JRC Scientific and Technical Reports

A Framework for assessing *in silico* Toxicity Predictions: Case Studies with selected Pesticides

Andrew Worth, Silvia Lapenna, Elena Lo Piparo, Aleksandra Mostrag-Szlichtyng and Rositsa Serafimova

EUR 24705 EN - 2011

JRC EUROPEAN COMMISSION

ihcp Institute for Health and Consumer Protection

The mission of the JRC-IHCP is to protect the interests and health of the consumer in the framework of EU legislation on chemicals, food, and consumer products by providing scientific and technical support including risk-benefit assessment and analysis of traceability.

European Commission
Joint Research Centre
Institute for Health and Consumer Protection

Contact information
Address: Via E. Fermi 2749, 21027 Ispra (VA), Italy
E-mail: andrew.worth@ec.europa.eu
Tel.: +39 0332 789566
Fax: +39 0332 786717

http://ihcp.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC62586

EUR 24705 EN
ISBN 978-92-79-19081-0
ISSN 1018-5593
doi:10.2788/29048

Luxembourg: Publications Office of the European Union

Printed in Italy

ABSTRACT

In the regulatory assessment of chemicals, the use of *in silico* prediction methods such as (quantitative) structure-activity relationship models ([Q]SARs), is increasingly required or encouraged, in order to increase the efficiency and effectiveness of the risk assessment process, and to minimise the reliance on animal testing. The main question for the assessor concerns the usefulness of the prediction approach, which can be broken down into the practical applicability of the method and the adequacy of the predictions. A framework for assessing and documenting (Q)SAR models and their predictions has been established at the European and international levels. Exactly how the framework is applied in practice will depend on the provisions of the specific legislation and the context in which the non-testing data are being used. This report describes the current framework for documenting (Q)SAR models and their predictions, and discuses how it might be built upon to provide more detailed guidance on the use of (Q)SAR predictions in regulatory decision making. The proposed framework is illustrated by using selected pesticide active compounds as examples.

LIST OF ABBREVIATIONS

AD	Applicability domain
ADME	Absorption, Distribution, Metabolism and Elimination
ANN	Artificial Neural Network
CRD	UK Chemicals Regulations Directorate
DfW	Derek for Windows
ECHA	European Chemicals Agency
EFSA	European Food Safety Authority
EU	European Union
GCMP	Good Computer Modelling Practice
JRC	Joint Research Centre
k-NN	K Nearest Neighbours
LOAEL	Lowest Observed Adverse Effect Level
LOO	Leave-one-out
LMO	Leave-many-out
OECD	Organisation for Economic Cooperation and Development
PBBK	Physiologically Based Biokinetic (model)
PPP	Plant Protection Product
QMRF	QSAR Model Reporting Format
QPRF	QSAR Prediction Reporting Format
QSAR	Quantitative Structure-Activity Relationship
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
ROC	Receiver Operating Characteristic
SAR	Structure-Activity Relationship
SVM	Support Vector Machine (computational method)
TTC	Threshold of Toxicological Concern

CONTENTS

1. Introduction

In order to efficiently and effectively assess the risks of large numbers of existing chemicals and new chemical entities, there is an increasing emphasis in the regulatory setting on the use of so-called "non-testing" methods, either as a supplement to, or as a substitute for, traditional testing methods. In particular, alternatives to animal methods are being developed to reduce the need for animal testing in pharmacology and toxicology. Non-testing methods are based on the premise that the properties (including physicochemical properties and biological activities) of a chemical depend on its intrinsic nature and can be directly predicted from its molecular structure or inferred from the properties of similar compounds whose activities are known.

Non-testing methods include a range of predictive approaches, including Structure-Activity Relationships (SARs), Quantitative Structure Activity Relationships (QSARs), chemical grouping and read-across, and computer-based tools based on the use of one or more of these approaches. Non-testing methods also include models for predicting the absorption, distribution, metabolism and elimination (ADME) characteristics of chemicals in biological organisms (Mostrag-Szlichtyng & Worth, 2010), even though these models are not based entirely on the intrinsic properties of chemicals, as well their fate in the environment.

The main question for the assessor when applying non-testing methods for regulatory purposes concerns the usefulness of the approach, which can be broken down into the practical applicability of the method and the adequacy of the predictions. Considerable progress has been made at the European Union (EU) and international levels to develop a harmonised framework for assessing and documenting non-testing methods and their predictions. Exactly how this framework is applied in practice will depend on the provisions of the specific legislation (e.g. chemicals, pesticides, biocides, cosmetics) and the context in which the non-testing data are being used (including, for example, whether a traditional testing method is being replaced, whether additional, supporting data are available, and the consequences of making an inaccurate prediction). The general framework leaves largely open the difficult question of how to determine the adequacy of predicted data, and there is a considerable need to develop detailed guidance on how the predictions generated by non-testing methods can be translated into regulatory decisions.

This report introduces the conceptual basis of SARs and QSARs, collectively referred to as (Q)SARs. The current international framework for (Q)SAR models and predictions is then described. The practical applicability of this framework is illustrated by focussing on a checklist of 10 key questions, with respect to some well known software tools and their predictions of genotoxicity of two case study compounds. The purpose of these case studies is to highlight some of the scientific issues that need to be considered, as well the difficulties encountered. This report is based partly on work carried out by the Joint Research Centre (JRC) in the context of a study funded by the European Food Safety Authority (EFSA). The full report of this study is publicly available from the EFSA website (JRC, 2010).

