Uncertainty in QSAR Predictions

Ullrika Sahlin

School of Natural Sciences, Linnaeus University, Kalmar, Sweden; Lund University, Centre of Environmental and Climate Research, Lund, Sweden

Summary — It is relevant to consider uncertainty in individual predictions when quantitative structure–activity (or property) relationships (QSARs) are used to support decisions of high societal concern. Successful communication of uncertainty in the integration of QSARs in chemical safety assessment under the EU Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) system can be facilitated by a common understanding of how to define, characterise, assess and evaluate uncertainty in QSAR predictions. A QSAR prediction is, compared to experimental estimates, subject to added uncertainty that comes from the use of a model instead of empirically-based estimates. A framework is provided to aid the distinction between different types of uncertainty in a QSAR prediction: quantitative, i.e. for regressions related to the error in a prediction and characterised by a predictive distribution; and qualitative, by expressing our confidence in the model for predicting a particular compound based on a quantitative measure of predictive reliability. It is possible to assess a quantitative (i.e. probabilistic) predictive distribution, given the supervised learning algorithm, the underlying QSAR data, a probability model for uncertainty and a statistical principle for inference. The integration of QSARs into risk assessment may be facilitated by the inclusion of the assessment of predictive error and predictive reliability into the "unambiguous algorithm", as outlined in the second OECD principle.

Key words: *applicability domain, knowledge-based uncertainty, probabilistic risk assessment, regression, uncertainty analysis.*

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

Introduction

Despite their many benefits, the widespread use of chemicals in our society is a serious threat to health and the environment. Given that high values are at stake, the prevailing knowledge gaps have led chemical regulation to adopt a cautious risk management approach (1). Cautiousness is, among other things, implemented in the treatment of knowledge-based uncertainty in hazard or risk assessments. In order to speed up the process of chemical regulation, and save resources and reduce animal testing, the EU Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) system allows the use of non-*in vivo* testing (i.e. *in vitro* or *in silico*-derived) information to support chemical safety assessments (2). The basis for knowledge goes from strong to weaker when *in vivo* experimental data are replaced by non-*in vivo* testing information (Appendix 1, Figure A1.1). The use of non-*in vivo* testing information is constrained by a lack of knowledge of the extent to which 'weaker' data are acceptable as a base (or for informing) risk management decisions. The added uncertainty that stems from the use of non-*in vivo* testing information, instead of *in vivo*-derived information, should be characterised in order to compensate for the lower strength of the information and to evaluate the confidence in the resulting decision support. Thus, the characterisation of uncertainty in non-*in vivo* testing information offers a way of promoting and facilitating the integration of non-*in vivo* testing methods in chemical regulation.

Quantitative structure–activity relationships (QSARs) produce non-*in vivo* testing information through the use of analogy predictions. A QSAR is quantitative, which means that it uses mathematical models, or algorithms, to make predictions. QSARs are some of the non-*in vivo* testing methods currently used in chemical regulation (3). QSAR predictions are commonly reported as a point estimate (4), but the added uncertainty compared to *in vivo* testing information is rarely given a full characterisation in hazard and risk assessment. For example, de Roode *et al*. (5) showed that it is common practice to apply QSARs to chemicals for which the model is not reliable, leading to low accuracy in the QSAR predictions. The application of QSARs in risk assessment raises the need to consider uncertainty in predictions in relation to the intended use of the QSAR (6).

A prerequisite for a successful consideration of uncertainty is to understand what is meant by a QSAR prediction and its uncertainty. QSARs can roughly be divided into two kinds of predictions:

classifications and regressions (7). A classification places a compound in one of at least two categories, such as biodegradable or not biodegradable. Uncertainty assessments for classifications can be based on contingency table statistics (8), and a Bayesian framework is used to assess the probability of making a correct classification, given descriptor values (7). In QSAR terminology, a 'regression' means modelling a continuous response, such as the boiling point. The assessment of uncertainty in QSAR regressions is hampered by the prevailing point prediction view on QSAR predictions. Assessment is further complicated, given the wide array of modelling approaches that, more or less, model uncertainty in predictions (9). As a start, there is a need for a common understanding of uncertainty and its assessment, along with a broad perspective on modelling algorithms.

The aim of this paper is to provide an overview of the assessment of uncertainty in predictions from QSAR regressions. The paper is structured to let the reader reflect on the meaning of a QSAR prediction, and its associated uncertainty, in a conceptual framework for its characterisation. The focus here is on the assessment of uncertainty, given the way that QSAR modelling commonly considers available *in vivo* experimental data — i.e. as experimentally-based point estimates. Even though it is relevant, the uncertainty that comes from both variability and measurement errors in *in vivo* experimental data cannot be quantitatively assessed, unless it is modelled and recognised in the QSAR data (see, for example, Tebby and Mombelli [10]). Furthermore, the focus on quantitative approaches, algorithms and metrics does not reduce the need for a validation of the reliability of QSAR predictions that is based on the knowledge of chemicals and their interactions with the environment and organisms, and is essential for a successful implementation of QSARs in chemical regulation. We end the discussion by looking at the current requirements for reporting uncertainty in QSAR predictions under the REACH system.

What is a QSAR Prediction?

A structure–activity relationship (SAR) is a non-*in vivo* testing method of predicting a chemical activity, or property, based on analogy reasoning, by saying that there is a correlation between a chemical's structure, its physical or chemical properties, and a measured biological activity (7, 11). A SAR is quantitative (i.e. is a QSAR), when there is a model to relate quantitative descriptors of chemical structure (X) to a quantitative measure of a property or activity (Y; 12). Models for QSARs can roughly be divided into parametric and non-parametric. A parametric model is defined by a mathematical equation. For example, a linear regression predicting Y conditioned to the value(s) on descriptor(s) X can be defined as:

$$Y|X = \beta_0 + \beta_1 X \qquad \text{[Equation 1]}$$

where β_0 and β_1 are parameters. A non-parametric model can be a prediction rule to arrive at $Y|X$.

These approaches have their strengths and weaknesses. Parametric models can be easier to interpret mechanistically, while non-parametric models are less constrained and may be better for extracting signals from QSAR data (13). The wide array of approaches for the use of parametric and non-parametric models in QSAR modelling exemplifies supervised learning algorithms. Supervised learning uses data on items for which response Y is known, to train an algorithm to predict items for which the response is unknown (14). A non-parametric model, such as nearest neighbour averaging, is trained to find an optimal way of defining what the nearest neighbours are. Training also occurs when fitting, or updating, a parametric model. The linear regression coefficients in Equation 1 are commonly specified on the basis of a QSAR data set. For example, an Ordinary Least Squares (OLS) regression is fitted by minimising the average squared difference between an observation and the model value (e.g. Montgomery *et al.* [15]). With the aim of facilitating the integration of QSARs in decision-making, and the treatment of uncertainty, we propose a conceptual framework to define a QSAR prediction from a statistical perspective, in which we suggest viewing QSARs as derived by supervised learning algorithms (Table 1, ID 1–8). Note that possible consensus modelling is to be seen as part of the supervised learning algorithm.

