the statistical models in QSAR are difficult to interpret. Statisticians rely on the probability models used in the analysis of randomized survey-samples and randomized experiments. In these applications, the probability model is known because it describes the *randomization* procedure specified in the study protocol (Freedman, Pursani, Purves; Freedman; Hinkelmann, Kempthorne). Because the probability distribution is designed by the statistician, there is no need to speculate about parametric models that are at best crude approximations (to the population from the data arose, often haphazardly). In randomized studies, the objective randomization allows a prescribed objective analysis.

In contrast, **QSAR models do not reflect a *known* probability distribution that is induced by *objective* randomization**.[14] New chemicals are not drawn randomly from a definite population; a model that assumes that the next chemical is another random selection from the same population that was randomly sampled for the previous chemicals has no plausibility. New chemicals are not *randomly* drawn but are *haphazardly* generated (by, for example, scientific progress and engineering skill). The elements of haphazard data-sets are plausibly described as "*dependent* and *differently distributed* (DDD)", rather than "*independent* and *identically distributed* (IID)" (Freedman).

In both the likelihood and the prior, **the probability models for QSAR represent *subjective, epistemic* probability-judgments, rather than objective facts**. As in other applications of subjective probability models in science, responsible application requires perspicacious modeling, which has been informed by earlier data and expert judgments.

The use of subjective parametric models for data is at best an approximation. Such models enable scientists economically to evaluate hypotheses and to suggest new hypotheses, which can be tested preferably by randomized studies (or at least on new data). As Charles S. Peirce wrote

> "Experience must be our chart in economical navigation; and experience shows that likelihoods are treacherous guides. Nothing has caused so much waste of time and means, in all sorts of researchers, as inquirers' becoming so wedded to certain likelihoods as to forget all the other factors of the economy of research; so that, unless it be very solidly grounded, likelihood is far better disregarded, or nearly so; and even when it seems solidly grounded, it should be proceeded upon with a cautious tread, with an eye to other considerations, and recollection of the disasters caused." (*Essential Peirce*, volume 2, pages 108–109)

A contemporary skeptical view of likelihood-models for data appears in the previously cited *Statistical Models* (by Freedman). In QSAR applications, the choice of a *t*-distribution or a normal distribution for the data's likelihood-function is a subjective choice by the modeler, which requires justification.

Besides exercising due care and caution with the likelihood function, statisticians must also tread cautiously with **prior probability**. Indeed, even the use of a prior probability-distribution has

---

[14] Plato defined knowledge as justified true belief; additional conditions are added in some definitions.

attracted controversy in statistics. At minimum, a neutral use of prior subjective-probability can consider priors representing all significant beliefs in the scientific community, individually with an emphasis on extreme views (or as a mixture of such individual beliefs).

The specification and examination of scientific prior probabilities is often prohibitively expensive, especially in using scientists' time. To reduce the costs of prior specification, practical statisticians use *default priors*, often called "*diffuse*", which spread the probability distribution over the entire parameter space, *enormously exaggerating uncertainty* to reduce the bias from the prior to a tolerable amount. Such conventional, diffuse priors appear for example in the three volumes of examples for WinBUGS/OpenBUGS.[15] The **use of such diffuse priors, *when approved by chemometricians and toxicologists* as exaggerating the uncertainty in the scientific community, is one apparently reasonable method** of providing the EU with the probability-distributions requested for decision-making; such diffuse priors should induce only negligible bias in resulting estimates (such as the posterior median or, if they exist, any posterior mean or mode).

However, diffuse priors that exaggerate uncertainty in the prior result in excessive uncertainty in the posterior. **The more diffuse the posterior, the more animal subjects** (e.g. humans) that **need to be allocated** in experiments to test a hypothesis (with prescribed power) or to estimate parameters (with prescribed confidence). In particular, the *t*-distributions with low degrees of freedom have widely spread tails whose enormity is difficult for non-statisticians to convey. Insofar as it is subjective, the selection of the degrees of freedom for the t-distribtion can dramatically change the posterior distribution, so this needs special attention from specialists and help from statisticians.

Thus, the prudent use of Bayesian QSAR models must consider **two conflicting goals, the need for objectivity and the desire to reduce the use of animals (especially humans) in experiments**.

Having cautioned the public of the limitations of subjective probability models, we now return to discussing regression modeling in practice.

[15] David Spiegelhalter, Andrew Thomas, Nicky Best, and Dave Lunn, *WinBUGS user manual*, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK, version 2.10 ed., April 2005.

Section D: Bayesian regression modeling

Despite the subjective choice of likelihood and prior in parametric-predictive statistics, there is surprising practical agreement about the modeling of distributions, at least for linear regression. Most books on Bayesian regression suggest multivariate (Gaussian) *normal* or (Student) *t* distributions for the likelihood of the data; most all suggest the same two distributions for the regression-coefficients' prior mean. Such prior-distributions respect the symmetries (with respect to orthogonal transformations) of the predictions, the data, and the prior.[16]

The multivariate *normal* distribution is a popular distribution for the data and for the regression coefficients. The multivariate *t*-distributions with fewer degrees of freedom (e.g., between 1 and 9 degrees of freedom) express greater prior-uncertainty about the true values of the regression coefficients; thus such *t*-distributions are often recommended as reducing the sensitivity of the posterior to changes in the median of the prior.

We have stated that the literature on Bayesian regression features widespread agreement on the distributions suitable for the mean of the prior-distribution for the regression coefficients. Disagreement and confusion arises in accounting for the *prior distribution on the* ***covariance matrix*** (which is conveniently induced by a prior distribution on the "***precision matrix***", which is the inverse of the covariance matrix).[17]

The literature on priors for the precision matrix provide counter-example to naive attempts automatically to provide "objective" priors for priors.[18] Such priors cause problems for especially for predictive modeling.[19] Such "non-informative" priors are derived from "principles" that have long been known known to lead to nonsense,[20] but which nonetheless survive as heuristics among the hurried.

---

[16] Bernardo, José M.; Smith, Adrian F. M. (1994). *Bayesian Theory*. Wiley.

[17] As an analogy to the normal distribution's "covariance matrix" of a normal distribution, the *t*-distribution has a "dispersion matrix" (regardless of the existence of finite variances or covariances).

[18] Heuristics for automatically choosing priors include (a) Jeffrey's rule, (b) maximum entropy, and other revivals of (c) Laplace's "principle" of insufficient reason, more properly known as the "base-rate fallacy").
Kass, R.E. and Wasserman, L.A. (1996) The selection of prior distributions by formal rules, *Journal of the American Statistical Association*, 91: 1343-1370.
Malay Ghosh, Objective Priors: An Introduction for Frequentists, *Statistical Science*. Volume 26, Number 2 (2011), 187-202.

[19] Ruo-yong **Yang** and James O. **Berger**, Estimation of a covariance matrix using the reference prior, *Annals of Statistics* 22 (1994), no. 3, 1195{1211. MR 1311972 (96b:62091)

[20] Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *Journal of the Royal Statististical Society Series B* **35** 189–233. .

We continue our discussion of a prior distribution for the covariance matrix (or for its inverse, the precision matrix); the plurality of priors illustrates difficulties of Bayesian statistics. Besides heuristics for "objective" priors, statisticians have used the Wishart distribution, which is computationally convenient (being a "conjugate" prior). Unfortunately, the Wishart prior does bias the posterior distribution against the equi-covariance structure that arises naturally in predictive inference (and sampling with replacement, etc.). Consequently, the Wishart prior inflates the mean squared error associated with point-estimates (the posterior median), when compared to the estimates from "*reference priors*", which are computed by simulations of random matrices.[21] Unfortunately, reference priors for the covariance matrix have not been implemented in OpenBUGS.

A posterior *t*-distribution results from a particular combination of "prior" infinite-measures (which are not probability measures) with prior probability-distributions, a normal likelihood, and computing a marginal posterior distribution.[22] However, the use of a (non-probability) infinite-measure as a "prior" typically leads to nonsense,[23] so the warranted use of this *t*-distribution predictive-model needs a statistically valid derivation.

Section E: Predictive *t*-distributions without Bayesian models

Ordinary least squares (OLS) produces mean-unbiased minimum-variance estimates of coefficients under the following assumptions:

- *Linearity*: The relationship between the response and the predictors is *linear*.
- *Random errors*: The *errors* each have mean zero, have the same variance (finite), and are *independent*.
  The results hold also if *independence* is generalized to *exchangeability* (Mancino, Pratelli):[24] A sequence of random variables is *exchangeable* if their joint distribution is invariant under permutations of indices. Sampling without replacement from a finite population generates exchangeable random-variables that are correlated.

---

[21] Yang and Berger.

[22] This derivation seems to have appeared in every standard book on Bayesian linear models.

[23] An exceptional allowance for infinite measures appeared in this monograph: Hartigan, J. A. 1983. *Bayes Theory*. New York: Springer-Verlag.

[24] M. E. **Mancino** and L. **Pratelli**, "Some Results of Stable Convergence for Exchangeable Random Variables in Hilbert Spaces", *Theory Probab. Appl.* 45, 329 (2001), DOI:10.1137/S0040585X97978270

Both assumptions deserve discussion particularly in QSAR modeling:.

- The *linearity* assumption may be false when there are nonlinear interactions among the variables, of course; however, a linear model provides an approximation, which usually provides better predictions than experts' judgments.[25]

- The *randomness* assumption for the error is more problematic, particularly in QSAR. An assumption of *exchangeability* would be warranted insofar as the chemicals were sampled (with equal probability) without replacement; since the randomness assumption cannot be seriously proposed for QSAR, the questions regarding the zero-mean and finite-variance assumptions are irrelevant.

We briefly sketch why, outside of QSAR, the OLS assumptions allow the use of predictive intervals.

If the randomness assumption holds, then the sample-mean of 30 observations of the response is well approximated by a normal distribution, according to simulation studies (Moore, McCabe), which yield stronger conclusions than even Berry-Essen refinements of the central limit theorem (Mancino, Pratelli; Hoffman-Jørgensen, p. 399).[26] The quality of the normal approximation enables the use of a predictive confidence-interval for the true value of the mean-response, without making further assumptions. As discussed in a first course in statistics, the *t*-distribution should be used for inference about the sample mean when the population-variance is unknown (Moore, McCabe; Freedman, Pursani, Purves).

[25] **Dawes**, R. M., **Faust**, D., and **Meehl**, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668-1674. Reprinted in T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 716-729). New York: Cambridge University Press, 2002.

[26] Jørgen Hoffman-Jørgensen's *Probability With a View Towards Statistics*, Volume I.

Appendix 2. Comments on the predictive distribution of a linear regression

Main author: Ullrika Sahlin

Integrating QSARs in risk assessment changes focus of statistical inference, from inference on model to inference on predictions. These comments explain the predictive inference of the regression models fitted by OLS in the two case-studies.

In the two case studies in this report used QSARs that were linear regressions developed in WP3. For each chemical and QSAR model, the authors reported a predictive distribution for response $\hat{Y}$. In particular, the authors report the predictive mean PRED($\hat{Y}$). This prediction has minimum-variance among all mean-unbiased estimators under the usual assumptions for the (*ordinary) least squares* (OLS) analysis of observational data (see Appendix 2).

Besides the point-estimate of the mean response, the authors also reported prediction intervals, which were based on the following summary statistics:

- the predictive mean PRED($\hat{Y}$),
- the predictive error SEP($\hat{Y}$),
- the number of data points in the training data set (n), and
- the number of descriptors in the linear regression model (p).