1.1 Conceptual basis of (Q)SAR models

(Q)SARs are theoretical models that are designed to predict the physicochemical, biological (e.g. toxicological) and fate properties of molecules from knowledge of chemical structure.

More specifically, a SAR is a qualitative relationship between a molecular (sub)structure and the presence or absence of a given biological activity, or the capacity to modulate a biological activity imparted by another substructure. The term substructure refers to an atom, or group of adjacently connected atoms, in a molecule. A substructure associated with the presence of a biological activity is also called a structural alert. A SAR can also be based on the ensemble of steric and electronic features considered necessary to ensure the intermolecular interaction with a specific biological target

molecule, which results in the manifestation of a specific biological effect. In this case, the SAR is sometimes called a 3D SAR or pharmacophore.

A QSAR is a quantitative relationship (often a regression model) between a biological activity (which may be categorical or quantitative) and one or more molecular descriptors that describe chemical structure in numerical terms and which are used as predictors of the biological activity. A molecular descriptor is a structural or physicochemical property of a molecule, or part of a molecule. A comprehensive review of molecular descriptors has been published by Todeschini & Consonni (2009).

Guidance on the regulatory application of (Q)SARs was developed to support the implementation of the REACH legislation in Europe, and has been published by the European Chemicals Agency (ECHA, 2008).

In addition to the formalised approach of QSAR analysis, it is possible to estimate chemical properties and endpoints by using a less formalised approach, based on the grouping and comparison of chemicals. The grouping approach can be used, for example, to support the results of QSAR analysis or to generate estimated data (and fill data gaps) assuming that, in general, similar compounds will exhibit similar biological activity (ECHA, 2008).

1.2 The adequacy of (Q)SARs

The REACH guidance for applying (Q)SARs provides a flexible framework according to which it is possible to use data from (Q)SAR models instead of experimental data if each of four main conditions is fulfilled (ECHA, 2008):

- the model used is shown to be scientifically valid;
- the model used is applicable to the chemical of interest;
- the prediction (result) is relevant for the regulatory purpose; and
- appropriate documentation on the method and result is given.

Thus, multiple, overlapping conditions must be fulfilled to use a (Q)SAR prediction instead of data generated by a standard experimental test, as illustrated in Figure 1. The extent to which these conditions can be relaxed for the indirect and supporting use of (Q)SAR data, remains to be established on the basis of experience.

The following sections will explain the considerations necessary for demonstrating model validity, applicability and adequacy. The need to provide “appropriate documentation” is fulfilled by the provision of QSAR reporting formats for models and their predictions. The former type of documentation is the QSAR Model Reporting Format (QMRF) and the latter is the QSAR Prediction Reporting Format (QPRF). To accompany the more detailed guidance, ECHA has published a summary guide on how to report QSAR models. (ECHA, 2010).

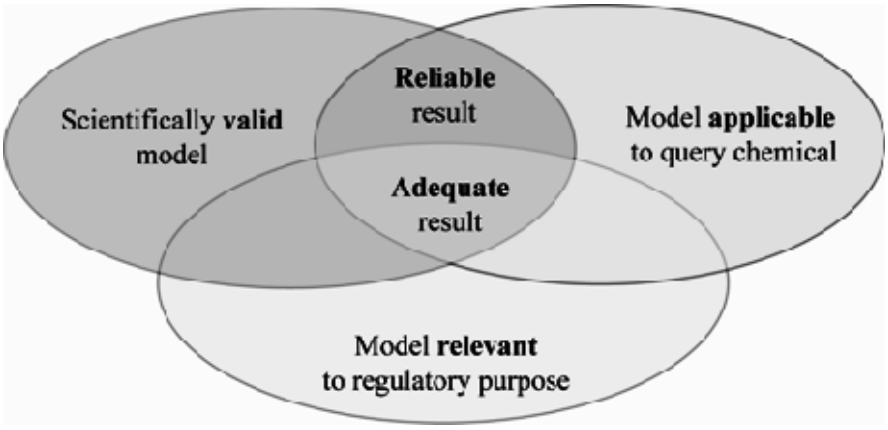


Figure 1 The overlapping considerations of validity, applicability and relevance needed to demonstrate (Q)SAR adequacy

1.3 Model validity

The first condition for using (Q)SARs is the demonstration of model validity. There is widespread agreement that models should be scientifically valid or validated if they are to be used in the regulatory assessment of chemicals. According to the OECD Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment, the term validation is defined as follows (OECD, 2005):

> "…the process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose"

This wide-ranging definition is intended to cover all kinds of traditional and alternative testing methods. In the context of (Q)SARs, this definition is rather abstract and difficult to apply. However, in the case of (Q)SARs, a set of five validation principles has been established by the OECD (OECD, 2007). The OECD principles for (Q)SAR validation state that in order:

> "to facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:
>
> 1. a defined endpoint;
> 2. an unambiguous algorithm;
> 3. a defined domain of applicability;
> 4. appropriate measures of goodness-of-fit, robustness and predictivity;
> 5. a mechanistic interpretation, if possible."

The OECD validation principles identify the types of information that are considered useful for the assessment of (Q)SARs for regulatory purposes. They constitute the basis of a conceptual framework, but they do not in themselves provide criteria for the regulatory acceptance of (Q)SARs. Fixed criteria are difficult, if not impossible, to define, given the highly context-dependent and variable ways in which non-testing data may be used. The intent of each principle is explained in Table 1.