Let us assume that Z denotes an unobserved response, and z the future value of Z once it has been observed; the error is the difference between the unobserved value z and the prediction of Z. Predictions are made for different reasons, which affect the information that is needed on the error. Consider the purpose of validating a model. The data on compounds not included in the training data set can be used for the external validation of a supervised learning algorithm. This means applying the model to predict compounds from a selected external data set, in order to derive performance measures based on a comparison of external predictions and known values. Depending on the choice of the performance measure, it may be enough to know only the values of Z and z. We refer to that as a QSAR prediction "without uncertainty" (Table 1, ID 9–13).

When the purpose is to apply the model for the prediction of a specific query compound to inform decision-making, more characteristics of the error $Z - z$ are needed. In this case, a QSAR prediction is derived from a supervised learning algorithm that,

Table 1: A framework for the definition of a QSAR prediction

ID	Description	Notation
1	Quantitative descriptor(s)	X [a]
2	Quantitative measure of a property or activity	Y
3	QSAR	Y | X [b]
4	Known values of Y	y
5	Specific values for the ith compound	$\{y,X\}_i$
6	QSAR data is a set of n compounds for which the quantitative property or activity is known	$\{y,X\}_{i = 1:n}$
7	Supervised learning algorithm	A
8	QSAR model is a supervised learning algorithm and QSAR data	Y | X, $\{y,X\}_{i = 1:n}$, A
9	Property or activity of query compound	Z
10	Known values of Z	z
11	Quantitative descriptor(s) of query compound	W [a]
12	QSAR external data is a set of nExt compounds with known values but not used to train the model	$\{z,W\}_{j = 1:nExt}$
13	**QSAR prediction without uncertainty** is a supervised learning algorithm, QSAR data and descriptors for the query compound	Z | W, $\{y,X\}_{i = 1:n}$, A
14	Algorithm to assess uncertainty	U
15	**QSAR prediction with uncertainty** is a supervised learning algorithm that includes the assessment of uncertainty, QSAR data (sometimes including external QSAR data) and descriptors for the query compound	a) Z | W, $\{y,X\}_{i = 1:n}$, AU or b) Z | W, $\{y,X\}_{i = 1:n}$, $\{z,W\}_{j = 1:nExt}$, AU

[a] *We assume that the values of X and W are always known.*
[b] *The symbol '|' stands for 'given'.*
Depending on the purpose of the prediction, the framework for the definition of a QSAR prediction can be with or without uncertainty.

in addition to the QSAR model (Table 1, ID 3), also includes an approach to assessing uncertainty (Table 1, ID 14), resulting in a QSAR prediction "with uncertainty" (Table 1, ID 15). Uncertainty is context-specific, and the next section will discuss the characteristics of uncertainty that are useful to support the integration of QSARs into chemical safety assessment under the REACH system.

What is Meant by Uncertainty in a Prediction?

Context

The meaning of QSAR uncertainty needs to be understood in relation to risk assessment in practice. Risk assessment is a science-based approach, but nevertheless, the characterisation of uncertainty rests upon assumptions and decisions taken by the risk assessor. In general, one can view uncertainty in a risk assessment as a reflection of the risk assessor's uncertainty in predicted quantities to express risk, given the available background knowledge (16). Thus, uncertainty, in the context of risk assessment, should be seen as a subjective judgement that can change when new data, models or expert knowledge add relevant information to the background knowledge. In order to tally the interpretation of uncertainties in input with that of the output in an assessment, the final interpretation of uncertainty in input parameters supported by QSAR predictions is an expression of the risk assessor's uncertainty in the value of that parameter, which can more-or-less be taken directly from the predictive inference of a QSAR (17). However, even though uncertainty with the purpose of supporting risk assessment should be assessed to reflect the risk assessor's uncertainty in a QSAR prediction, its assessment can more-or-less be based on data and probabilistic modelling, and is not constrained by the QSAR algorithm. Even though uncertainty is subjective, there is a need for unambiguous algorithms for its assessment, in order to inform and support the final choice in its characterisation.

Uncertainty in QSAR predictions is a major concern, especially when these predictions could influence human and animal lives, as well as the safety

of environmental systems. Approaches to the assessment of uncertainty must therefore be as correct as possible, which poses a need to evaluate the quality of assessments of uncertainty, of which transparency, repeatability and rationale are relevant characteristics. We seek useful guidance for the characterisation of uncertainty in a QSAR prediction, in terms of its definition, characterisation, assessment and evaluation, with the purpose of supporting decision-making (Appendix 1, Table A1.1).

Chemical safety assessment asks for knowledge-based uncertainty in QSAR prediction in relation to information requirements and uncertainty analysis. The information must fulfil several requirements, before it can support a chemical safety assessment (18). In particular, to make sure that a reliable prediction is obtained, the European Chemicals Agency (ECHA) requires that the validity of the selected QSAR has been assessed, and requires verification that the chemical to be evaluated falls within the QSAR's applicability domain (12). The latter requirement is a qualitative characterisation of uncertainty (which we refer to as predictive reliability), and is related to the use of a QSAR for the prediction of a specific chemical.

Uncertainty analysis is conducted to evaluate the need to refine an assessment, and to inform the risk assessor of the magnitude of the risk. Uncertainty can be analysed in three tiers, with increasing precision of the quantification of uncertainty — from deterministic, worst (plausible case), to probabilistic. Probabilistic means that uncertainty is given a full characterisation by specifying and quantifying, through probabilities, the likelihoods of all possible values for an input parameter. Uncertainty analysis distinguishes between parameter uncertainty, model uncertainty and scenario uncertainty (19). Out of these three, uncertainty in a QSAR prediction is most closely associated with parameter uncertainty, which according to the ECHA "is the uncertainty involved in the specification of numerical values". Parameter uncertainties include measurement errors, sample uncertainty, selection of the data used for assessing the risk, and extrapolation uncertainty, which can be a consequence of "the use of alternative methods (e.g. a QSAR model, an *in vitro* test, or read-across for similar substances) or the use of assessment factors (e.g. inter-species, intra-species, acute to chronic, route to route, lab to field extrapolation)". The need to quantify uncertainty in QSAR predictions by probabilities was pointed out by Walker *et al.* (20), by suggesting "that errors need to be evaluated when applying QSARs by providing confidence intervals that take into consideration the uncertainty associated with the estimate", and we note that a confidence interval presumes an underlying probability distribution.

The need to characterise uncertainty in a QSAR prediction is both qualitative, by expressing our confidence in the use of a prediction to support decision-making, and quantitative, by believing in the values that the predicted property or activity may take after observation (9). Qualitative uncertainty is related to the reliability in individual predictions, and we refer to this as 'predictive reliability' to avoid confusion with 'reliability' in a more general context (Appendix 1, Table A1.1). Quantitative uncertainty is associated with the predictive error, which is a measure that describes the distance between a point prediction and the actual value, and may change from compound to compound. Both predictive reliability and predictive error are sensitive to the degree to which a prediction is an extrapolation from a model.

The following two examples are provided to aid the discussion of the two types of uncertainties.