The prediction $\hat{Y}$ was distributed according to its predicted distribution

$$\hat{Y} \sim \text{PRED}(\hat{Y}) + t_{n-p-1}\, \text{SEP}(\hat{Y}), \qquad (1)$$

where t_{n-p-1} stands for the t-distribution with $n - p - 1$ degrees of freedom. The t-distribution is the result of inference on regression coefficients and of the variance, and can be derived analytically given prescribed conditions. The predictive error is estimated as

$$\text{SEP}(\hat{Y})^2 = \sigma^2(1+W^t(X^tX)^{-1}W), \qquad (2)$$

where σ^2 is the variance in model errors and $(X^tX)^{-1}$ is the information matrix (Box and Tiao 1992). Thus, the descriptors in the training data set X, and of the compound to be predicted W are needed to specify the predictive error. An intercept term is present in the model matrix X and the query compound descriptors W. Note that predictive error is assessed by multiplying model variance with leverage (diagonal form the "hat" projection-matrix) and a one (i.e. 1). This one (i.e. 1) adds uncertainty to the error in prediction. Given a correct model, the prediction interval of the predictive distribution covers the true value of the population mean (of the future responses) with the probability prescribed by the confidence level. The adding of the error

uncertainty is important when predicting a new and not yet observed activity or property of a chemical.

If the purpose is rather to test a hypothesis about the slope of the regression line, then the type-I error probability ("alpha") is derived from the observational error (rather than from the predictive error). The predictive distribution describes the uncertainty in the expected (future) response and also the model error; model error depends on model structure and the truth, and (in sampling-theory statistics) is estimated using near-replicates in the data set. A disadvantage of frequentist practice is its rigid use of one estimated model, which is used to estimate model error; in contrast, (Bayesian) probability-based modeling allows the consideration of a continuum of models, each of which's plausibility is quantified using the posterior probability density.

The *Bayesian lasso* provides an example of predictive inference. In our model, the endpoint Y is supposed to be normally distributed; the prior distribution on the regression coefficients is a product of univariate normal distributions; the infinite (non-probability) measure puts a constant value of $1/\sigma^2$ for the model variance. This Bayesian model was called "the Bayesian lasso" since it modifies "the lasso", a method for sparse or penalized regression. The Bayesian lasso has good frequentist performance, in limited testing (Park and Casella 2008; Hastie, Tibshirani et al. 2009).

Analytical solutions to Bayesian inference quickly becomes computational complex. It is therefore common to make Bayesian inference by Markov Chain Monte Carlo (MCMC) sampling from the posterior distribution, i.e. joint distribution of regression coefficients and variance (see DeGroot and Schervish 2002 for an introduction).

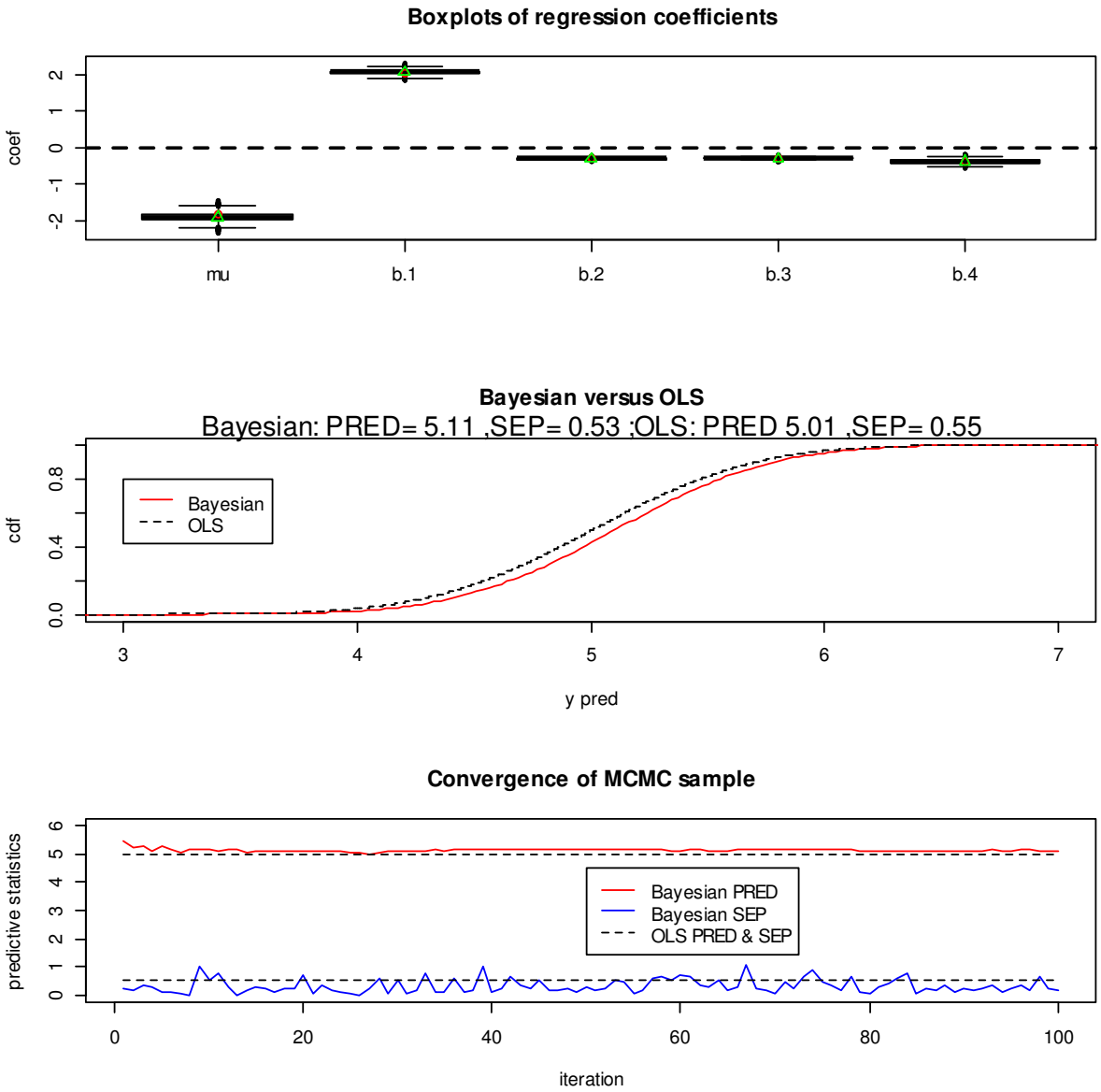


Figure A2.1. Uncertainty in regression coefficients from both the Bayesian lasso and ordinary least-squares (green triangle) (upper). The corresponding predictive distributions (middle). The sample-path for the predictive mean (PRED) and predictive error (SEP) of the Bayesian lasso (lower).

We implemented the Bayesian lasso for QSAR using the R-package monomvn of Gramacy and Pantaleo (2010). First, we computed the posterior distribution of the linear regression coefficients and model variance predictions using Markov Chain Monte Carlo[1]. The posterior distribution is represented by a sample for the stationary distribution of an appropriate Markov chain. Second, these sampled posterior parameters each defined a predictive distribution. For each sampled parameter-value, we computed new data (using the poster-values, for the mean and variance). Third, we repeated this procedure for a new MCMC sample: the union of such samples from predictive distributions constituted the (computed) posterior predictive distribution. The method was implemented on the QSAR data for the multiple linear regression for K_{OH} (Gramatica, Pilutti et al. 2004). The results from the Bayesian lasso were similar to the results based on sampling-theory with a normal distribution, which coincide with ordinary least-squares estimates (Figure A1.1).

A review of the simulations showed agreement between the predictive distributions from the Bayesian lasso and the t-distribution (Figure A2.1).

There are minor differences, of course. In the Bayesian lasso, the normal-prior on the regression coefficients were associated with a distinction between the posterior mean and the mode, which coincide in the sampling-distribution approach. Using the assumption that the data are drawn from a normal distribution, the sampling-theory estimates (OLS) the mode of the likelihood function. This effect can be seen in Figure A2.2 where exactly the same analysis was run as in Figure A2.1, but with the normal priors on the regression coefficients replaced by non-informative priors. As the variance of the prior normal-distribution on the regression coefficients increases without bound, the mode of the posterior approaches the posterior mean. For very large variances for the prior, the prior appears to be “flat” to the human eye, which led some to consider “uniform” (non-probability) priors, such as the Bayesian lasso’s infinite-measure on the variance; however, such non-probability priors lead to nonsense on other models, we warn.

[1] Posterior distributions are difficult to compute in closed form. When the parameter space exceeds 10, numerical quadrature quickly becomes impractical; then simulation becomes the usual method of evaluating integrals, such as the evaluating the posterior distribution. One of the most popular simulation methods is Markov Chain Monte Carlo (MCMC), which is described in calculus-based statistics books (e.g. DeGroot and Schervish 2002).

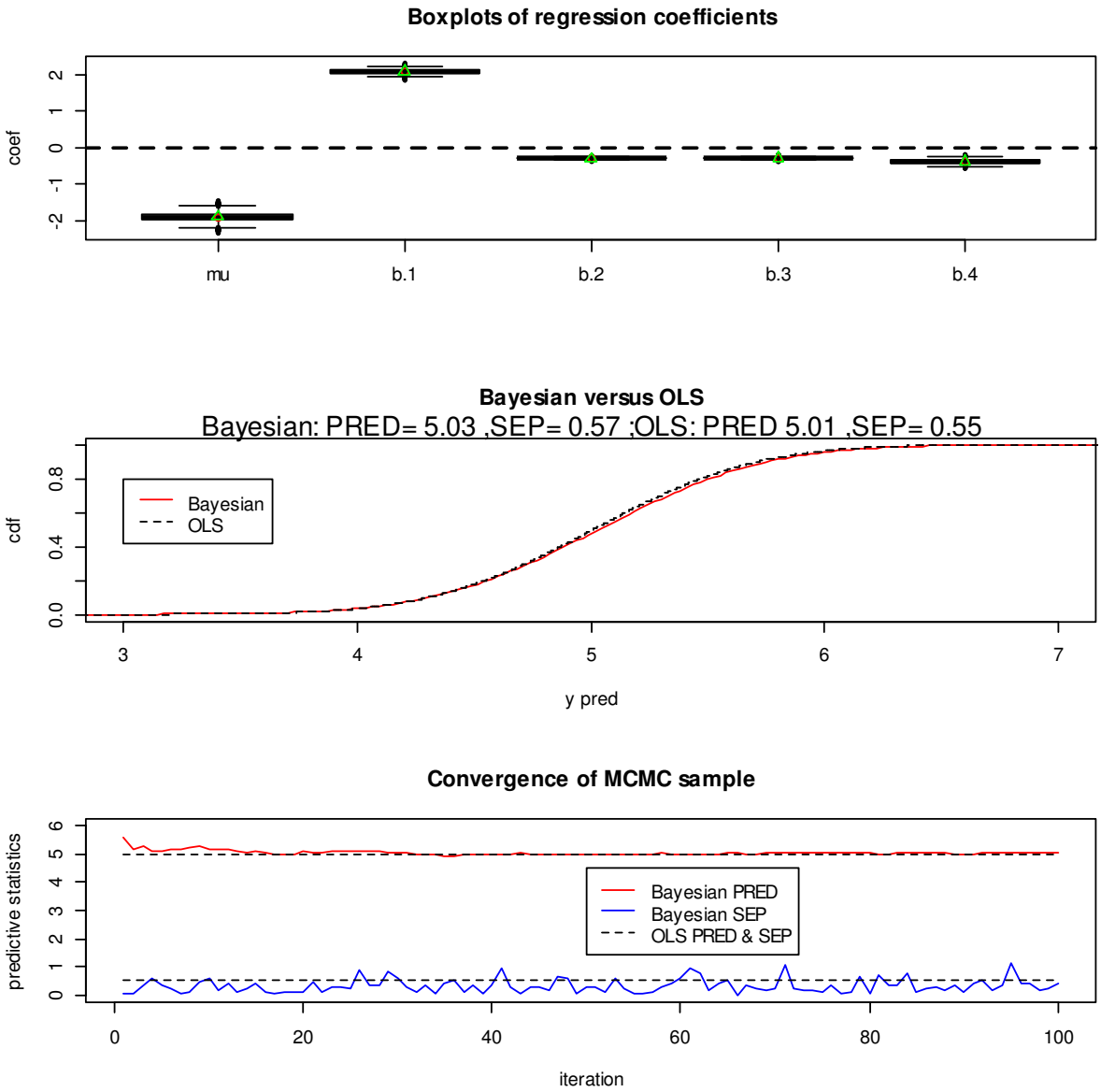


Figure A2.2. Uncertainty in regression coefficients from both the Bayesian lasso with non-informative priors and ordinary least-squares (green triangle) (upper). The corresponding predictive distributions (middle). The sample-path for the predictive mean (PRED) and predictive error (SEP) of the Bayesian lasso (lower).