The assessment of (Q)SAR model validity should therefore be performed by reference to the OECD principles for the validation of (Q)SARs. The validation exercise itself may be carried out by any person or organisation. The guidance on QSAR validation published by the OECD (OECD, 2007) is also summarised in the REACH guidance on the use of (Q)SARs (ECHA, 2008).

Based on a review of QSARs in the scientific literature, Dearden *et al.* (2009) identified as many as 21 types of common "error" which the authors associate with improper model development, assessment and documentation.[9] As can be seen in Table 2, some of these errors are related to the choice and treatment of the data that are included in the training and test sets (e.g. relevance, heterogeneity, spread in values, accuracy); other errors to the choice of descriptors (e.g. comprehensibility), and others to the statistical methodology (e.g. overfitting, suitable choice of statistics); and others to the overall transparency and transferability. This paper provides technical considerations that are useful for the model developer, some of which are also suitable for routine checking by the model end-user.

As an example, a common technical error in the modelling of toxicological endpoints is that QSAR models should be based on molar units, rather than on weight or concentration units. This is because physical and biological effects depend on the number of molecules present, and not on how much they weigh. However, this is often overlooked by the model developer because the underlying data, generated from guideline studies, are generally expressed in concentration or weight units. However, it is a simple exercise to convert values derived from dose-response curves (typically median values) from mg/kg to molar units, and this often improves the models derived. If, however, the response is measured at a fixed weight dosage for all chemicals, then conversion is useless, because each chemical will have been tested at a different molar dose, making comparison impossible. This one example illustrates the detail that is ideally needed in a comprehensive assessment of QSAR models. In developing the QSAR reporting formats, the main challenge was to define a level of detail in reporting models and their predictions that represents a workable compromise between what is desirable scientifically and what is workable and sufficient to support the regulatory assessment of models.

Information on (Q)SAR model validity, including peer-reviewed documentation , is available the JRC QSAR Model Database (http://qsardb.jrc.it). This database is intended to be a repository of potentially useful information on models that are characterised according to the internationally accepted format of the QMRF. It is not intended to be an inventory of officially accepted or adopted (Q)SAR models, since in the EU there is no formal process for their validation and acceptance (Worth, 2010).

1.4 Model overfitting - causes, consequences and diagnostics

An important consideration when assessing the validity of statistically-based models concerns the problem of overfitting. This is covered by OECD Validation Principle 4. The goodness-of-fit of a QSAR model reflects: a) the ability of the model predictors to account for the variance of the response in the training set; and b) the statistical significance of the model. The optimal model should express a balance between the complexity and relevance of the applied predictors/methodology and the resulting benefits in terms of performance. According to the principle of parsimony (Occam's Razor), the optimal model should be based on the minimum necessary information and nothing more. Otherwise, it can result in one of two extremes: either the model is over-fitted, i.e. too complex and simply modelling noise, or the model is under-fitted, i.e. too simplistic and lacking vital information (Hawkins, 2003, Gramatica, 2007). Moreover, if the model is not statistically significant, it should not be used for predictive purposes. These considerations do not apply to models that are entirely knowledge-based rather than statistically-based.

There are two main causes of overfitting, resulting in a redundant level of model complexity, which does not improve (and may even diminish) the model performance. The first main cause is the improper selection of independent variables by: a) including more predictors than are necessary to capture the variance of the response (Aptula *et al.*, 2005); b) using predictors that are inter-correlated (collinear); c) using predictors that are irrelevant and correlated with the response "by chance", without being meaningful and predictive (Topliss & Edwards, 1979). Another main cause of overfitting is related to the choice of modelling technique that is: a) more complicated than necessary to find the relationships between the descriptors and the response; or b) not suitable to describe particular dependencies.

Overfitting is undesirable for a variety of reasons. Using superfluous and/or irrelevant predictors can result in the loss of valuable information, adding random variation to the predictions (i.e. prediction error) and incomplete or misleading interpretation of the modelled phenomenon. The use of collinear predictors means that the same information is being used more than once to fit the model and this leads to an overestimation of its importance. Overfitting usually results in unpredictable errors, which are related to the noise in the dataset (due to the selected variables and methods), rather the modelled property (Tetko *et al.*, 2008). The problem of overfitting is more likely to occur, and harder for the non-specialist to diagnose, in the case of methods capable of handling large amounts of correlated information and noisy variables. Examples of such methods are Support Vector Machines (SVM) and Artificial Neural Networks (ANN).

To assess the goodness-of-fit of a model, the predictions for chemicals in the training set are used. However, models that are overfitted typically show worse predictivity for independent data than their internal validation statistics imply. Thus, internal validation should ideally be followed by external validation. The main indication of an overfitted model is its high internal "predictivity" (error of estimation much lower than the experimental error), accompanied by a lower (or lack of) external predictivity.

To avoid overfitting, it is necessary to apply certain statistical validation techniques providing fitness functions that allow verification of the appropriate form (relevance) of the model (in terms of the variables and methodology used) and which enable the optimal model complexity to be found. Clearly, statistical expertise is required to apply and interpret the results of such techniques.