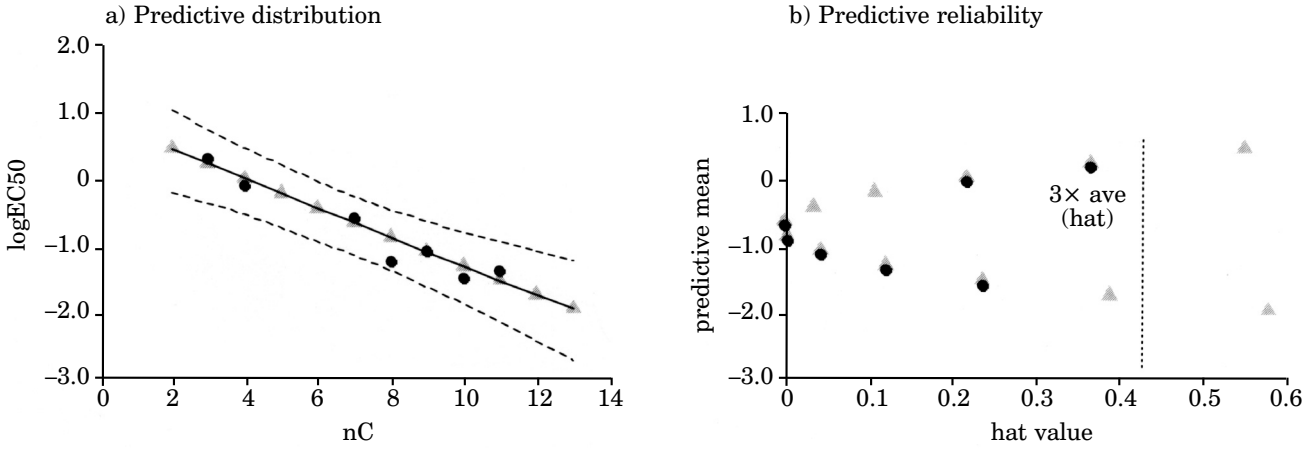
Example 1

The first example is that of a QSAR for the prediction of aquatic toxicity (log EC50 [48 hours] for *Chydorus sphaericus*), based on the fluorinated carbon chain length (nC) for perfluorinated carbons (PFCs). The QSAR training data is taken from Ding *et al.* (4), and contains seven PFCs with the following carbon chain lengths: 3, 4, 7, 8, 9, 10 and 11. This is a typical small QSAR data set that poses challenges to the quantitative assessment of uncertainty. Due to the limited QSAR data, the model has not gone through any external validation. Nevertheless, predictions may be supported by the clear mechanistic understanding associated with this QSAR. Ding *et al.* (4) fitted a QSAR by OLS regression, and provided predictions without uncertainty (Table 1, ID 13). QSAR predictions with uncertainty (Table 1, ID 15) were modelled by Bayesian linear regression (Figure 1a). Predictive reliability was evaluated by considering the range of descriptor values, and by comparing hat values, which are located on the diagonal of the information matrix, to a cut-off of three times the average hat value (Figure 1b).

Example 2

The second example consists of QSAR predictions for aquatic toxicity to fish (LC50 [96 hours] for *Onchorhynchus mykiss*) for triazoles and benzotriazoles, by using descriptors generated by DRAGON 6.0. This example is representative of a case where the number of descriptors is large, which has led to the great variety of QSAR algorithms for finding the best combination of descriptors for predictions and techniques to evaluate the degree of extrapolation. In this example, a multivariate analysis by Partial

Figure 1: PFC QSAR predictions with uncertainty



——— = *predictive mean;*---------= *95% confidence region;* ● = *training data;* ▲ = *predictions.*

The prediction in the PFC example is characterised by a) the predictive distribution and b) calculated hat values (i.e. distances to the applicability domain) to evaluate predictive reliability. Three times the average hat value for the training data set constitutes a possible cut-off for acceptable predictive reliability.

Least Squares regression was used to reduce a high dimensional descriptor space to a few latent variables for capturing the most relevant variation within chemical structures. Predictions with uncertainty were then generated by Bayesian modelling on the latent variables by Markov Chain Monte Carlo (MCMC) sampling of a Bayesian Lasso (21). A MCMC sample of the predictive distribution (Appendix 1, Figure A1.2a) can be generated by a short script in open access software, such as R (22), that can be attached to the reporting of a QSAR. Predictions with uncertainty and predictive reliability were generated for training and test data sets (Figure 2).

Qualitative Uncertainty: Predictive Reliability

Definition

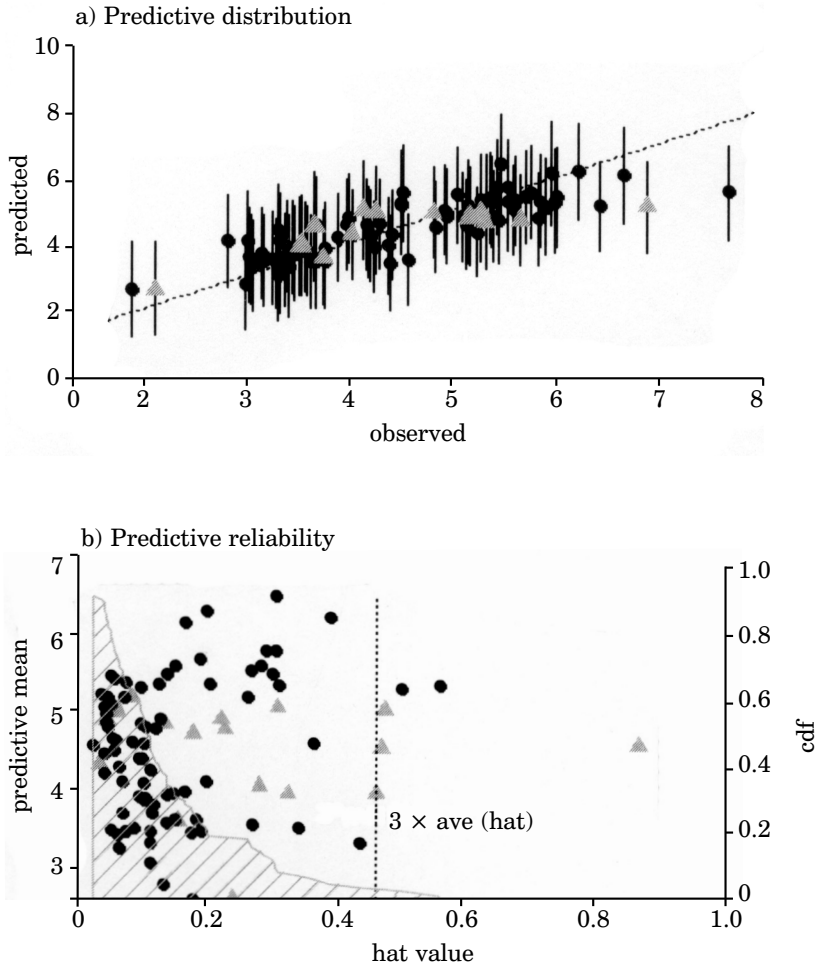
Predictive reliability is a judgement of the strength of the non-*in vivo* testing information, which depends on the general reliability of a QSAR, and whether the QSAR is suitable for predicting the compound in question. Predictive reliability is commonly evaluated on the basis of validation of the QSAR model and the extent to which an individual prediction is an extrapolation from the QSAR. Before being considered in risk assessment, the acceptance of a QSAR should have been verified according to the so-called OECD principles (23), which, for example, ask for measures of predictivity, goodness-of-fit and robustness (7, 24). To aid this judgement, various measures can be combined into an index that includes crucial factors in relation to the intended application (25). Only QSARs with an approved QSAR Model Reporting Format (QMRF; 26) are accepted for use in the REACH system, and we refer to more details in the QMRF on how to validate a QSAR.