This example demonstrates that the use of the t-distribution defined in Equation 1 based on the OLS estimates is an adequate assessment of the predictive distribution based on Bayesian predictive inference when the Bayesian linear regression model are given non-informative priors[2]. However, informed Bayesian inference may be an alternative. Note that these results apply to the linear regression model with symmetrically distributed errors, as soon as other assumptions on variance of errors or the symmetry of distributions does not hold more complicated models are needed for predictive inference.

Our brief simulation study shows that, under the assumption (hopefully checked) of independent random errors that follow a normal distribution, similar results follow from either the Bayesian lasso or the sampling-distribution based t-distribution (Equation 1). The first appendix raises concerns that about this assumption of independent and identically distributed errors, especially for QSAR.

References

Box, G. E. P. and G. C. Tiao (1992). Bayesian inference in statistical analysis. New York, Wiley.

DeGroot, M. H. and M. J. Schervish (2002). Probability and statistics. Boston, Addison-Wesley.

Gramacy, R. B. and E. Pantaleo (2010). "Shrinkage Regression for Multivariate Inference with Missing Data, and an Application to Portfolio Balancing." Bayesian Analysis **5**(2): 237-262.

Gramatica, P., P. Pilutti, et al. (2004). "Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling." J Chem Inf Comput Sci **44**(5): 1794-1802.

Hastie, T., R. Tibshirani, et al. (2009). The elements of statistical learning : data mining, inference, and prediction. New York, NY, Springer.

Park, T. and G. Casella (2008). "The Bayesian Lasso." Journal of the American Statistical Association **103**(482): 681-686.

[2] The author would like to emphasis that this example is not a state of the art, but a way to motivate current practice from the basis of predictive inference.

QSAR models in a probabilistic risk assessment framework CADASTER deliverable 4.1 Application of QSAR models for probabilistic risk assessment

Appendix 3. Risk Characterisation Ratio (RCR) and Expected Risk (ER)

Main author: Tom Aldenberg, RIVM , April 20, 2012

When a chemical is manufactured or imported in quantities of more than 10 tonnes per year, it is required to conduct a chemical safety assessment (CSA) and to prepare a chemical safety report (CSR). This chemical safety assessment generally is achieved along two separate lines of evidence: Hazard Assessment (HA) and, if a substance is classified as dangerous, or assessed to have PBT or vPvB properties, in addition Exposure Assessment (EA). The two assessments are integrated in the Risk Characterisation (RC) stage of the chemical assessment, as shown in Fig. A.[1].

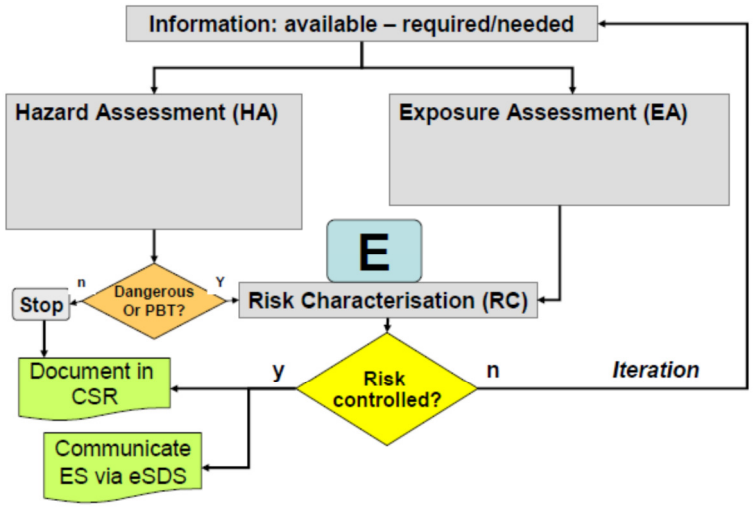


Fig. A. Information flowchart of a Chemical Safety Assessment. Lines of evidence through Hazard Assesment (HA) and Exposure Assessment (EA) meet at the Risk Characterisation (RC) stage that may either end the assessment, or lead to further iterations of the assessment ([1]).

In the Risk Characterisation phase, a central role is played by the Risk Characterisation Ratio (RCR). REACH Chemical Safety Assessment hinges on the RCR. In this Section, we will present a modeling view on the RCR and show its relationship to another measure of risk called: Expected Risk (ER).

Basically, the RCR is a ratio of an exposure value divided by a no-effect value ([1]):

$$RCR = \frac{PEC}{PNEC}, \text{ or } \frac{Exposure}{DNEL}, \quad (1)$$

where the first expression aims at environmental RC, the second addresses human health RC.[1] Although our considerations apply to both fields to a large extent, we are focusing on the environmental RC.

Basically, the RCR is a ratio of two fixed numbers, so-called point estimates, assuming that the risk of the chemical is 'controlled', if the RCR is below 1, and not controlled, if it is above 1 ([1]). If the RCR is close to 1, or exceeds 1, it is necessary to refine the chemical safety assessment through a stepwise approach ([2]):

1) At Level 1 (Qualitative Uncertainty Analysis), all uncertainties are treated qualitatively, by way of listing the different sources of uncertainty and variability.
2) At Level 2 (Deterministic Uncertainty Analysis), alternative point estimates are generated, through a series of reasonable worst-case exposure assumptions, and by varying factors for the determination of the hazard.
3) At Level 3 (Probabilistic Uncertainty Assessment), the aim is to determine the probability that the RCR is exceeded, allowing for the fact that both effect and exposure are probabilistic quantities.

This paper addresses Level 3 (Probabilistic Uncertainty Assessment) of risk characterisation in chemical safety assessment ([2]). The advantage of probabilistic risk assessment is that –consistent with the probabilistic nature of risk– more accurate chemical risk estimates can be obtained, compared to assessments based on worst-case assumptions, which brings in an unknown degree of conservatism.

Disadvantages of Probabilistic Risk Assessment (PRA) are: (1) Increased data requirements; (2) Increased calculation efforts; (3) Lack of experience among risk assessors and lack of guidance; (4) Needs to adapt risk communication procedures. The fear is expressed that PRA may be difficult, time-consuming, and expensive to carry out ([3], p. 23). This may certainly be the case for complex models and complicated assessments with high stakes.

However, we are of the opinion that there is reason for optimism.

First of all, the REACH Uncertainty analysis chapter R.19 ([2]) is a great step forward, as it clearly defines, what levels of uncertainty analysis to distinguish, and what the qualifiers 'deterministic' and 'probabilistic' mean in the context of chemical safety assessment.

Second, conceptually more sophisticated insights have been proposed in recent years ([4]; [5]; [6]; [7]; [8]). This Section extends and systematises probabilistic chemical safety assessment on the basis of exposure and effect distributions, as reviewed and analysed previously ([4]; [5]).

[1] PEC: predicted environmental concentration; PNEC: predicted no-effect concentration; DNEL: derived no-effect level.

We will explain the relationship of Expected (ecological) Risk with probabilistic RCR in the REACH Uncertainty Analysis Guidance ([2], Figs. R.19-5 and R.19-6). Readily usable computer program code, developed in the free statistical software environment 'R' ([9]) are available from Tom Aldenberg on request.

For illustrations, we use a model in which both the exposure distribution, as well as the no-effect distribution, is a Normal distribution over $\log_{10}$ concentration, denoted as x. This is our *canonical* model. However, the theory is developed in full generality. The environmental no-effect distribution may be a Species Sensitivity Distribution (SSD) of chronic data, e.g. NOECs, or acute data (that can be extrapolated with a safety factor). With the environmental risk assessment in mind, the basic idea is that cumulative values of the SSD, if suitably assessed, can be interpreted as a dose-response curve, the response being the (potential) Fraction (of species, or taxa) Affected (FA). Issues pertaining to SSDs are treated in recent monographs on the subject ([10]; [11]).

To gain an impression of the risk due to overlapping distributions, one may plot overlays of the density functions (PDFs), or of the cumulative distribution functions (CDFs), the horizontal axis preferably on a logarithmic scale, here $\log_{10}$ (Fig. B).

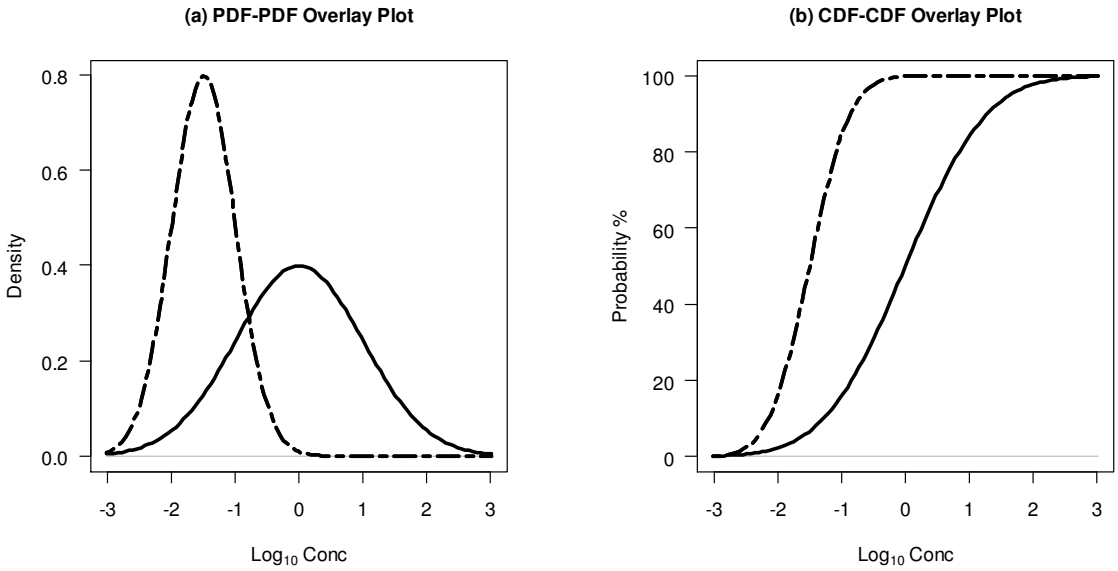


Fig. B. (a) Overlay plot of exposure density function (dashed), and no-effect density function (continuous); (b) Overlay plot of exposure cumulative distribution in % (dashed), and no-effect cumulative distribution in % (continuous).

It definitely makes analysis easier to think of these distributions as densities over a logarithmic scale, then to consider explicit lognormal distributions over untransformed concentrations.