The initial splitting of data into training and test sets is of vital importance, since it determines the data which will be used to fit the model. Ideally, both the training and test sets should cover the entire range of the considered chemical space, defined by the predictor values, on the one hand, and by the endpoint values, on the other (Daszykowski *et al.*, 2002). The most popular algorithms used to design the optimal size and composition of training and test sets are: Kennard-Stone (Kennard & Stone, 1969), Duplex (Snee, 1977), D-optimal distance (Cook & Nachtsheim, 1980), repeated test set technique (Boggia *et al.*, 1997) and self-organising map (Kohonen, 1998). However, a practical problem is that high-quality data are not always sufficiently available to divide into independent training and test sets, which means that most if not all of the available data are used to train the model.

1.4.1 Regression models

In order to assess the predictive ability of a regression model in the absence of an external test set, a number of internal validation techniques can be used (Cramer e*t al.*, 1988). The most popular techniques are based on leave-one-out (LOO) and leave-many-out (LMO) cross-validation methods (Osten, 1988), as well as bootstrap resampling (Wehrens *et al.*, 2000). Although the LOO cross-validation is probably the most widely used technique, it often provides an overly optimistic simulation of a model's predictive ability, especially when large datasets are used. In such cases, more realistic (lower) estimates of predictivity can be obtained by LMO cross-validation, which introduces a larger perturbation in the studied dataset. Similarly, the fitness statistics provided by the bootstrapping method are indicative of the model's predictivity. Models which are based on "chance correlation" can be identified by the response permutation test (Y-scrambling; Lindgren *et al.*, 1996) or the QUIK rule (Todeschini *et al.*, 1999).

The statistical significance of a model can be estimated by its *F*-values for a given number of degrees of freedom. The probability of the model equation being significant increases with the *F*-value. A statistically significant model ensures that the highest possible predictivity is obtained with the minimal number of predictors (Occam's razor) and that the predictors are orthogonal (uncorrelated). A model having low significance statistics may be underfitted, based on "chance correlations", noisy predictors or intercorrelated predictors, and will probably show discrepancies between its goodness-of-fit and its predictivity. In order to assess the statistical significance of each regression coefficient, a t-test (for a given number of degrees of freedom) should be performed. This allows identification of

those predictors that significantly contribute to the explanation of the response variable and those that are meaningless. The main fitness functions performed for various types of models and possible ways of interpreting them are summarised in Table 3.

1.4.2 Classification models

The goodness-of-fit of a classification model can be assessed in terms of its Cooper statistics, namely: sensitivity, specificity, concordance (accuracy), positive/negative predictivity, as well as false positive/false negative rates. For individual classification models, the relevant Cooper statistics should be significantly greater than a pre-defined threshold (typically 50%, but a lower percentage may be acceptable if the emphasis is on positive/negative predictivity rather than sensitivity/specificity). In order to compare the performances of a number of classification models, the Receiver Operating Characteristic (ROC) curve, plotting the true positive rate against the false positives rate, is often used. In the ROC curve, a model in the top left-hand corner is ideal (fully predictive), whereas a model along the diagonal is producing predictions nor better than chance.

Methods for checking whether a model is overfitted can be automatically applied by available statistical software packages. Nevertheless, the user will need sufficient knowledge about the diagnostic rules and the underlying methods in order to interpret the statistical. Different statistical approaches may provide different statistics; for example, different values of the same statistics can result from varying the composition of the training and test sets. Thus, transparency is necessary to document the procedure of model development and validation process. However, exactly how much transparency is needed for regulatory purposes is still a matter of debate.

Apart from the statistical approaches to assess whether the model is overfitted or not, which require statistical expertise to apply and interpret, some more simple and straightforward “rules of thumb” can be followed based on the information available to the assessor:

1. to estimate the model's goodness-of-fit, internal predictivity and statistical significance, the following statistics should be provided: n, r^2 (R^2), q^2 (Q^2), R^2_{adj}, s, F statistics including p-values;
2. the number of chemicals (n) to predictors should be at least 5:1 (Topliss and Costello, 1972);
3. for transparent mechanistic interpretation of the model, the maximal number of predictors (except group contributions and electropological state indices) should not exceed 5-6 (Dearden *et al*., 2009);
4. the standard error of the estimate (s) should not be significantly less than the known experimental error for the predicted endpoint;
5. the difference between the $R^2(Y)$ and $Q^2(Y)$ values should not exceed 0.3.

1.5 Model applicability

Assessment of model validity is a necessary but not sufficient step in assessing the adequacy of a (Q)SAR prediction. Assuming that the model of choice is considered valid, a second essential step is to demonstrate the applicability of the model to the chemical of interest. The evaluation of model applicability is related to the evaluation of the reliability of prediction for the chemical of interest, since a valid (Q)SAR is associated with at least one defined applicability domain in which the model makes estimations with a defined level of accuracy (reliability). When applied to chemicals within its applicability domain, the model is expected to give reliable results (or results with a defined level of reliability). Conversely, if a model is applied to a chemical outside its applicability domain, it is likely that the estimated result will either be unreliable or of unknown reliability.

The applicability domain of a model is a multi-faceted concept, and can be broken down into: a) a descriptor domain; b) a structural fragment domain; c) a mechanistic domain; and d) a metabolic

domain. In other words, the reliability of a prediction is constrained by whether the chemical of interest has: a) descriptor values within predefined ranges; b) structural fragments that are "known" to the model; c) its predefined mode and/or mechanism of action; and d) the likelihood that it may undergo transformation or metabolism, and the characteristics of any products.

There is no unique measure of model reliability, and no criteria for (Q)SAR reliability have been established in regulatory guidance. Model reliability should be regarded as a relative concept, depending on the context in which the model is applied. In other words, a greater or lesser degree of reliability may be sufficient for a given regulatory application. This implies that the applicability domain can be defined to suit the regulatory context.