Given an acceptable QSAR, the qualitative uncertainty that needs to be characterised is the extent to which a QSAR is an extrapolation. The reliability in an individual prediction is evaluated in relation to the chemical domain over which a QSAR has been built. This domain is defined by molecular descriptors, structural fragments (e.g. chemicals fragments that are not represented in the QSAR training data set), the domain defined by mechanisms such as mode-of-actions associated with the QSAR, and, if relevant, metabolic domains (18). The extent to which a specific QSAR is applicable to predict a specific compound is judged according to several criteria. Below, we refer to the establishment of a model's so called applicability domain based on measured similarity between descriptor values only, which is to be used in combination with other criteria to evaluate the chemical domain.

Characterisation and assessment

Predictive reliability can be assessed in several ways, which are roughly divided into measures of

Figure 2: BTAZ QSAR predictions with uncertainty



——— = *95% confidence interval;* ● *=training data;* ▲ *= test data; cdf = cumulative distribution function.*

A prediction in the BTAZ example is characterised by a) its predictive distribution for training and test data sets, and b) its predictive reliability either evaluated by the distance to the applicability domain (hat value) or the density of the applicability domain seen by the empirical distribution (marked area) of hat values for the training data.

extrapolation and measures of performance. The former includes various metrics based on descriptors for describing the similarity between a compound and the QSAR training data set (27, 28). The applicability domain is a region in chemical domain determined by the training set and the algorithm that is used. It is used to evaluate the degree to which a compound is an extrapolation, given that it is judged to be inside the chemical domain based on other information. For example, even though the QSARs in Example 1 are derived for compounds representative of the chemical class PFCs, there are some PFCs that become further extrapolations compared to other PFCs.

According to OECD principles, the applicability domain must have been given a clear definition. Metrics such as the distance or density of the applicability domain are relative, and need to be compared to a critical threshold, to judge whether the predictive reliability is acceptable or not (see example in Figures 1b and 2b). Where to position a cut-off is a subjective judgement that allows the assessment of predictive reliability to be adapted to its context.

Performance measures include non-probabilistic measures of predictive performance, such as the standard deviation in ensemble predictions (29), or sensitivities (30), and probabilistic measures of predictive performance, such as the local coverage (hit rates or empirical confidence levels). Standard deviations in ensemble predictions, or sensitivities, can point out the predictions that are expected to be relatively more uncertain, without being a measure of the magnitude of uncertainty related to error. Such performance measures are useful, as

they not only capture the degree of extrapolation (since predictions, in general, will be more uncertain with increasing extrapolation), but may also identify regions of the applicability domain that are more uncertain compared to others. As a side-effect, trying to explain why some predictions are given with relatively high uncertainty can be useful for the identification of important knowledge gaps in the chemistry behind a QSAR.

Probabilistic measures of predictive performance can be assessments of the confidence in the prediction. In Bayesian statistics, confidence denoted by $1 - \alpha$ is the probability that an interval covers the true value, where α is the probability of the true value falling outside the interval. Empirical estimates of confidence, the (empirical) coverage, can be assessed by counting the number of hits in a data set, where each data point is associated with an interval with the same level of confidence (Figure 3). A measure of predictive performance is, compared to a metric of the applicability domain, more informative in judging predictive reliability, but the latter is often used to assess the former. For example, local coverage can be assessed by dividing the applicability domain into regions based on the distance from the centre of the applicability domain. This was done by Tong *et al.* (31), who assessed coverage for a given confidence level over different regions of the applicability domain, defined by extrapolation measured by the proportion of items in the training data set that are further away than the item to be predicted. Their finding — that coverage was lowest for the most extreme region of the applicability domain — is intuitive, but difficult to put into practice. Good empirical estimates of predictive performance in the most extreme regions cannot be obtained, since, by definition, there are fewer data points in those regions. Predictive reliability aims to support a qualitative uncertainty that comes from an extrapolation. No matter how much we try, there will always be a need for judgements to fully assess predictive reliability.

Evaluation

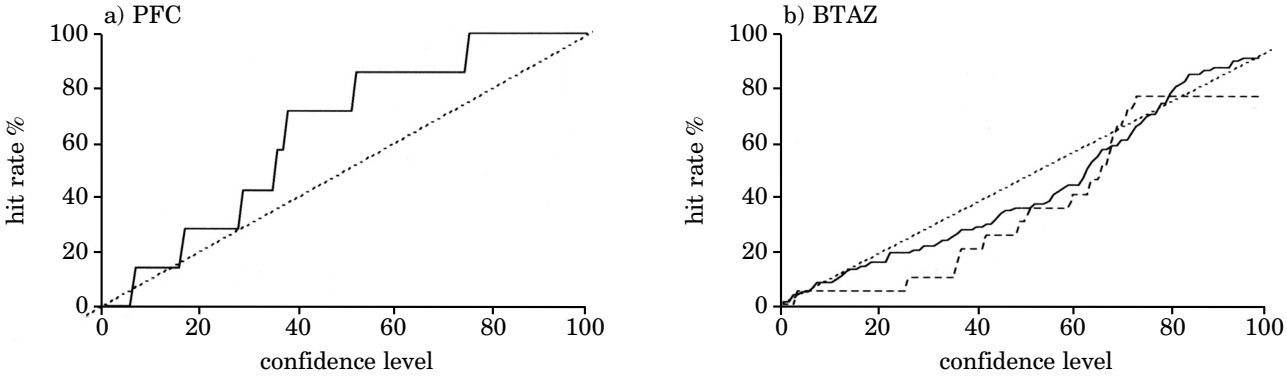
Many measures of predictive reliability are relative. A measure of predictive reliability can be inferred from the model's ability to describe what it is supposed to describe, based on a correlation between another measure of the applicability domain, or on observations of errors in associated predictions. Performance measures can be empirically evaluated by sampling or re-sampling. When predictive reliability is judged to be unacceptably low, an alternative is to use the QSAR prediction in a weight-of-evidence (WoE) approach, where a consensus prediction is reached after consideration of all the available information, including that from other non-*in vivo* testing methods (32).

Quantitative Uncertainty: Predictive Error

Definition and characterisation

Quantitative uncertainty is related to the error in a prediction, which for a regression is the differ-

Figure 3: Empirical coverage



a) ——— *= training;* ····· *= 1:1.*
b) ——— *= training;* - - - - - - - *= test;* ····· *= 1:1.*

The coverage was evaluated for a) the training data set in the PFC example and b) the training and test data sets for the BTAZ example.