Suppose, one would quantify the overlap through Monte Carlo analysis by simulating random values from either distribution (assuming independence), and counting how often exposure concentrations exceed no-

effect concentrations. The probability that random variable X exceeds random variable N is called the Probability of Failure (POF) in reliability engineering (e.g. [12]):

$$\mathrm{POF} = \Pr[X > N]. \tag{2}$$

The analytical expression for *any pair* of exposure and no-effect distributions (not necessarily Normal) comes in two versions:

$$\Pr[X > N] = \int_{-\infty}^{\infty} \mathrm{CDF}_N(x) \cdot \mathrm{PDF}_X(x)dx, \tag{3}$$

$$\Pr[X > N] = \int_{-\infty}^{\infty} \left(1 - \mathrm{CDF}_X(x)\right) \cdot \mathrm{PDF}_N(x)dx. \tag{4}$$

([13]; [12]; [14]).

For many probability distributions, it will be more efficient to evaluate the probability of failure integral by direct numerical integration than by Monte Carlo simulation. But, the latter interpretation better explains its meaning.

If both exposure and no-effect variables have Normal distributions:

$$\begin{aligned} X &\sim \mathrm{Normal}(\mu_X, \sigma_X) \\ N &\sim \mathrm{Normal}(\mu_N, \sigma_N), \end{aligned} \tag{5}$$

Eq. (3) further simplifies to:

$$\Pr[X > N] = \Phi\left(\frac{\mu_X - \mu_N}{\sqrt{\sigma_X^2 + \sigma_N^2}}\right) \tag{6}$$

([12]; [15]; [4]), where Φ is the standard Normal cumulative distribution function

Van Straalen realized that the POF integrals –which he called *Ecological* Risk– have a graphical interpretation ([14]; [4]; [16]). We have two graphical representations, corresponding to the two POF expressions in Eqs. (3) and (4), as shown in Fig. C. Fig. C(a) displays an overlay of exposure density and cumulative no-effect distribution, according to Eq. (3). Fig.(b) exhibits an overlay of the complementary cumulative exposure distribution with the no-effect density function.

Fig. C(a) is similar to Fig. R.19-5 in the ECHA Uncertainty Analysis Guidance ([2], p. 28).

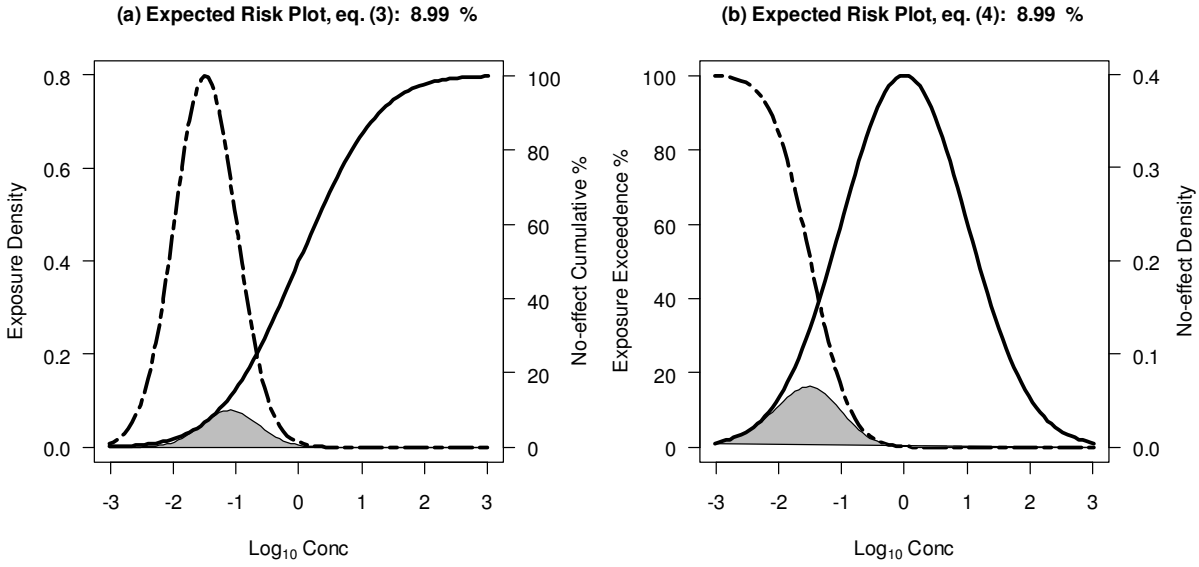


Fig. C. (a) Expected risk plot according to Eq. (3): exposure density (dashed) and no-effect cumulative distribution (continuous); (b) Expected risk plot according to Eq. (4): exposure complementary cumulative (or exceedence) distribution (dashed) and no-effect probability density (continuous). Expected risk is the area under the curve of the product of the respective probability functions, here equal to 8.99%. Plot (a) corresponds to Fig. R.19-5 in the ECHA Uncertainty Analysis Guidance ([2], p. 28).

In both cases, two different vertical axes are needed, as density values (unit: $1/\log_{10}$ concentration) differ from cumulative probabilities (unit: fraction, expressed in percent). A *complementary* cumulative probability is the probability that a random ($\log_{10}$) concentration is exceeded. It is equal to one (100%) minus the cumulative distribution value at each point, and therefore a descending function. We use the word *exceedence* after the *exceedence profile plot* ([17]),

By way of a scaling argument, we demonstrated that ER depends on only two parameters ([4], p.72). We chose to standardise the no-effect distribution, to become standard Normal, as the exposure distribution may be rather variable, depending on circumstances, or scenarios. This leads to the starred distributions, X^* and N^*, parameterised as follows:

$$\begin{cases} \mu_{X^*} = \dfrac{\mu_X - \mu_N}{\sigma_N}, & \sigma_{X^*} = \dfrac{\sigma_X}{\sigma_N}, \\ \mu_{N^*} = \dfrac{\mu_N - \mu_N}{\sigma_N} = 0, & \sigma_{N^*} = \dfrac{\sigma_N}{\sigma_N} = 1. \end{cases} \tag{7}$$

The expected risk, Eq. (6), then simplifies to:

$$\mathrm{ER} = \Phi\left(\frac{\mu_{X^*}}{\sqrt{\sigma_{X^*}^2 + 1}} \right) \tag{8}$$

Table 1 tabulates ER as a function of exposure mean μ_{X*} and standard deviation σ_{X*}, relative to standardized no-effect: $\mu_{N*}=0, \sigma_{N*}=1$ (adapted from [4], p. 73).

Table 1
Expected Risk (ER) in percent for Normal $\log_{10}$ exposure and no-effect distributions, as function of $\log_{10}$ exposure mean μ_{X*} and standard deviation σ_{X*}, relative to standardized no-effect: $\mu_{N*}=0, \sigma_{N*}=1$

$\sigma_{X*} \setminus \mu_{X*}$	**–5.0**	**–4.5**	**–4.0**	**–3.5**	**–3.0**	**–2.5**	**–2.0**	**–1.5**	**–1.0**	**–0.5**	**0.0**
0.0	0.00	0.00	0.00	0.02	0.13	0.62	2.28	6.68	15.87	30.85	50.00
0.2	0.00	0.00	0.00	0.03	0.16	0.71	2.49	7.07	16.34	31.20	50.00
0.5	0.00	0.00	0.02	0.09	0.36	1.27	3.68	8.99	18.55	32.74	50.00
1.0	0.02	0.07	0.23	0.67	1.69	3.85	7.86	14.44	23.98	36.18	50.00
1.5	0.28	0.63	1.33	2.61	4.80	8.28	13.36	20.27	28.95	39.08	50.00
2.0	1.27	2.21	3.68	5.88	8.99	13.18	18.55	25.12	32.74	41.15	50.00

Up to now, we have considered Expected Risk on the $\log_{10}$ concentration scale. It can be demonstrated that, on the original concentration scale, the calculation of Expected Risk will have exactly the same value. So, the focus on Normal distributions over the $\log_{10}$ concentration axis immediately translates to their lognormal counterparts. The argument can be extended to any monotonic scale transformation.

Relating Deterministic Risk Characterisation Ratio and Expected Risk

We will now address the problem, why deterministic RCRs, as defined in the REACH guidance ([1]; [2]) are very hard to interpret, quantitatively, despite their simplicity to calculate. We present a numerical example, based on the canonical PRA model of two Normal distributions for both $\log_{10}$ exposure and $\log_{10}$ no-effect.

Table 2 presents two hypothetical risk characterisations A and B, with given Normal exposure and no-effect distributions. The parameters are given in raw (un-standardized) $\log_{10}$ units, and standardized $\log_{10}$ units, scaled to the no-effect distribution.

Table 2
Two hypothetical risk assessments, case A and B, with Normal exposure and no-effect distributions. Raw (un-standardized) parameters are on the $\log_{10}$ scale. Standardized parameters are scaled to the parameters of the no-effect Normal distribution according to Eq. (7). Expected risk equals 8.99% in all cases; expected risk plots in Fig. . Cases A and B are essentially the same.

		Case A		Case B	
	Parameters	Raw	Standard.	Raw	Standard.
Exposure	μ_X	2.00	–1.50	0.00	–1.50
	σ_X	0.25	0.50	1.00	0.50
No-effect	μ_N	2.75	0.00	3.00	0.00
	σ_N	0.50	1.00	2.00	1.00
	ER%	8.99	8.99	8.99	8.99

Fig. D displays the expected risk plots for cases A and B, both raw (un-standardised), (a) and (b), as well as standardized to the no-effect distribution, (c) and (d). The expected risk is 8.99% in all cases.