The assessment of whether a given model is applicable to a given chemical can be broken down into the following specific questions:

1. is the chemical of interest within the scope of the model, according to the defined applicability domain of the model?
2. is the defined applicability domain suitable for the regulatory purpose?
3. how well does the model predict chemicals that are similar to the chemical of interest?
4. is the model estimate reasonable, taking into account other information?

The importance of having an explicit definition of the model domain becomes apparent when addressing question 1. In practice, there is often limited information concerning the descriptor, structural fragment, mechanistic, and metabolic domains.

The second question arises because most currently available models were not tailor-made for current regulatory needs and inevitably incorporate biases which may or may not be useful, depending on the context of prediction. A model can be (deliberately or inadvertently) biased toward certain classes of chemicals), or toward a certain type of prediction (e.g. a model optimised to correctly identify positives at the expense of correctly identifying negatives). Such biases do not affect the validity of the model, but they do affect its applicability for specific purposes. Information on these biases can therefore help the user determine whether the model is suitable.

The third question provides a simple way of checking whether a model is appropriate by checking its predictive capability for one or more analogues that are similar to the one of interest and for which measured values exist. This is effectively using a read-across argument to support the reliability of the (Q)SAR prediction.

A more generic check, expressed by question 4, is whether the predicted value seems "reasonable", based on any other information available. This is an appeal to an expert judgement, supported with argumentation.

The judicious application of these questions to assess the applicability of a (Q)SAR model is by no means straightforward, and needs specialised expertise. Software applications that generate (Q)SAR estimates vary in the extent and manner to which they incorporate and report applicability domain considerations.

1.6 Model adequacy

The preceding two sections explain that in order for a (Q)SAR result to be adequate for a given regulatory purpose, the estimate should be generated by a valid model, and the model should be applicable to the chemical of interest with the necessary level of reliability. Fulfilment of these two conditions is necessary but not sufficient for demonstrating adequacy. At present, there is no detailed and firm guidance on how to demonstrate adequacy, but some general considerations are offered in the REACH guidance (ECHA, 2008). This is partly a reflection of the fact that more experience is needed at the regulatory level to expand on existing guidance, but also that the concept of adequacy, by its very nature, means that only general considerations will be possible. In any case, to demonstrate the

adequacy of a QSAR estimate generated by a valid and applicable model some additional argumentation is required.

One piece of argumentation is that the model endpoint should be relevant for the regulatory purpose. For some models, in which the model predicts directly the regulatory endpoint (e.g. an acute toxicity LD50 value), the relevance is self-evident. However, in the case of many QSAR models, and especially a new generation of QSAR models that are focusing on predicting lower-level mechanistic endpoints (Cronin *et al.*, 2009), an additional extrapolation is needed to relate the modelled endpoint (e.g. nucleophilic reactivity towards DNA or proteins) to the endpoint of regulatory interest (e.g. mutagenicity or sensitisation).

The relevance and reliability of a given prediction need to be assessed in relation to a particular regulatory purpose, taking into account the availability of other information in the context of a weight-of-evidence assessment. In other words, the question is whether the totality of information is sufficient to reach a regulatory conclusion, and if not, what additional information (possibly including new test data) is needed to reduce the uncertainty and increase confidence in the conclusion. This should take account of the "severity" of the decision (the "principle of proportionality") as well as the possible consequences of reaching a "wrong" conclusion ("principle of caution or conservativeness"). Thus, the amount and quality of information that is required depends on the uncertainty in the data, the severity of the regulatory decision, and the consequence of being wrong. It follows that the determination of adequacy is based not only on scientific argumentation, but also on a policy decision. For this reason, it does not make sense to develop absolute criteria for assessing adequacy that are acceptable in all regulatory decision-making contexts.

2. Proposed framework for assessing *in silico* predictions

Although absolute criteria cannot be meaningfully formulated for assessing the adequacy of (Q)SAR predictions for regulatory purposes, it is possible to highlight a minimal set of important questions that can applied in a practical manner by the risk assessor. For illustrative purposes, a checklist of 10 questions is proposed in Table 4. Questions 1-6 are related to the model, whereas questions 7-10 are related to the model prediction. This set of questions is not intended to be definitive - individual questions could be skipped or additional questions could be added, depending on the needs of a particular regulatory framework and the context in which the predictions are being used.

2.1 Method for applying the framework

To illustrate the application of the checklist, the 10 questions have been applied to the predictions of genotoxicity for two case study chemicals, methyl parathion and sodium nitroguaiacolate, obtained with a range of popular software tools, including:

- a tool based on expert rules, Derek for Windows;
- tools based on statistical methodologies: CAESAR, Lazar, TOPKAT, HazardExpert, and the formerly named ToxBoxes (now called ACDToxSuite);
- and a hybrid tool (Toxtree).

These tools were selected for illustrative purposes. Many other software tools also make predictions of genotoxiicty, and there is a vast literature of published models, as reviewed elsewhere (Serafimova *et al.*, 2010).

2.1.1 Derek for Windows

Derek for Windows (DfW) is a SAR-based system is developed by Lhasa Ltd, a non-profit company and educational charity (https://www.lhasalimited.org/). DfW contains over 50 alerts covering a wide range of toxicological endpoints in humans, other mammals and bacteria. An alert consists of a toxicophore (a substructure known or thought to be responsible for the toxicity) and is associated with literature references, comments and examples. A key feature of DfW is the transparent reporting of the reasoning underlying each prediction.