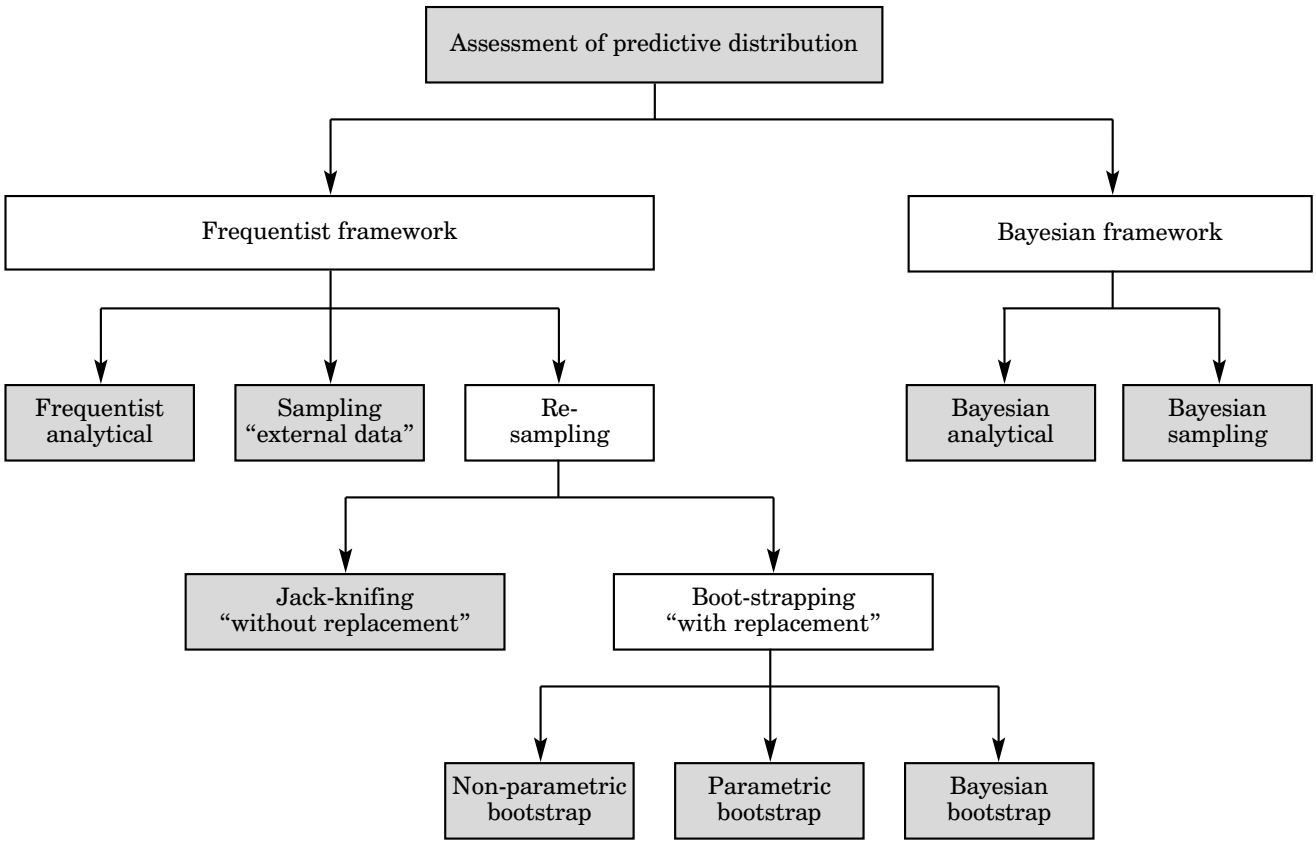
ence between $Z - z$, where Z has not been observed. The aim is to quantify uncertainty in the error associated with an individual prediction by a probability distribution that we believe reflects the values that the error in a prediction may take after the property or activity has been observed. The probability distribution for the predictive error is, in this context, referred to as the predictive distribution.

Assessment

Approaches to assessing the predictive distribution can be made under different statistical frameworks, by using more or less data-intensive methods, with more or less specific probabilistic models for uncertainty (Figure 4). Sampling theory assesses the predictive error based on a representative sample. Such (frequentist) inference rests on assumptions of independent and, for example, identically distributed observations, in combination with a probabilistic assumption of characteristics of uncertainty. Whenever there is doubt in any of these assumptions, users of frequentist inference run into problems. The Bayesian paradigm for inference *assigns*, instead of *assuming*, a probabilistic model for observations, and assigns models for uncertainty in parameters (so-called 'priors'). Bayesian inference uses Bayes' rule to update expert knowledge with information from empirical observations. The result is a well-defined probabilistic model of uncertainty. In case of doubt, 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 judgement.

A third approach is to assign a probability distribution for the predictive error, based on expert judgement only. This can, for example, be based on experience from *in vivo* testing, or based on combinations of different sources of information. Sampling theory and solid expert judgement can be seen as extreme kinds of Bayesian inference. Expert judgement of uncertainty can be seen as Bayesian modelling of the error with no updating, i.e. based on prior distribution only. The difference between frequentist inference and Bayesian inference in parametric and non-hierarchical linear regressions is usually negligible, given a large data set or non-informative priors. Under a weak signal

Figure 4: Approaches for the assessment of the predictive distribution of a QSAR prediction



in the QSAR data, the differences between predictive distributions generated by the Bayesian Lasso and the Student-*t* distribution following frequentist statistical inference from an OLS prediction (e.g. Montgomery *et al.* [15]) in Figure A1.2 (Appendix 1), may be highly sensitive to the specification of informative priors. Note that the comparison between OLS and Bayesian inference is made for a given set of descriptors. Perhaps the largest difference between the two approaches lies in the selection of descriptors.

In order to assess the predictive distribution, a probability model can be included in, or added to, the supervised learning algorithm for the prediction (Table 1, ID 14 and ID 15). There is a need to acknowledge and discuss the conditions and suitability of different approaches for including the assessment of predictive distribution in the QSAR algorithm. Some examples are described in Appendix 4 in CADASTER Deliverable 4.2 "Guidance on using QSAR models for probabilistic risk assessment" (available at www.cadaster.eu).

Evaluation

In the same way as a QSAR algorithm needs to be verified, the algorithm for assessing the predictive error from the predictive distribution needs to be evaluated for a particular QSAR. Sampling and resampling are data-rich methods, suitable when there are many data points in the domain. As the size of the samples becomes smaller, the need for parametric specification of a QSAR model, including the probabilistic model for uncertainty, increases. Some modellers may feel uncomfortable with adding prior information or making parametric specifications. It is therefore important that a model is evaluated by, for example, comparing it to a less constrained alternative. The burden of information and resources for calculations should be in relation to the required level of detail of the quantified uncertainty (19, 33). Before suggesting a resource-demanding and complex approach for the assessment of uncertainty, it may be relevant to evaluate it in relation to a simpler assessment. A candidate for a rule of thumb, which can be read out from information provided in the QMRF, is to assign a Gaussian distribution with the point prediction, and a reported value on mean square error of prediction as its first and second moment (9).

An algorithm that provides predictions with uncertainty can be evaluated by looking for a one-to-one correspondence between empirical coverage for different levels of confidence and the assigned confidence levels (see examples in Figure 3). There are also relative methods based on likelihood-based measures that can be used to make relative comparisons between alternative algorithms (e.g. Tetko *et al.* [29]), or as weights for the model averaging of alternative predictions in consensus modelling (34).