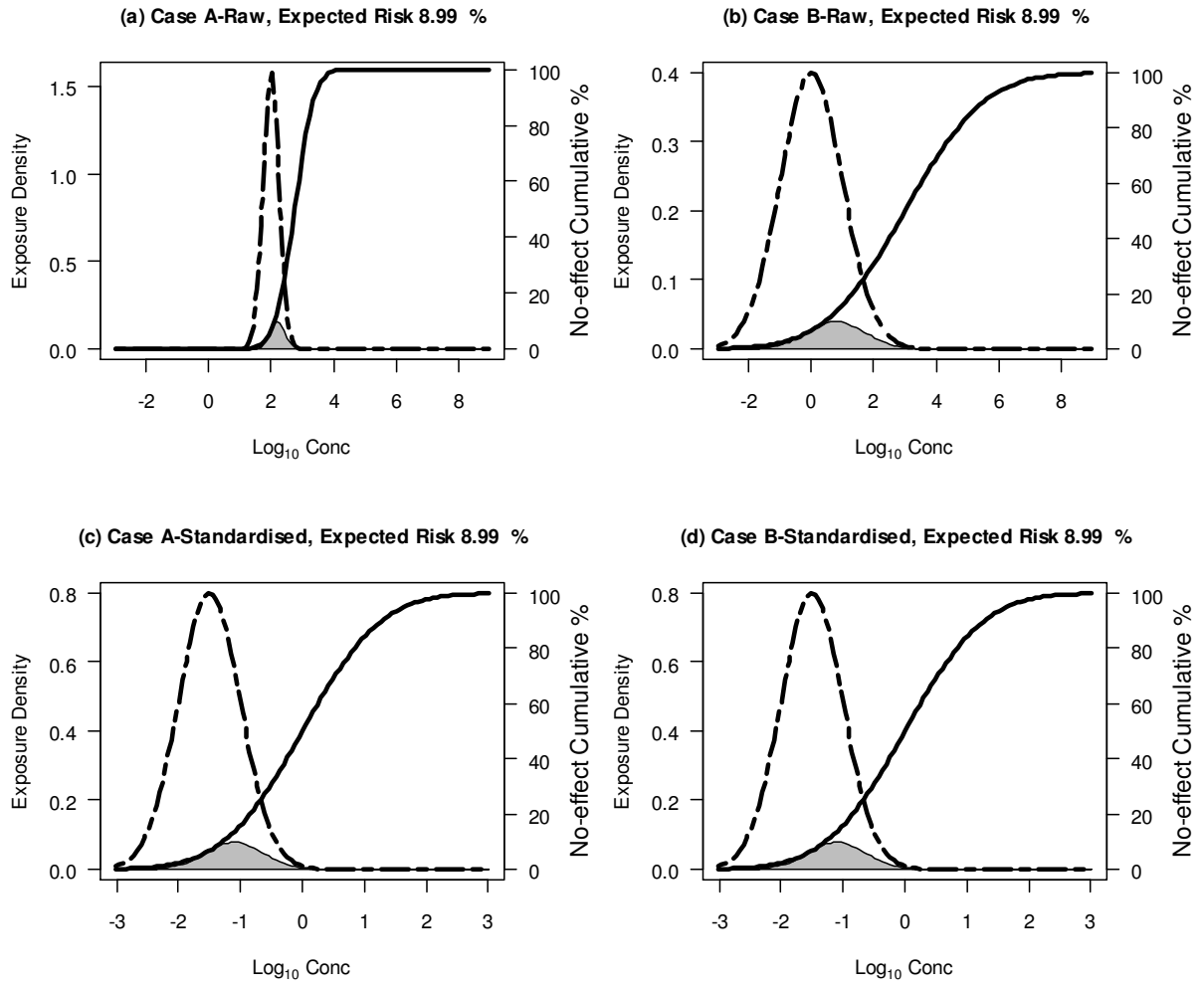


Fig. D. Expected risk plots for two hypothetical risk assessments, case A and B, with Normally distributed exposure and no-effect distributions (parameters in Table 2); (a) and (b) plotted over raw (un-standardised) $\log_{10}$ concentration; (c) and (d) standardised to the no-effect distribution, which reveals that the cases are identical. The expected risk is 8.99% in both case A and case B (cf. Table 1).

Within this hypothetical (environmental) setting, we compare the expected risk to a deterministic assessment by imposing quantitative worst case assumptions by means of the 95th percentile of the exposure distribution ([2], p. 15)[2] and the 5th percentile of the no-effect distribution ([18], p. 23).

The $RCR_{95/5}$ is calculated as:

[2] In previous TGD documents, the 90th percentile was often advocated for exposure. However, the exact percentile choice is not critical in the argument.

$$\mathrm{RCR}_{95/5} = \frac{\mathrm{PEC}_{95}}{\mathrm{PNEC}_5} = \frac{10^{\mu_X + 1.645 \cdot \sigma_X}}{10^{\mu_N - 1.645 \cdot \sigma_N}} \tag{9}$$

It follows that

$$\log_{10}(\mathrm{RCR}_{95/5}) = \mu_X - \mu_N + 1.645 \cdot (\sigma_X + \sigma_N) \tag{10}$$

The results for both cases, A and B, are given in Table 3.

Table 3
Deterministic risk characterisation ratio, $\mathrm{RCR}_{95/5}$, for cases A and B from Table 2, as plotted in Fig. . PEC_{95}, PNEC_5, $\mathrm{RCR}_{95/5}$, and $\log_{10}(\mathrm{RCR}_{95/5})$ calculated from Eq. (9) and (10). The deterministic $\mathrm{RCR}_{95/5}$ in case B is 28 times the one in case B, but the standardized $\mathrm{RCR}_{95/5}$-values are equal. The expected risk (ER) is 8.99% in all cases.

		Case A		Case B	
	Estimates	Raw	Standard.	Raw	Standard.
Exposure	PEC_{95}	257.78	0.21	44.16	0.21
No-effect	PNEC_5	84.63	0.02	0.51	0.02
	$\mathrm{RCR}_{95/5}$	**3.05**	**9.28**	**86.10**	**9.28**
	$\log_{10}(\mathrm{RCR}_{95/5})$	0.48	0.97	1.94	0.97
	ER%	8.99	8.99	8.99	8.99

The deterministic $\mathrm{RCR}_{95/5}$, calculated from Eq. (9), turns out to be 3.05 in case A, versus 86.10 in case B, which is 28 times as high. After standardisation on the $\log_{10}$ scale, the $\mathrm{RCR}_{95/5}$ becomes 9.28 in both cases, A and B.

Why the deterministic $\mathrm{RCR}_{95/5}$ yields different answers in cases A and B, and both agree after standardisation at yet another value, can be examined by substituting the standardisation equations (7) into Eq. (9):

$$\begin{aligned} \mathrm{RCR}_{95/5} &= 10^{\mu_X + 1.645 \cdot \sigma_X - \mu_N + 1.645 \cdot \sigma_N} \\ &= 10^{\left(\mu_{X*} + 1.645 \cdot (\sigma_{X*} + 1)\right) \cdot \sigma_N} \\ &= \left(10^{\mu_{X*} + 1.645 \cdot (\sigma_{X*} + 1)}\right)^{\sigma_N} \end{aligned} \tag{11}$$

Apparently, the deterministic RCR does not only depend on the standardized Normal exposure parameters, μ_{X*} and σ_{X*}, which determine the expected risk, but also on the un-standardised $\log_{10}$ no-effect distribution standard deviation: σ_N .

It follows from Eq. (9) that the standardised deterministic $\mathrm{RCR}_{95/5}$ is equal to

$$\mathrm{RCR}^*_{95/5} = 10^{\mu_{X*} + 1.645 \cdot (\sigma_{X*} + 1)}, \tag{12}$$

so that we can relate standardised and un-standardised deterministic $\mathrm{RCR}_{95/5}$-values, as follows:

$$\mathrm{RCR}_{95/5} == \left(\mathrm{RCR}^{*}_{95/5}\right)^{\sigma_N}. \tag{13}$$

Since σ_N in case B is 2.0/0.5 = 4 times the one in case A (Table 2), it follows from Eq. (13) that $RCR_{95/5}$ in case B becomes the 4th power of $RCR_{95/5}$ in case A: $3.05^4 = 86.1$. Note that the $\log_{10}(RCR_{95/5})$ in case B (1.94) is 4 times that in case A (0.48). The standardised $\log_{10}(RCR^{*}_{95/5}) = 0.97$ is twice that of case A, and one half the one in case B.

An important observation is that a change of concentration unit, e.g. dividing untransformed concentrations by 1000, will only affect PEC and PNEC, but not RCR, nor ER. This is because, in this case, the $\log_{10}$ means, μ_X and μ_N, both shift three units to the left on the $\log_{10}$ axis, the difference staying the same, while the $\log_{10}$ standard deviations, σ_X and σ_N are unchanged.

It follows from Eq. (13) that, if the standardised $RCR_{95/5}$ is above trigger value 1, as in Table 3, then all un-standardised $RCR_{95/5}$ are also above 1, for any σ_N, although the values differ. If the standardised $RCR_{95/5}$ is below 1, all of them are. Note also that for very small $\sigma_N \to 0$, $RCR_{95/5}$ values approach 1, from either side.

Thus, $RCR_{95/5}$ grows (diminishes) with increasing σ_N, if

$$\log_{10}\left(\mathrm{RCR}^{*}_{95/5}\right) = \mu_{X*} + 1.645 \cdot (\sigma_{X*} + 1) \tag{14}$$

is positive (negative). The logarithmic standardised $RCR_{95/5}$ was calculated as $0.97 > 0$, for both cases A and B (Table 3), explaining all $RCR_{95/5}$ in Table 3 to be above 1.

Probabilistic Risk Characterisation Ratio

In Section 2, we alluded to the Monte Carlo interpretation of the probability of failure, as given in Eq. (2) for general exposure and no-effect distributions: the probability that random variable *X* (exposure) exceeds random variable *N* (no-effect). The result, which we interpreted as the expected risk, is just a number between 0% and 100%.

One can also examine the difference between random exposure *X* and random no-effect *N* on the $\log_{10}$ scale. This is also a random variable: $X - N$, to be interpreted as the probabilistic risk characterisation ratio, on the $\log_{10}$ scale ([5]):

$$\begin{aligned} RCR &= \frac{10^X}{10^N}, \\ \log_{10}(RCR) &= X - N. \end{aligned} \tag{15}$$

Now, PEC and PNEC, formerly considered as values, resolve into exposure and no-effect random variables with distributions.

For general, as yet unspecified, independent exposure and no-effect distributions, the density function of $\log_{10}(RCR)$ is known as the so-called *convolution* integral ([19], p. 185; [20], p. 137):

$$\mathrm{PDF}_{\log_{10} RCR}(v) = \int_{-\infty}^{\infty} \mathrm{PDF}_N(x-v) \cdot \mathrm{PDF}_X(x)dx\,, \tag{16}$$

while the cumulative distribution function is

$$\mathrm{CDF}_{\log_{10} RCR}(v) = \int_{-\infty}^{\infty} \left(1-\mathrm{CDF}_N(x-v)\right) \cdot \mathrm{PDF}_X(x)dx\,. \tag{17}$$

This expresses the probability density and cumulative distribution of $\log_{10}(RCR)$ at values $v = x - n$, in terms of the densities of the component distributions in the risk characterisation.[3] Cumulative $\log_{10}(RCR)$ distribution functions have been evaluated in both environmental and human risk assessment ([21]; [22]; [23]).

The exceedence probability function, i.e. the probability of $\log_{10}(RCR)$ to exceed zero, follows from Eq. (17):

$$1-\mathrm{CDF}_{\log_{10} RCR}(v) = \int_{-\infty}^{\infty} \mathrm{CDF}_N(x-v) \cdot \mathrm{PDF}_X(x)dx \tag{18}$$

([4]).

If we compare Eq. (18) to Eq. (3) for the probability of failure, or expected risk, we see that the probability of $\log_{10}(RCR)$ to exceed zero is equal to the expected risk in risk characterisation. This is a very important result.

Consequently, the probability of random variable RCR to exceed trigger value 1 is equal to the expected risk, as graphically displayed in the Van Straalen plots (Figs C and D). Note that this holds for arbitrary exposure and no-effect distributions.

As an example $\log_{10}(RCR) = X - N$ distribution, we return to the canonical risk characterisation model of two independent Normal distributions for exposure and no-effect, Eqs.(5). The difference between two Normal distributions is again Normal ([19], p. 194, [24]). The Normal distribution of $\log_{10}(RCR)$ is known to be

[3] *RCR* is denoted in italics, as it is a random variable, in contrast to the deterministic RCR.

$$\log_{10}(RCR) \sim \text{Normal}\left(\mu_X - \mu_N,\ \sqrt{\sigma_X^2 + \sigma_N^2}\right). \tag{19}$$

If we standardize to the no-effect distribution, Eq. (7), we get:

$$\log_{10}(RCR) \sim \text{Normal}\left(\mu_{X*},\ \sqrt{\sigma_{X*}^2 + 1}\right). \tag{20}$$

The probability of the probabilistic RCR to exceed trigger value 1, i.e. the probability of its logarithm to exceed 0, is equal to the expected risk for the canonical model, cf. Eq. (6):

$$\Pr[RCR > 1] = \Pr[\log_{10}(RCR) > 0] = \Phi\left(\frac{\mu_X - \mu_N}{\sqrt{\sigma_X^2 + \sigma_N^2}}\right), \tag{21}$$

or, for the standardized case, cf. Eq. (8):

$$\Pr[RCR > 1] = \Pr(\log_{10}(RCR) > 0) = \Phi\left(\frac{\mu_{X*}}{\sqrt{\sigma_{X*}^2 + 1}}\right). \tag{22}$$

The parameters of the $\log_{10}(RCR)$ Normal distribution for case A and B in Table 2 are given in Table 6; the distributions are plotted in Fig..