All the rules in DfW are based either on hypotheses relating to mechanisms of action of a chemical class or on observed empirical relationships (Sanderson & Earnshaw, 1991). Information used in the development of rules includes published data and suggestions from toxicological experts in industry, regulatory bodies and academia. The toxicity predictions are the result of two processes. The program first checks whether any alerts in the knowledge base match toxicophores in the query structure. The reasoning engine then assesses the likelihood of a structure being toxic. There are nine levels of confidence: certain, probable, plausible, equivocal, doubted, improbably, impossible, open, and contradicted. DfW can be integrated with Lhasa's Meteor software, which makes predictions of fate, thereby providing predictions of toxicity for both parent compounds and their metabolites.

DfW predictions are knowledge-based, based on the application of alerts and reasoning rules. The final toxicity assessment is a result of a two-part process: (i) the program checks whether any alerts from the knowledge base appear in the query compounds, and (ii) the reasoning model is applied in order to determine the likelihood of the compound's toxicity (expressed as the level of likelihood). If no alerts from the knowledge base can be matched against query structure, the program displays a message "Nothing to report".

Genotoxicity alerts in DfW include alerts for mutagenicity (in bacteria and mammals) and alerts for chromosome damage based on the in vitro chromosomal aberration assay and including effects that do

not involve direct DNA damage (inhibition of DNA synthesis/repair, spindle function disruption, reactive oxygen species generation, energy depletion, thiol reactivity, intercalation).

In order to make the results from DfW comparable with other results, we converted the output into three categories: active, equivocal and not active, as in the following table.

Interpretation of Derek toxicity predictions

Level of likelihood	Interpretation of the results
Certain	active
Probable	active
Plausible	active
Equivocal	equivocal
Doubted	not active
Improbable	not active
Impossible	not active
Open	not active
Contradicted	not active
Nothing to report	not active

2.1.2 CAESAR

CAESAR comprises a series of statistically-based models developed within EU-funded CAESAR project (http://www.caesar-project.eu). The models have been implemented into open-source software and made available for online use via the web. Predictions can be made for five endpoints: mutagenicity (Ames), carcinogenicity, developmental toxicity, skin sensitisation, and the bioconcentration factor.

The CAESAR prediction of mutagenicity is based on the SVM approach and the Kazius/Bursi database (http://www.cheminformatics.org/datasets/bursi). The SVM modelling is followed by an “expert facility” filter based on Benigni/Bossa rules, applied to the compounds presumed safe by SVM. The filter combines two sets of structural alerts with different distinguishing features: the former (the "sharp" one) has the aim to enhance the prediction accuracy attempting a precise identification of misclassified False Negatives (FN), the latter (the “suspicious” one) continues with the FN removal in such a way that this does not noticeably reduce the original prediction accuracy by generating too many False Positives (FP) as well. Compounds picked out by the first checkpoint are classified as “mutagenic” (i.e. active), and those picked out by the second are classified as “suspicious” (i.e. equivocal). Unaffected ones are finally classified as “non-mutagenic” (i.e. inactive).

2.1.3 ToxBoxes

ToxBoxes (now called ACD/Tox Suite), marketed by ACD/Labs and Pharma Algorithms, provides predictions of various toxicity endpoints including human *Ether-à-go-go* Related Gene (hERG) channel inhibition, genotoxicity, cytochrome P450 (CYP3A4) inhibition, Estrogen Receptor (ER) binding affinity, irritation, rodent acute lethal toxicity (LD50), aquatic toxicity, and organ-specific health effects (http://www.acdlabs.com/products/admet/tox/). The predictions are associated with confidence intervals and probabilities, thereby providing a numerical expression of prediction reliability. The software incorporates the ability to identify and visualize specific structural toxicophores, giving insight as to which parts of the molecule are responsible for the toxic effect. It also identifies analogues from its training set, which can also increase confidence in the prediction. The algorithms and datasets are not disclosed.

The predictions of genotoxicity by ToxBoxes are based on the probability of query compounds to be genotoxic in Ames test. The training data used in the software originate from Chemical Carcinogenesis Research Information (CCRIS) and Genetic Toxicology Data Bank (GENE-TOX), containing the results of Ames genotoxicity assays for several strains of *S. typhimurium* (TA97, TA98, TA100, TA102, TA104, TA1535, TA1537, TA1538 and also E. coli strain WP2 uvrA), with or without metabolic activation. In establishing this training set, a compound was considered genotoxic if at least one of Ames results was positive; otherwise, the compound was considered non-genotoxic. In case of inconsistent results from different assays, the data were evaluated by experts and in some cases had been labelled as inconclusive. The final training set exceeded 8000 compounds with standardised Ames genotoxicity values. A neural network model was built using structural fragments as descriptors. Molecules were decomposed into atomic and chain-based fragments (chains of interconnected atoms). Fragments containing 2 to 5 atoms, present in at least 10 training set molecules were used to develop the model. The model makes a prediction if the chemical structure is more than 75% covered by fragments in the training set. For each compound, the "probability of positive Ames test" and a so-called "Ames test reliability index" are provided.