Current requirements for reporting uncertainty in QSAR predictions

The communication of uncertainty may be assisted by the use of formats for documentation of a QSAR prediction that provide useful information to facilitate the assessment of uncertainty. Nowadays, the documentation of a QSAR prediction is framed by the QMRF (26) and the QSAR Prediction Reporting Format (QPRF; 35). The QMRF is designed to ensure that the QSAR to be integrated in the chemical safety assessment under REACH fulfils the OECD principles, which means that it must have: a) a defined endpoint; b) an unambiguous algorithm; c) a defined applicability domain; d) appropriate measures of goodness-of-fit, robustness and predictivity; and e) if possible, a mechanistic interpretation.

We asked whether the requirement for an unambiguous algorithm also includes the assessment of predictive error, either as the predictive distribution or as a point estimate of the variance in the predictive distribution. According to the current version of the QPRF, it does not — a prediction as a point value is enough. The document states: "Report the predicted value (including units) obtained from the application of the model to the query chemical" (paragraph 3.2.d). The uncertainty of the prediction should, "if possible, be commented on, taking into account relevant information (e.g. variability of the experimental results)" (paragraph 3.4). According to the required information on uncertainty, there is no need to specifically address the added quantitative uncertainty when going from *in vivo* testing to non-*in vivo* testing information. Furthermore, the addition of the term "if possible" introduces a loose end to this requirement. It might be that comments on the uncertainty in a prediction would be more successful when given with reference to a conceptual framework of what uncertainty means, such as the one provided in this paper.

Another question is the extent to which the limits of the defined applicability domain in the QMRF restrict the possibility of context-specific judgement on the predictive reliability of a QSAR prediction. According to the QPRF, they do not. It is sufficient to discuss whether the query chemical falls within the applicability domain of the model as defined in the corresponding QMRF (paragraph 3.3.a). The QPRF contains an optional fourth paragraph, where the adequacy of a QSAR prediction is considered in relation to the regulatory context. Here, the regulatory decision is allowed to frame the interpretation of the results from the model, and both quantitative and qualitative uncertainty can be reported.

We conclude that the current reporting formats, including requirements for assessing and reporting qualitative uncertainty related to the extrapolation and the applicability domain, are well established, and there are possibilities for context-specific subjective judgement for its characterisation. We did not find any mandatory requirement for reporting quantitative uncertainty related to the accuracy or error in the predictions. A possible reason is that it is difficult to specify exactly what to report, as the characterisation of uncertainty is different between regressions and classifications, and for qualitative, semi-quantitative and quantitative models. The reporting of uncertainty belongs to the prediction format, but it may be relevant to report algorithms for assessing uncertainty in the model format as well. Introducing a requirement for the assessment of uncertainty during later stages of the production of a QSAR prediction, i.e. after the QMRF has been made, may hinder risk assessment, since many frequently used algorithms are adopted to generate point predictions. Thus, introducing the need to consider uncertainty in the final stage of the production of a QSAR prediction may be problematic, when the QSAR algorithm must be altered and thereby the QRMF is no longer valid. The option is to alter the QSAR (keeping the QSAR data fixed) and use it in a WoE approach, which forces the need to also consider other available information as well. Another option is to build QSAR models that predict with uncertainty (Table 1, ID 15), i.e. that from the start include an unambiguous and evaluated algorithm for the assessment of uncertainty in predictions.

Conclusions

The integration of QSARs into probabilistic risk assessment is possible, given proper assessments of quantitative and qualitative uncertainties in a QSAR prediction, referred to in this paper as predictive error and predictive reliability (Appendix 1, Table A1.1). Uncertainty in QSAR predictions is context-dependent, and is closely linked to the background knowledge in a risk assessment, since we have the option of strengthening background knowledge by further testing. Predictions must be performed with care, and the use of different bases for predictive inference must properly acknowledge the limitations in QSAR data. Both quantitative and qualitative uncertainties are influenced by the extent to which a prediction is an extrapolation for the chemical domain of a QSAR.

One aim within the CADASTER project has been to provide guidance on the integration of QSARs in hazard and risk assessment. Principles and bases in the assessment of uncertainty in QSAR predictions are often described and communicated for classification models (20). The focus in CADASTER has therefore been toward the development and evaluation of approaches for assessing uncertainty in predictions from QSAR regressions, i.e. models that predict a continuous endpoint as opposed to a categorical endpoint. The discussion here was aimed at reaching a common understanding and initiating a 'model-free' guidance for the assessment of uncertainty in QSAR regressions (Table 1, Figure 4 and Appendix 1, Figure A1.1). Adapting a Bayesian interpretation of uncertainty in a QSAR prediction conforms to the interpretation of uncertainty in probabilistic risk assessment. The framework provided to explain the meaning of uncertainty in a QSAR prediction is based on the idea that a separation between predictive error and predictive reliability makes it possible to both apply models and to discuss their reliability in a constructive way. We suggest that the integration of QSARs into risk assessment would be facilitated if the assessment of uncertainty was included in the unambiguous modelling algorithm, as outlined in the second OECD principle.

Acknowledgements

The author is grateful to the CADASTER partners — in particular, Tom Aldenberg, Nina Jeliazkova, Tomas Öberg and Igor Tetko. James E. Blevins and Johan Lindström contributed with useful discussions about statistical methods. This work has been funded by the European Seventh Framework Programme through the CADASTER project FP7-ENV-2007-1-212668.

References

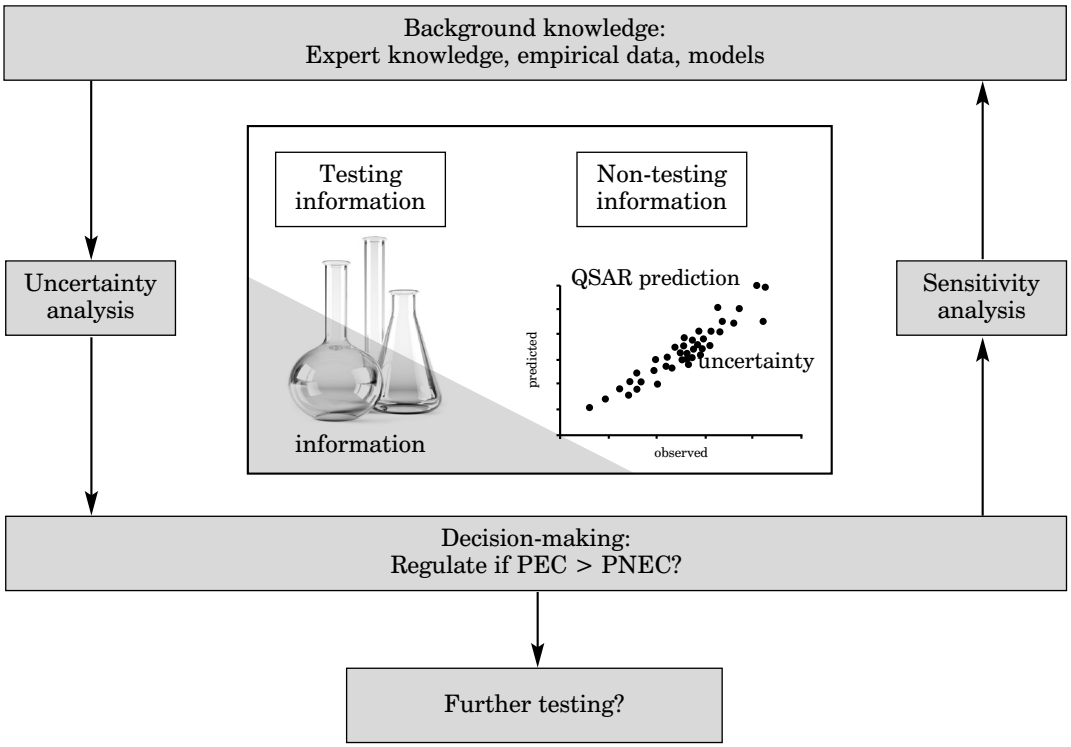
1. Sandin, P., Bengtsson, B-E., Bergman, Å., Brandt, I., Dencker, L., Eriksson, P., Förlin, L., Larsson, P., Oskarsson, A., Ruden, C.S.N.M., Södergren, A., Woin, P. & Hansson, S.O. (2004). Precautionary defaults — A new strategy for chemical risk management. *Human & Ecological Risk Assessment* **10**, 1–18.
2. European Parliament (2006). *Regulation (EC) No 1907/2006* of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending *Directive 1999/45/EC* and repealing *Council Regulation (EEC) No 793/93* and *Commission Regulation (EC) No 1488/94* as well as *Council Directive 76/769/ EEC* and *Commission Directives 91/155/EEC*, *93/67/EEC*, *93/105/EC* and *2000/21/EC*. *Official Journal of the European Union* **L396**, 30.12.06, 1–849.
3. ECHA (2011). *The Use of Alternatives to Testing on Animals for the REACH Regulation*, 4pp. Helsinki, Finland: European Chemicals Agency. Available at:

http://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2011_summary_en.pdf (Accessed 25.01.12).
4. Ding, G.H., Fromel, T., van den Brandhof, E.J., Baerselman, R. & Peijnenburg, W.J. (2012). Acute toxicity of poly- and perfluorinated compounds to two cladocerans, *Daphnia magna* and *Chydorus sphaericus*. *Environmental Toxicology & Chemistry* **31**, 605–610.
5. De Roode, D., Hoekzema, C., de Vries-Buitenweg, S., van de Waart, B. & van der Hoeven, J. (2006). QSARs in ecotoxicological risk assessment. *Regulatory Toxicology & Pharmacology* **45**, 24–35.
6. ECETOC (1998). *QSARs in the Assessment of the Environmental Fate and Effect of Chemicals*, 143pp. Brussels, Belgium: European Centre for Ecotoxicology and Toxicology of Chemicals.
7. Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T., McDowell, R.M. & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives* **111**, 1361–1375.
8. Fielding, A.H. & Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**, 38–49.
9. Sahlin, U., Filipsson, M. & Oberg, T. (2011). A risk assessment perspective of current practice in characterizing uncertainties in QSAR regression predictions. *Molecular Informatics* **30**, 551–564.
10. Tebby, C. & Mombelli, E. (2012). A kernel-based method for assessing uncertainty on individual QSAR predictions. *Molecular Informatics* **31**, 741–751.
11. Walker, J.D., Jaworska, J., Comber, M.H., Schultz, T.W. & Dearden, J.C. (2003). Guidelines for developing and using quantitative structure–activity relationships. *Environmental Toxicology & Chemistry* **22**, 1653–1665.
12. ECHA (2009). *Practical Guide 5: How to Report (Q)SARs,* 11pp. Helsinki, Finland: European Chemicals Agency.
13. Sprenger, J. (2011). Science without (parametric) models: The case of bootstrap resampling. *Synthese* **180**, 65–76.
14. Hastie, T., Tibshirani, R. & Friedman, J.H. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (2nd ed.), 746pp. New York, NY, USA: Springer.
15. Montgomery, D.C., Peck, E.A. & Vining, G.G. (2001). *Introduction to Linear Regression Analysis* (3rd ed.), 152pp. New York, NY, USA: Wiley-Blackwell.
16. Aven, T. (2010). Some reflections on uncertainty analysis and management. *Reliability Engineering & System Safety* **95**, 195–201.
17. National Research Council (2009). *Science and Decisions: Advancing Risk Assessment*, 424pp. Washington, DC, USA: National Academies Press.
18. ECHA (2008). *Guidance on Information Requirements and Chemical Safety Assessment Chapter R.6: QSARs and Grouping of Chemicals*, 134pp. Helsinki, Finland: European Chemicals Agency.
19. ECHA (2008). *Guidance on Information Requirements and Chemical Safety Assessment Chapter R.19: Uncertainty Analysis*, 36pp. Helsinki, Finland: European Chemicals Agency.
20. Walker, J.D., Carlsen, L. & Jaworska, J. (2003). Improving opportunities for regulatory acceptance of QSARs: The importance of model domain, uncertainty, validity and predictability. *QSAR & Combinatorial Science* **22**, 346–350.
21. Gramacy, R.B. & Pantaleo, E. (2010). Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing. *Bayesian Analysis* **5**, 237–262.
22. R Development Core Team (2008). *R: A language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
23. OECD (2007). *Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure–Activity Relationship [(Q)SAR] Models in the Assessment of New and Existing Chemicals*, 79pp. Paris, France: Organisation for Economic Co-operation and Development.
24. Gramatica, P. (2007). Principles of QSAR models validation: Internal and external. *QSAR & Combinatorial Science* **26**, 694–701.
25. Schultz, T.W., Netzeva, T.I. & Cronin, M.T.D. (2004). Evaluation of QSARs for ecotoxicity: A method for assigning quality and confidence. *SAR & QSAR in Environmental Research* **15**, 385–397.
26. JRC (2008). QSAR Model Reporting Format (Version 1.2), 9pp. Ispra, Italy: Institute for Health and Consumer Protection. Available at: http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/qrf/QMRF_version_1.2.pdf (Accessed 25.01.13).
27. Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D., Schultz, T., Stanton, D.W., van de Sandt, J.J., Tong, W., Veith, G. & Yang, C. (2005). Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. The report and recommendations of ECVAM Workshop 52. *ATLA* **33**, 155–173.
28. Jaworska, J., Nikolova-Jeliazkova, N. & Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *ATLA* **33**, 445–459.
29. Tetko, I.V., Sushko, I., Pandey, A.K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D. & Varnek, A. (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 & Modeling* **48**, 1733–1746.
30. Bosnic, Z. & Kononenko, I. (2009). An overview of advances in reliability estimation of individual predictions in machine learning. *Intelligent Data Analysis* **13**, 385–401.
31. Tong, W.D., Xie, W., Hong, H., Shi, L., Fang, H. & Perkins, R. (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.
32. ECHA (2010). *Practical Guide 2: How to Report Weight of Evidence,* 21pp. Helsinki, Finland: European Chemicals Agency. Available at: http://echa.europa.eu/documents/10162/13655/pg_report_weight_of_evidence_en.pdf (Accessed 17.01.13).
33. Jager, T., Vermeire, T.G., Rikken, M.G. & van der Poel, P. (2001). Opportunities for a probabilistic risk assessment of chemicals in the European

Union. *Chemosphere* **43**, 257–264.
34. Johnson, J.B. & Omland, K.S. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution* **19**, 101–108.
35. JRC (2008). *QSAR Prediction Reporting Format (QPRF) (Version 1.1, May 2008)*, 6pp. Ispra, Italy: Institute for Health and Consumer Protection. Available at: http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/qrf/QPRF_version_1%201_DEREK_SS.pdf (Accessed 25.01.13).