Table 6

Normal distribution parameters for the probabilistic $\log_{10}$ risk characterisation ratio (RCR), Eq. (19) and (20). Case A and B are defined in Table 2. The worst-case shift of the deterministic $\log_{10}(\text{RCR}_{95/5})$, Eq. (10), with respect to the mean of the probabilistic $\log_{10}(RCR)$ and its ratio to the standard deviation are used in the main text to explain the position of the deterministic RCR relative to the variability of the probabilistic RCR.

		Case A		Case B	
	Parameters	Raw	Standard.	Raw	Standard.
Mean	$\mu_X - \mu_N$	–0.750	–1.500	–3.000	–1.500
St.Deviation	$\sqrt{\sigma_X^2 + \sigma_N^2}$	0.559	1.118	2.236	1.118
Worst-case Shift	$1.645 \cdot (\sigma_X + \sigma_N)$	1.234	2.468	4.935	2.468
Ratio	Shift/St.Dev.	2.207	2.207	2.207	2.207
	Pr[$RCR > 1$] %	8.99	8.99	8.99	8.99

Fig. shows probabilistic logarithmic risk characterisation ratio plots, for cases A and B (Table 2), both for the raw (un-standardized) data, as well as standardized to the no-effect distribution (parameters in Table 6). Fig. corresponds to Fig. R.19-6 in the ECHA Uncertainty Analysis Guidance ([2], p. 31).

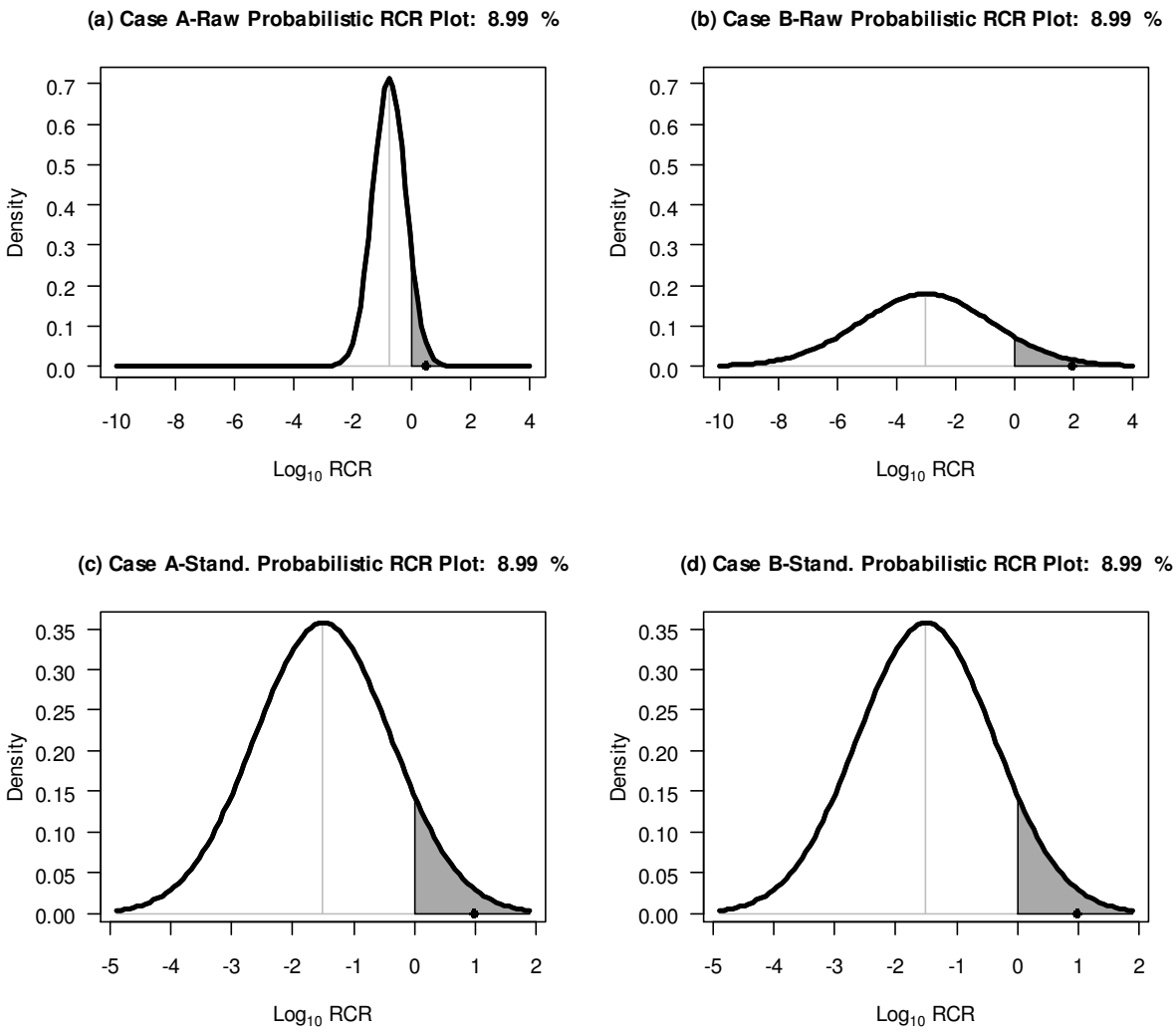


Fig. E. Normal distribution probabilistic $\log_{10}$ risk characterisation ratio (RCR) plots, cases A and B (Table 2), both for un-standardized (raw) data, as well as standardized to the no-effect distribution (parameters in Table 6). The shaded area is the probability of the RCR to exceed trigger value 1, and equals the expected risk (ER): 8.99% in each case. The dots are the deterministic $\log_{10}RCR_{95/5}$ values calculated in Table 3: 0.48 in (a), 1.94 in (b), and 0.97 in (c) and (d). The standardized plots reveal that case A and B are identical essentially. These Figs. are examples of Fig. R.19-6 in the ECHA Uncertainty Analysis Guidance ([2], p. 31).

The lower panels of Fig. E, the standardized cases, show that cases A and B of Table 2 and Table 6 are really the same. The dots in Fig. are the deterministic RCRs, i.e. $\log_{10}(RCR_{95/5})$, as calculated in Eq. (10) and (14). The respective $\log_{10}$-values are: (a) 0.48; (b) 1.94; (c) and (d) 0.97, cf. Table 3. The value 0.97 in (c) and (d) represents the standardised deterministic RCR: $\log_{10}(RCR^{*}_{95/5})$.

We can now analyze, how the probabilistic RCR and the deterministic RCR relate, and how this in fact explains the seemingly disparate values of the latter, and, even more important, how this can be repaired.

Thus, we compare the deterministic $\log_{10}(RCR_{95/5})$ in Eq. (10), with the probabilistic $\log_{10}(RCR)$ from Eq. (19). The leading term of the deterministic version, $\mu_X - \mu_N$, is the mean of the probabilistic version. The worst-case assumptions, due to the 95th percentile of exposure and the 5th percentile of no-effect, add $1.645 \cdot (\sigma_X + \sigma_N)$ to the mean, which is called the 'worst-case shift' in Table 6.

The ratio of this shift to the standard deviation of the probabilistic RCR distribution on the $\log_{10}$-scale equals

$$\frac{1.645 \cdot (\sigma_X + \sigma_N)}{\sqrt{\sigma_X^2 + \sigma_N^2}} = \frac{1.645 \cdot (\sigma_{X*} + 1)}{\sqrt{\sigma_{X*}^2 + 1}}. \qquad (23)$$

Since this ratio only depends on the standardized exposure standard deviation, the relative position of the deterministic $\log_{10}$RCR on the probabilistic $\log_{10}RCR$ axis will be the same for two cases that are identical after standardisation. For $\sigma_{X*} = 0.5$, this ratio equals 2.207 (cf. Table 6). Fig. visualizes this equivalence, as the deterministic value occupies the same position relative to the distribution and to the trigger value.

In Table 3, we saw that comparing deterministic RCRs is very difficult, as seemingly different cases lead to different RCRs, which, however, after standardisation to the no-effect distribution, turn out to be identical. The expected risk already shows its equivalence before standardisation, though. Probabilistic RCR values experience the same problem, as the location and scale of the distribution depend on the particular case parameterisations. However, the *probability* of RCR-values to exceed the trigger value 1 is the same for cases equivalent through scaling. The maths reveals that this probability is equal to the expected risk.

Level 2 worst-case deterministic RCRs *can* be quantitatively compared, if we standardise them by scaling exposure to the no-effect distribution. However, this would imply using distribution means and variability information from a (preliminary) Level 3 probabilistic assessment. In fact, we must recognise that the estimation of a worst-case $RCR_{95/5}$ by assessment of the variability of exposure and no-effect involves Level 3 considerations as well.

References

1. ECHA *Guidance on information requirements and chemical safety assessment. Part E: Risk Characterisation*; European Chemicals Agency (ECHA): Helsinki, 2008.
2. ECHA *Guidance on information requirements and chemical safety assessment. Chapter R.19: Uncertainty analysis*; European Chemicals Agency (ECHA): Helsinki, 2008.
3. Van Leeuwen, C. J., General introduction. In *Risk Assessment of Chemicals: An Introduction, 2nd Edition*, Van Leeuwen, C. J.; Vermeire, T. G., Eds. Springer: Dordrecht, 2007; pp 1-36.
4. Aldenberg, T.; Jaworska, J. S.; Traas, T. P., Normal species sensitivity distributions and probabilistic ecological risk assessment. In *Species Sensitivity Distributions in Ecotoxicology*,

Posthuma, L.; Suter II, G. W.; Traas, T. P., Eds. CRC/ Lewis Publishers: Boca Raton, 2002; pp 49-102.
5. Verdonck, F. A. M.; Aldenberg, T.; Jaworska, J. S.; Vanrolleghem, P. A., Limitations of current risk characterization methods in probabilistic environmental risk assessment. *Environmental Toxicology and Chemistry* **2003,** *22*, 2209-2213.
6. EUFRAM *Concerted action to develop a European Framework for probabilistic risk assessment of the environmental impacts of pesticides. Volume 1. Framework and worked examples*; Central Science Laboratory, York, UK. www.eufram.com: 2006.
7. EUFRAM *Concerted action to develop a European Framework for probabilistic risk assessment of the environmental impacts of pesticides. Volume 2. Detailed Reports on Role, Methods, Reporting & Validation*; Central Science Laboratory, York, UK. www.eufram.com: 2006.
8. Verdonck, F. A. M.; Souren, A.; Van Asselt, M.; Van Sprang, P. A.; Vanrolleghem, P. A., Improving uncertainty analysis in European Union risk assessment of chemicals. *Integrated Environmental Assessment and Management* **2007,** *3*, (3), 333-343.
9. R-Project *R: A language and environment for statistical computing*, R Development Core Team, R Foundation for Statistical Computing, http://www.R-project.org.: Vienna, Austria, 2011.
10. Posthuma, L.; Suter II, G. W.; Traas, T. P., *Species Sensitivity Distributions in Ecotoxicology*. Lewis Publishers: Boca Raton, 2002.
11. Solomon, K. R.; Brock, T. C. M.; De Zwart, D.; Dyer, S. D.; Posthuma, L.; Richards, S. M.; Sanderson, H.; Sibley, P. K.; Van den Brink, P. J., *Extrapolation Practice for Ecotoxicological Effect Characterization of Chemicals*. SETAC, CRC Press: Boca Raton, 2008.
12. Ang, A. H.-S.; Tang, W. H., *Probability Concepts in Engineering Planning and Design. Volume II: Decision, Risk, and Reliability*. Wiley: New York, 1984.
13. Freudenthal, A. M.; Garrelts, J. M.; Shinozuka, M., The analysis of structural failure. *Journal of the Structural Division, ASCE* **1966,** *92*.
14. Van Straalen, N. M., New methodologies for estimating the ecological risk of chemicals in the environment. In *Proceedings 6th International IAEG Congress*, Price, D. G., Ed. Balkema: Rotterdam, 1990; pp 165-173.
15. Suter II, G. W.; Vaughan, D. S.; Gardner, R. H., Risk assessment by analysis of extrapolation error: a demonstration for effects of pollutants on fish. *Environmental Toxicology and Chemistry* **1993,** *2*, 369-378.
16. Van Straalen, N. M., Theory of ecological risk assessment based on species sensitivity distributions. In *Species Sensitivity Distributions in Ecotoxicology*, Posthuma, L.; Suter II, G. W.; Traas, T. P., Eds. Lewis Publishers: Boca Raton, 2002; pp 37-48.
17. Giesy, J. P.; Solomon, K. R.; Coats, J. R.; Dixon, K. R.; Giddings, J. M.; Kenaga, E. E., Chlorpyrifos: ecological risk assessment in North American aquatic environments. *Reviews in Environmental Contamination and Toxicology* **1999,** *160*, 1-129.
18. ECHA *Guidance on information requirements and chemical safety assessment. Chapter R.10: Characterisation of dose [concentration]-response for environment*; European Chemicals Agency (ECHA): Helsinki, 2008.
19. Mood, A. M.; Graybill, F. A.; Boes, D. C., *Introduction to the Theory of Statistics*. Third ed.; McGraw-Hill: Tokyo, 1974; p 564.
20. Hsu, H. P., *Probability, Random Variables, and Random Processes*. Schaum's Outline Series, McGraw-Hill: New York, 1997.
21. Jager, T.; Vermeire, T. G.; Rikken, M. G. J.; Van der Poel, P., Opportunities for a probabilistic risk assessment in the European Union. *Chemosphere* **2001,** *43*, 257-264.