The method suggested by the vendor was adopted to convert the probability values into binomial ones (actives or inactives) according to the following rules:

(i) if the "Probability of positive Ames test" is bigger than 0.7, then the compound is a predicted mutagen (i.e. active);

(ii) if the "Probability of positive Ames test" is smaller than 0.3, then the compound is a predicted non-mutagen (i.e. inactive);

(iii) if the "Probability of positive Ames test" is between 0.7 and 0.3, then the result is predicted as equivocal.

2.1.4 Lazar

Lazar is an open-source software programme that makes predictions of toxicological endpoints (currently, mutagenicity, human liver toxicity, rodent and hamster carcinogenicity, and Maximum Recommended Daily Dose) by analysing structural fragments in a training set (Helma, 2006; Maunz & Helma, 2008). It is based on the use of statistical algorithms for classification (k-nearest neighbours and kernel models) and regression (multi-linear regression and kernel models). In contrast to traditional k-Nearest Neighbour (k-NN) techniques, Lazar treats chemical similarities not in absolute values, but as toxicity dependent values, thereby capturing only those fragments that are relevant for the toxic endpoint under investigation. The Lazar algorithm works by building an instance-based local model that excludes the chemical being predicted from its local training set. Lazar performs automatic applicability domain estimation and provides a confidence index for each prediction, and is usable without expert knowledge. Lazar runs under Linux and a web-based prototype is also freely accessible (http://lazar.in-silico.de/).

The mutagenicity predictions by Lazar are based on a k-NN algorithm and two datasets: Kazius/Bursi (http://www.cheminformatics.org/datasets/bursi/) and the so-called "Benchmark Data Set for In Silico Prediction of Ames Mutagenicity" (http://ml.cs.tu-berlin.de/toxbenchmark/). Each prediction is associated with a prediction confidence (between 0 and 1), which gives information about the presence/absence of studied compounds within the applicability domain (AD) of the model. The developer proposed a confidence value higher than 0.025 as a reasonable hard cut-off for compounds within the AD. The accuracy of prediction decreases with the confidence value.

2.1.5 TOPKAT

TOPKAT is a QSAR-based system, developed by Accelrys Inc. (http://accelrys.com/), makes predictions of a range of toxicological endpoints, including mutagenicity, developmental toxicity, rodent carcinogenicity, rat chronic LOAEL Lowest Observed Adverse Effect Level (LOAEL), rat Maximum Tolerated Dose (MTD) and rat oral LD_{50}. The QSARs are developed by regression analysis for continuous endpoints and by discriminant analysis for categorical endpoints. TOPKAT models are derived by using a range of two-dimensional molecular, electronic and spatial descriptors. TOPKAT estimates the confidence in the prediction by applying the patented Optimal Predictive Space (OPS) validation method. The OPS is TOPKAT's formulation of the model applicability domain - a unique multivariate descriptor space in which a given model is considered to be applicable. Any prediction generated for a query structure outside of the OPS space is considered unreliable.

The TOPKAT mutagenicity model was developed from compounds assayed according to the US EPA GeneTox protocol (i.e. tested against five strains of *Salmonella typhimurium* using the Histidine Reversion Assay). A chemical is labelled a mutagen if a positive response is observed against one or more strains. A chemical is considered a non-mutagen if a negative response is observed in all of these five bacterial strains. Therefore, when a query structure is assessed by TOPKAT to be a non-mutagen (computed probability of mutagenicity between 0.0 and 0.3), it indicates that there is a high probability of the query chemical producing a negative response in the Histidine Reversion Assay against all of the five bacterial strains. It is important to note that a non-mutagen assessment by TOPKAT does not mean that the query chemical will be a non-mutagen in other mutagenicity tests, such as the micronucleus and Chinese Hamster Ovary tests. As suggested by the vendor, probability values can be converted into binomial ones (actives or inactives) according to the following rules:

(i) if computed probability of mutagenicity greater than 0.7, then the compound is considered to be a mutagen (i.e. active);

(ii) if computed probability of mutagenicity smaller than 0.3, then the compound is considered to be a non-mutagen (i.e. inactive);

(iii) if computed probability of mutagenicity between 0.3 and 0.7, then the prediction is equivocal.

2.1.6 HazardExpert

HazardExpert is a module of the Pallas software developed by CompuDrug (http://compudrug.com/). It predicts the toxicity of organic compounds based on toxic fragments, and it also calculates bioavailability parameters (logP and pKa). It is a rule-based system with an open knowledge base, allowing the user to expand or modify the data on which the toxicity estimation relies. It covers the following endpoints relevant to toxicity assessment: carcinogenicity, mutagenicity, teratogenicity, membrane irritation, immunotoxicity and neurotoxicity

The results of mutagenicity predictions by HazardExpert (Pallas v 3.3.2.4) are provided as relative percentage toxicity values. On the basis of the ranges of the results the authors proposed the classification of chemicals as "highly probable", "probable", "uncertain" and "not probable" to express mutagenic activity. In order to compare the HazardExpert predictions with the results of other software tools we treated "highly probable" and "probable" chemicals as active**,** "uncertain" chemicals as equivocal**,** and "not probable" ones as not active, as in the following table.

Interpretation of HazardExpert mutagenicity predictions

The range of relative percentage toxicity [%]	Toxic Class	Classification	Interpretation of the results
100-60	1	Highly probable	active
59-48	2A	Probable	active
47-36	2B	Probable	active
35-3	3	Uncertain	equivocal
2-0	4	Not probable	not active

2.1.7 Toxtree

Toxtree is a flexible and user-friendly open-source application that places chemicals into categories and predicts various kinds of toxic effect by applying decision tree approaches. Toxtree can be downloaded from the JRC (http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/index.php?c=TOXTREE) and from Sourceforge (https://sourceforge.net/projects/toxtree/). Toxtree has been developed by the JRC in collaboration with various consultants, in particular Ideaconsult Ltd (Sofia, Bulgaria). A key feature of Toxtree is the transparent reporting of the reasoning underlying each prediction.