Appendix 1

Figure A1.1: The use of non-*in vivo* testing information for risk assessment

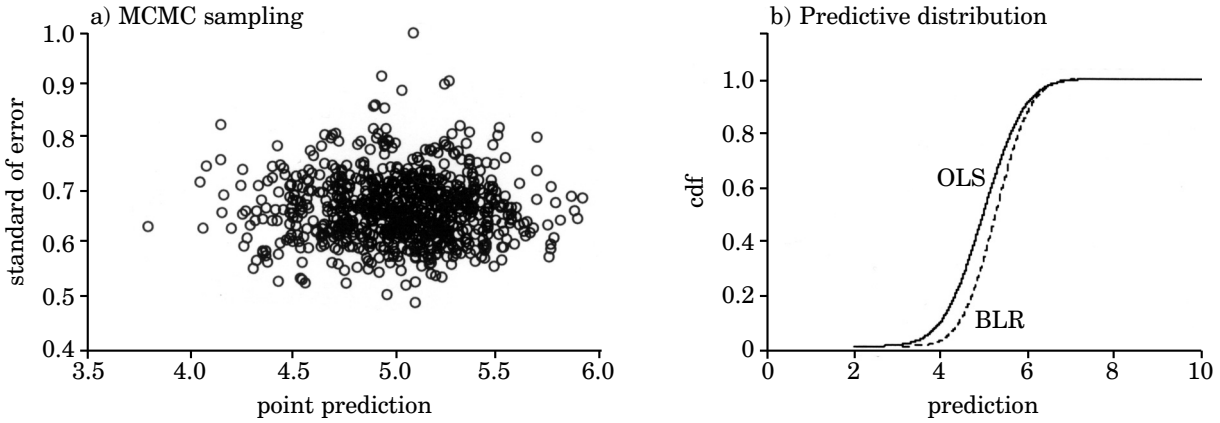


*Non-*in vivo *testing information supports risk assessment with background knowledge of lower strength compared to testing information. The influence of the added uncertainty in non-testing information on risk, and the need to obtain testing information to support a risk assessment, can be evaluated by uncertainty and sensitivity analysis and uncertainty in, for example, QSAR predictions quantified by probabilities.*

Table A1.1: The characterisation of uncertainty in a QSAR prediction

Uncertainty	Predictive error	Predictive reliability
Definition	Magnitude of the added error in a prediction compared to experimentally based estimate.	Confidence in the use of a model for predicting a specific compound.
Characteristic	Quantitative probability distribution.	Qualitative judgement of confidence (e.g. high or low).
Assessment	Probabilistic modelling of the error based on sampling, re-sampling, or probability theory, potentially in combination with expert judgement.	Confidence assessed by expert judgement (informed by relative measures such as density, distance and variation in perturbed predictions) or empirical coverage.
Evaluation	Empirical coverage for a chosen level of confidence or likelihoods for an external data set (relative).	Difficult to evaluate a qualitative judgement *per se*. Alternative measures of predictive reliability can be evaluated for their abilities to capture a trend in predictive error or perceived lower reliability.

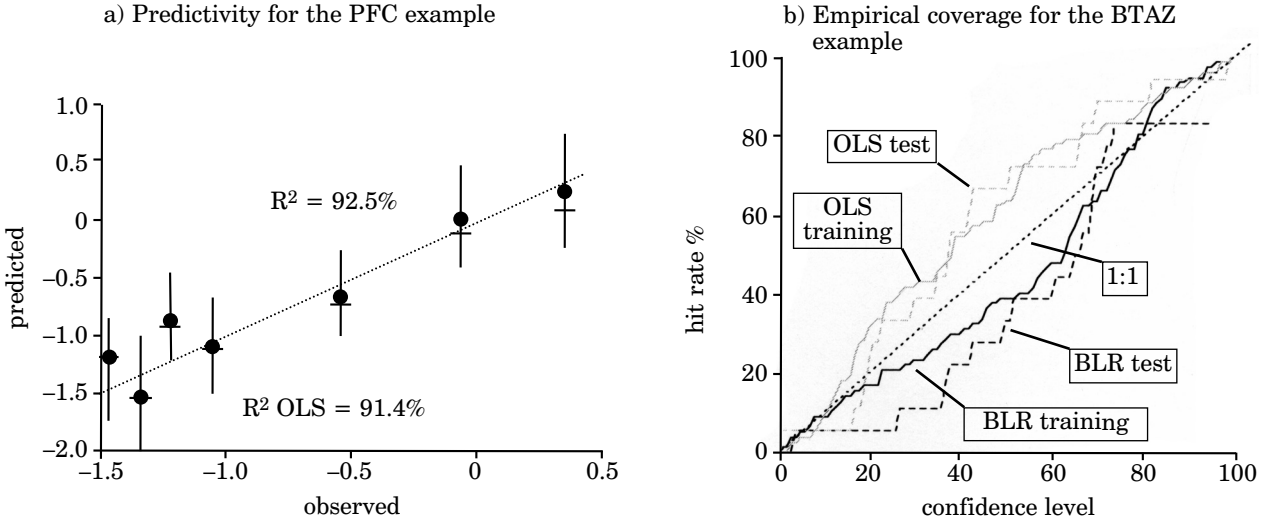
Figure A1.2: Illustration of a BTAZ QSAR prediction



BLR = Bayesian Linear Regression; cdf = cumulative distribution function.

*a) A Markov Chain Monte Carlo (MCMC) sample of a triazole predicted by the BTAZ QSAR, and b) the resulting predictive distribution which is compared to a predictive distribution based on an Ordinary Least Squares (OLS) regression fitted to the same descriptors. The predictive distribution for the OLS is assessed by assuming errors to be independent and identically distributed as Normal Distribution with fixed variance, which generates a predictive distribution being a Student-*t*.*

Figure A1.3: A comparison between Bayesian Linear regression and Ordinary Least Squares regression



a) Even though the differences between Bayesian and Frequentist point predictions are small (represented by the filled dots and vertical lines = expected posterior with confidence interval from BLR; – = OLS point predictions), these rely on different statistical principles and only BLR quantifies uncertainty. b) The performance of the uncertainty assessment can be evaluated by relative comparison of alternative approaches, in this case, BLR is compared to OLS. In this example, OLS results in relatively wider predictive distributions, while one could consider changing the kind of probability distribution for the BLR.

Table A1.2: Further details on alternative approaches to assessment of the predictive distribution

Bayesian analytical	Result for Student-*t*
Bayesian sampling	e.g. MCMC
Sampling	External data set
Re-sampling without replacement	LOO
Re-sampling with replacement	
Non-parametric bootstrap	Modified residuals by using the empirical distribution of the residuals.
Parametric bootstrap	Modified residuals assuming a Gaussian distribution.
Bayesian bootstrap	Link to Bayesian sampling.