22. Bodar, C. W. M.; De Bruijn, J. H. M.; Vermeire, T. G.; Van der Zandt, P. T. J., Trends in risk assessment of chemicals in the European Union. *Hum. Ecol. Risk Asess.* **2002,** *8*, (7), 1825-1843.
23. Vermeire, T. G. Evaluating Uncertainties in an Integrated Approach for Chemical Risk Assessment under REACH: More Certain Decisions? Utrecht University, Utrecht, 2009.
24. Slob, W., Uncertainty analysis in multiplicative models. *Risk Analysis* **1994,** *14*, (4), 571-576.

Appendix 4. Supplementary material for case studies

Contents

Chemicals selected

Triazoles: Tebuconazole (CAS 107534), Difenoconazole (CAS 119446), Triazamate (CAS 112143), Bromuconazole (CAS 116255), and Metconazole (CAS 125116)

PBDEs: BDE-03 (4-MonoBDE, CAS 101553); BDE-28 (2,4,4'-TriBDE, CAS 41318756); BDE-47 (2,2',4,4'-TetraBDE, CAS 5436431

Supporting Information: Uncertainties in a Triazole Risk Assessment based on QS(A)PRs

Table S1:

QS(A)PR model predictions with their SEP (or geometric mean and standard deviation for the half-lives in water) and distribution for Tebuconazole, Triazamate, Bromuconazole, Difenoconazole, and Metconazole

Input Parameter	Distribution	Descriptive Statistic	Tebuconazole	Triazamate	Bromuconazole	Difenoconazole	Metconazole
Log K_{oc} (L/kg)	Student-t	Pred. SEP	$3.42 \cdot 10^{0}$ $5.58 \cdot 10^{-1}$	$2.43 \cdot 10^{0}$ $5.58 \cdot 10^{-1}$	$4.24 \cdot 10^{0}$ $5.58 \cdot 10^{-1}$	$4.83 \cdot 10^{0}$ $5.60 \cdot 10^{-1}$	$3.52 \cdot 10^{0}$ $5.59 \cdot 10^{-1}$
Log WS (mg/L)	Student-t	Pred. SEP	$1.53 \cdot 10^{0}$ $5.74 \cdot 10^{-1}$	$2.68 \cdot 10^{0}$ $6.33 \cdot 10^{-1}$	$1.43 \cdot 10^{0}$ $5.92 \cdot 10^{-1}$	$1.17 \cdot 10^{-1}$ $5.70 \cdot 10^{-1}$	$1.54 \cdot 10^{0}$ $5.75 \cdot 10^{-1}$
MP (°C)	Student-t	Pred. SEP	$1.01 \cdot 10^{2}$ $3.15 \cdot 10^{1}$	$8.28 \cdot 10^{1}$ $3.24 \cdot 10^{1}$	$1.00 \cdot 10^{2}$ $3.19 \cdot 10^{1}$	$1.09 \cdot 10^{2}$ $3.17 \cdot 10^{1}$	$1.19 \cdot 10^{2}$ $3.18 \cdot 10^{1}$
Log VP (mmHg)	Student-t	Pred. SEP	$-7.94 \cdot 10^{0}$ $9.69 \cdot 10^{-1}$	$-8.06 \cdot 10^{0}$ $9.83 \cdot 10^{-1}$	$-6.95 \cdot 10^{0}$ $9.90 \cdot 10^{-1}$	$-9.74 \cdot 10^{0}$ $9.95 \cdot 10^{-1}$	$-8.66 \cdot 10^{0}$ $9.73 \cdot 10^{-1}$
Log $1/k_{OH}$ (cm^3s^{-1} /mol)	Student-t	Pred. SEP	$1.09 \cdot 10^{1}$ $4.38 \cdot 10^{-1}$	$9.54 \cdot 10^{0}$ $4.38 \cdot 10^{-1}$	$1.17 \cdot 10^{1}$ $4.41 \cdot 10^{-1}$	$1.15 \cdot 10^{1}$ $4.43 \cdot 10^{-1}$	$1.10 \cdot 10^{1}$ $4.38 \cdot 10^{-1}$
$t_{1/2,water}$ (d^{-1})	Log-normal	Geo. mean Geo. sd.	$8.50 \cdot 10^{1}$ $3.51 \cdot 10^{0}$	$1.49 \cdot 10^{1}$ $7.45 \cdot 10^{0}$	$8.50 \cdot 10^{1}$ $3.51 \cdot 10^{0}$	$8.80 \cdot 10^{1}$ $3.46 \cdot 10^{0}$	$8.50 \cdot 10^{1}$ $3.51 \cdot 10^{0}$
-Log LC50 O. Mykiss (mol/L)	Student-t	Pred. SEP	$5.11 \cdot 10^{0}$ $4.98 \cdot 10^{-1}$	$5.56 \cdot 10^{0}$ $5.25 \cdot 10^{-1}$	$5.01 \cdot 10^{0}$ $5.05 \cdot 10^{-1}$	$5.69 \cdot 10^{0}$ $4.95 \cdot 10^{-1}$	$5.00 \cdot 10^{0}$ $4.93 \cdot 10^{-1}$
-Log EC50 D. Magna (mol/L)	Student-t	Pred. SEP	$4.83 \cdot 10^{0}$ $5.06 \cdot 10^{-1}$	$3.65 \cdot 10^{0}$ $5.65 \cdot 10^{-1}$	$4.41 \cdot 10^{0}$ $5.12 \cdot 10^{-1}$	$5.34 \cdot 10^{0}$ $5.31 \cdot 10^{-1}$	$4.84 \cdot 10^{0}$ $5.11 \cdot 10^{-1}$
-Log EC50 P.Subcapitata (mol/L)	Student-t	Pred. SEP	$5.36 \cdot 10^{0}$ $5.38 \cdot 10^{-1}$	$5.30 \cdot 10^{0}$ $5.56 \cdot 10^{-1}$	$4.80 \cdot 10^{0}$ $5.28 \cdot 10^{-1}$	$5.92 \cdot 10^{0}$ $5.86 \cdot 10^{-1}$	$5.43 \cdot 10^{0}$ $5.40 \cdot 10^{-1}$

Table S2:

HAT values of the QS(A)PR predictions indicating whether a prediction is in or outside the applicability domain (AD).

Outcome Measure	in or out AD	Tebuconazole	Triazamate	Bromuconazole	Difenoconazole	Metconazole
Log K_{oc} (L/kg)	HAT AD	0.007 in	0.006 in	0.007 in	0.015 in	0.009 in
Log WS (mg/L)	HAT AD	0.059 in	0.296 ** out	0.118 in	0.045 in	0.073 in
MP (°C)	HAT AD	0.027 in	0.089 in	0.058 in	0.045 in	0.047 in
Log VP (mmHg)	HAT AD	0.075 in	0.107 in	0.123 in	0.134 in	0.084 in
Log $1/k_{OH}$ (cm^3s^{-1} /mol)	HAT AD	4 in	0 in	3 border	7 out	4 in
-Log LC50 O. Mykiss (mol/L)	HAT AD	0.034 in	0.075 in	0.056 in	0.043 in	0.029 in
-Log EC50 D. Magna (mol/L)	HAT AD	0.028 in	0.107 in	0.034 in	0.066 in	0.036 in
-Log EC50 P.Subcapitata (mol/L)	HAT AD	0.0519 in	0.2498 in	0.0703 in	0.1009 in	0.0551 in

**very high HAT value

Supporting information: Non-testing versus testing based risk assessment on three PBDEs

Table S3: Comparison of non-testing versus testing based (5^{th}; 50^{th}; 95^{th}) percentiles of log PEC for PBDEs with different scenarios in fresh water (based on approx. 1ton/year).

Substance	Without Photolysis (–log PEC, mg/L)		With Photolysis (–log PEC, mg/L)	
	QSPR-based	Exp. and QSPR-based	QSPR-based	Exp. and QSPR-based
BDE-03	9.61;10.28;11.11	N/A	9.71;10.36;11.17	N/A
BDE-28	8.07 ;8.58 ;9.11	8.33;8.46; 8.76	8.48; 8.99; 9.52	8.77; 8.87; 9.10
BDE-47	7.85; 8.33; 8.84	8.27; 8.44; 8.75	8.53; 9.02; 9.53	8.99; 9.13; 9.41

Table S4

substance	MW	CAS							
BDE-003	249.1	101-55-3							
		EC50 (mg/l)		log(EC50) mg/l	logHC5acute	gHC5chronic	logPNEC	PNEC(SSD)	PNEC (mg/l)
Predicted Acute (ECOSAR)	LC50 Fish (96hr)	0.500		-0.301					
	LC50 Daphnid (48hr)	0.430		-0.367					
	EC50 Algae (96hr)	0.740		-0.131	**-0.50**	**-2.50**	**-3.20**	**6.29E-04**	**4.30E-04**
				mean SSD pred acute	-0.266		**-2.97**	mean SSD pred acute	
				SD SSD pred acute	0.122		**0.12**	SD SSD pred acute	
Experimental Acute	LC50 Daphnid (48hr)	0.28-0.48	0.360	-0.444					
	LC50 Fish (96hr) (lepomis macrochir	4-6.1	5.900	0.771	**-1.84**	**-3.84**	**-4.54**	**2.86E-05**	**3.60E-04**
				mean SSD exp acute	0.164		**-2.54**	mean SSD exp acute	
				SD SSD exp acute	0.859		**0.86**	SD SSD exp acute	
		ChV (mg/l)	EC (mg/l)	g (NOEC)					
Predicted Chronic (ECOSAR)	LC50 Fish (96hr)	0.063	0.045	-1.351					
	LC50 Daphnid (48hr)	0.080	0.057	-1.247					
	EC50 Algae (96hr)	0.460	0.325	-0.488		**-1.94**	**-2.64**	**2.28E-03**	
				mean SSD pred chronic		-1.029	**-1.73**	mean SSD pred chronic	
				SD SSD pred chronic		0.471	**0.47**	SD SSD pred chronic	
BDE-028	406.9	41318-75-6							
		EC50 (mg/l)		log(EC50) mg/l	logHC5acute	gHC5chronic	logPNEC	PNEC(SSD)	PNEC (mg/l)
Predicted Acute (ECOSAR)	LC50 Fish (96hr)	0.121		-0.917					
	LC50 Daphnid (48hr)	0.121		-0.917					
	EC50 Algae (96hr)	0.310		-0.509	**-1.24**	**-3.24**	**-3.94**	**1.15E-04**	**1.21E-04**
				mean SSD pred acute	-0.781		**-3.48**	mean SSD pred acute	
				SD SSD pred acute	0.236		**0.24**	SD SSD pred acute	
Experimental Acute	LC50 Nitocra spinipes (48hr)	0.072	0.072	6.54					
	NOEC (6 days) Nitocra spinipes	0.0002	0.0002	9.10					**7.20E-05**
		ChV (mg/l)	EC (mg/l)						
Predicted Chronic (ECOSAR)	LC50 Fish (96hr)	0.016	0.011	-1.946					
	LC50 Daphnid (48hr)	0.026	0.018	-1.736					
	EC50 Algae (96hr)	0.226	0.160	-0.796		**-2.68**	**-3.38**	**4.17E-04**	
				mean SSD pred chronic		-1.493	**-2.19**	mean SSD pred chronic	
				SD SSD pred chronic		0.612	**0.61**	SD SSD pred chronic	
BDE-047	485.8	5436-43-1							
		EC50 (mg/l)		log(EC50) mg/l	logHC5acute	gHC5chronic	logPNEC	PNEC(SSD)	PNEC (mg/l)
Predicted Acute (ECOSAR)	LC50 Fish (96hr)	0.024		-1.620					
	LC50 Daphnid (48hr)	0.026		-1.585					
	EC50 Algae (96hr)	0.027		-1.569	**-1.64**	**-3.64**	**-4.34**	**4.56E-05**	**2.40E-05**
				mean SSD pred acute	-1.591		**-4.29**	mean SSD pred acute	
				SD SSD pred acute	0.026		**0.03**	SD SSD pred acute	
Experimental Acute	LC50 (Fundulus heterocliticus)	>0.1	0.100	-1.000					
	NOEC (Fundulus heterocliticus)		0.050	-1.301					**1.00E-04**
		ChV (mg/l)	EC (mg/l)						
Predicted Chronic (ECOSAR)	LC50 Fish (96hr)	0.003	0.002	-2.673					
	LC50 Daphnid (48hr)	0.007	0.005	-2.305					
	EC50 Algae (96hr)	0.087	0.062	-1.211		**-3.54**	**-4.24**	**5.78E-05**	
				mean SSD pred chronic		-2.063	**-2.76**	mean SSD pred chronic	
				SD SSD pred chronic		0.761	**0.76**	SD SSD pred chronic	

Sensitivity analyses

Sensitivity analysis is considered an important tool to show how the results of a risk assessment are dependent on the assumptions made (Aven, 2010). Just looking at the uncertainty in the input parameter is not enough to judge its influence on risk assessment. Even though uncertainty in a QSAR predicted parameter may be relatively small, it can have a large effect on the risk assessment due to potential error propagation. The risk assessment example in Walker (chapter 11) pointed out the effect of error (or uncertainty) propagation, which means that parameters often have a large influence on the uncertainty in the output. This needs to be considered when applying the QSAR and becomes crucial when QSAR models are developed using descriptor data estimated by other QSARs, thus increasing the potential for error propagation.

Sensitivity analyses of individual QSAR models with regard to their contributions in the overall risk assessment framework were carried out in for different purposes and with different measures. Sensitivity analysis are useful for many purposes and a good reference is Saltelli's book "Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models" (Saltelli, 2004), but with respect to the contribution to QSAR models on the overall risk assessment the sensitivity analysis should aim to evaluate robustness of the decisions on which the QSAR models are allowed to influence. Saltelli (Saltelli, 2002) writes "sensitivity analysis should not focus on the model output as such, but rather on the answer that the model is supposed to provide". Within CADASTER three kinds of purposes have been identified, namely:

1. To evaluate the contribution of QSAR uncertainty to total uncertainty in risk assessment output e.g. to identify important sources of uncertainty in a risk assessment,
2. To compare the consequences of using non-testing information instead of testing information, e.g. to decide whether further testing is needed or the evaluate the possible errors of using non-testing information,
3. To model the influence of compounds-specific properties and activities on risk for a class of compounds e.g. for risk screening or prioritization for testing.

Examples of different sensitivity measures are:

Correlation

The Spearman rank correlation between an uncertain input parameter and an output variable. Spearman rank correlations are used to identify the influence (sign and magnitude) of each input

parameter on the output. Since it is a correlation it can only detect monotonic influences, i.e. positive or negative.

Variance decomposition

For example, derive how much of the variance in the output that can be explained by variance in the input.

Pinched uncertainty versus full uncertainty

The influence of uncertainty in an input parameter can be found by comparing a statistic based on the outcome of the risk assessment before and after pinching the uncertainty in an input parameter. The outcome statistics are the inter-percentile width (e.g. between 95^{th} and 5^{th} percentiles), the percentile of the output distribution up which decisions are base, or if the pinching result in the passing of a decision threshold.

Value of information

The value of reducing uncertainty seen by the change in expected loss based on a decision analysis with the probabilistic risk as input (Jaworska et al., 2010).

Strategy to design a cross-compound sensitivity analysis

The purpose with the strategy is support analysis where Probabilistic Risk Assessments (PRA) are done on a representative set of compounds from a chemical class (such as one of the four CADASTER classes). Multiple compound PRA are intended to be used to search for compound specific characteristics that can be related to regulatory variables of risk, and show which parameters that are most influential on both risk and its uncertainty, and to open up for a general comparison of PRA based on testing and non-testing (model predictions) information.
The suggested strategy is as follows: for a given chemical class the PRA is based on K QSAR models to predict input parameters p1 to pK. These are the design variables. As design criteria we want to have compound spread out over the space. Here a D-optimal design is a good suggestion, where we also uses square and cross terms of p1 to pK to avoid only sampling from the outer range. In order to do apply the design we need a list of possible candidate compounds for the class. CADASTER has such list for at least the classes with PFCs and BTAZs. Let us say that we have a list of N compounds. For each of the QSAR models in the PRA we calculate the predicted means of that parameter, and decide to what extent each compound is in the AD. We

also flag when a compound is present in the training or testing set used to build the QSAR model. Here it may be necessary to select other design variables such as molecular weight. The compounds cas number (or similar), predicted input parameters p1 to pK (the design variables), and flags on AD (1 if inside AD) and QSARdata (1 if part of training or testing data set) can be placed in an Excel file. We apply D-optimal design limited to finite candidate set e.g. using the facility provided on the CADASTER web tool. The strategy is to seek D-optimality and have a high proportion of selected compounds that are in the AD and with QSAR data. In summary to implement this we need to

1) Compile a list of candidate compounds
2) For every QSAR in the PRA generate descriptors and predict the corresponding input parameter
3) Flag if compound is in or out of AD
4) Flag if compound is in the QSAR data
5) Decide how many compounds to select
6) Put in an Excel sheet and apply D-optimal design to suggest a candidate selection e.g. using the CADASTER web tool.

Monte Carlo simulations

If the amount of uncertainty about the true value of the parameter is known, the influence on the model output can be quantified. This type of analysis is sometimes referred to as probabilistic (as opposed to deterministic). One technique that can be used for probabilistic analysis is Monte Carlo simulation. In a Monte Carlo simulation, the deterministic values of variable and/or uncertain input parameters are replaced by probability distributions. A few commonly used distribution types in Monte Carlo simulation are:

- The normal distribution is a symmetrical distribution that is characterized by two parameters: the mean and the standard deviation. The body length of the individuals in a population is an example of a variable that is often characterized using a normal distribution.
- The lognormal distribution is a positively skewed distribution which is characterized by two parameters: the mean and the standard deviation. The lognormal distribution is widely used to describe natural phenomena that are restricted to positive values. Most values cluster around or below the mean, but there are few outliers with an extreme value. The distribution is called lognormal because the logarithm of the individual values follows a normal distribution.

- The triangle distribution is an artificial distribution type which is used if only three characteristic values are available: a minimum, a most likely and a maximum value. The probability density (or frequency) distribution of this distribution type has a triangle shape.
- A uniform distribution is also an artificial distribution type which is used if only two values are available: a minimum and a maximum. All values between the minimum and the maximum are assigned the same probability density, resulting in a probability density (or frequency) distribution that looks like a rectangle.

After distributions have been specified for all uncertain model parameters, the Monte Carlo simulation can be initiated. In the first simulation step, the computer draws one value from each predefined probability distribution. These values (also referred to as "realizations") are used to calculate the output parameters and the result is stored. This process of drawing values and calculating output parameters is repeated many times, e.g. 10,000. Each repetition is called an iteration. The set of output values gathered after 10,000 iterations gives an impression of the variation in the output caused by uncertainty in the input parameters. Crystal Ball® is a Microsoft Excel® plug-in that can be used to perform Monte Carlo simulations. Monte Carlo simulations of an assessment model specified in Excel is alternatively possible by the RDCOMClient package in the R which is a free software environment for statistical computing and graphics (R Development Core Team, 2008). R compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. A code for Monte Carlo simulations in R has been developed in the project and it has been checked to be compatible to Crystal ball for the simulations done in the case-studies of PBDEs.

To correlate or not

Correlation between input parameters may alter the outcome of a Monte Carlo simulation, and it is relevant to ask when correlations between QSAR predicted input parameters should be considered. We have several options:

1. Do not consider correlations between input parameters
2. Consider correlation between input parameters predicted based on a common QSAR models such as log Kow
3. Consider correlation between input parameters that are believed to be correlated

4. Consider correlation between input parameters which predictive error is believed to be correlated.

In the view that QSAR predictions are precise and uncertainty refers to errors and not actual variation in parameters, it can be argued that correlating parameters is unnecessary if QSAR models are independent, since the predictive uncertainty describes the error and not the actual value taken by the parameter. Correlation between QSAR predicted input parameters may be important when there is a suspicion of predictive errors being correlated, but we are not aware of anyone discussing such systematic bias in QSAR predictions.

Lognormal distribution

Uncertainty in biodegradation was assigned a lognormal distribution. The lognormal distribution in Crystal ball uses the arithmetic mean μ and standard deviation σ, by default. For applications where historical data are available, it is more appropriate to use the logarithmic mean μ_{log} and logarithmic standard deviation σ_{log}, or the geometric mean and geometric standard deviation. The relations between these estimates are

$$\mu_{log} = \ln\left(\frac{\mu}{\exp\left({(\sigma_{log})^2}/{2}\right)}\right)$$

$$\sigma_{log} = \sqrt{ln\left(\exp\left(2 * ln\frac{\sigma}{\mu}\right) + 1\right)}$$

Geometric mean = $\exp(\mu_{log})$, Gemotric std dev = $\exp(\sigma_{log})$, Median = geometric mean.

Aven T. (2010) Some reflections on uncertainty analysis and management. *Reliability Engineering & System Safety* 95: 195-201.

Jaworska J, Gabbert S and Aldenberg T. (2010) Towards optimization of chemical testing under REACH: A Bayesian network approach to Integrated Testing Strategies. *Regulatory Toxicology and Pharmacology* 57: 157-167.

R Development Core Team. (2008) R: A language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.

Saltelli A. (2002) Sensitivity analysis for importance assessment. *Risk Analysis* 22: 579-590.

Saltelli A. (2004) *Sensitivity analysis in practice : a guide to assessing scientific models,* Hoboken, N.J.: Wiley.