In this study, Toxtree v 1.60 was used. The mutagenicity predictions generated by Toxtree v. 1.60 are based on a decision tree implementing the Benigni/Bossa rules (Benigni et al., 2008) and rules for the *in vivo* micronucleus assay (Benigni et al, 2010). In addition, Toxtree applies the following QSAR models to query chemicals belonging to the classes of aromatic amines or alpha,beta-unsaturated aldehydes: (i) QSAR6 - mutagenic activity of aromatic amines in the *Salmonella typhimurium* TA100 strain (Ames test); (ii) QSAR8 - carcinogenic activity of the aromatic amines in rodents (summary activity from rats and mice); (iii) QSAR 13 - mutagenic activity of alpha,beta-unsaturated aldehydes in the *Salmonella typhimurium* TA100 strain (Ames test). There are certain exceptions in the application of these QSARs, namely QSAR6 and QSAR8 in Toxtree v 1.60 apply to aromatic amines with the exclusion of aromatic amines having a sulphonic group on the same ring, and QSAR13 applies to alpha,beta-unsaturated aldehydes excluding cyclic alpha,beta-unsaturated aldehydes.

The structural rules in Toxtree are based largely on expert knowledge rather than statistically derived from training sets. However, the Benigni-Bossa rulebase includes some QSARs in addition to the structure-based rules: QSAR6 (Ames mutagenicity of aromatic amines) has 111 chemicals in its training set, QSAR8 (rodent carcinogenicity of aromatic amines) has 64 training set chemicals, and QSAR13 (Ames mutagenicity of alpha,beta-unsaturated aldehydes) has 20.

2.2 Results obtained in the application of the framework

The results are given in Tables 5-11. Many of the questions are straightforward to answer, provided that the background documentation is available (QMRF, software manual and/or research publications). However, a few noteworthy points are elaborated in the following paragraphs:

The predicted endpoint is not always defined in detail; for example, sometimes a generic prediction of genotoxic potential is made but the underlying mechanistic effect may not be clearly identified (e.g. Ames mutagenicity, chromosome aberration). In the case of expert systems, the predictions are likely to be based on heterogeneous datasets of expert conclusions based on data from multiple test methods: the conclusions for genotoxic potential are likely to vary between assessors and over time, especially when the criteria for assessing the raw data have changed or were not clear in the first place. In the case of statistical models, this is a source of variability in the training set, which will inevitably affect the reliability of prediction.

The question of model overfitting is not applicable to all types of models, especially knowledge-based models which encode human knowledge and statistical models that employ instance-based (nearest neighbour) learning algorithms. However, in general this consideration is applicable to statistically based models.

Comparing the prediction for the chemical of interest with predictions made for similar chemicals helps to assign confidence to the prediction, even when information on the statistical characteristics or the mechanistic basis of the model are missing. Some software tools (e.g. CAESAR, ToxBoxes) provide information on analogues. In such cases, it is recommended to consider the analogues provided by the software since these are likely to reflect the applicability domain of the model. However, additional and/or different analogues might also be considered. Other software tools (e.g. Toxtree) do not provide any information on analogues, so in these cases, the user needs to use other resources to find analogues. It is useful to know whether the selected analogues are in the training set of the model. This does not mean that they are not appropriate choices. On the contrary, if the model training set contains analogues of the chemical of interest, this increases confidence that the model is applicable since it covers the same chemical space.

At present, there is no firm guidance on how to select an appropriate set of analogues - it is left largely to expert judgement. However, a few rules-of-thumb can be proposed:

a) at least two analogues should be selected, including both known positives and negatives;

b) in general, 2-5 analogues should be sufficient for the evaluation;

c) the experimental data for the analogues selected should be reliable and appropriate (e.g. same effect/endpoint) for the comparison;

Computer-based search tools can be used to find analogues. Freely available tools include PubChem (http://pubchem.ncbi.nlm.nih.gov/), ChemSpider (http://www.chemspider.com), AMBIT (http://ambit.sourceforge.net/intro.html) and AIM (http://www.epa.gov/oppt/sf/tools/aim.htm). These tools do not always provide links to relevant and reliable experimental data. Freely downloadable tools such as the OECD QSAR Toolbox (http://www.qsartoolbox.org/) and Toxmatch (http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/index.php?c=TOXMATCH) can also be used (provided that a suitable dataset is already included or imported in the software). While these tools can assist the user in finding analogues, it is still necessary to judge whether the analogues are appropriate (or which analogues are most appropriate). Worked examples on how to use AMBIT and Toxmatch are given in Jeliazkova *et al.* (2010).

2.3 Use of analogue data in the assessment of predictions

An illustration of the use of analogue data is provided by predictions for methyl parathion and sodium nitroguaiacolate. Methyl parathion is a well-predicted chemical, in the sense that all software tools correctly predict it to be a mutagen (or indeterminate in the case of ToxBoxes). Conversely, sodium nitroguaiacolate is a poorly-predicted chemical, in the sense that all software tools incorrectly predict it to be a mutagen (or indeterminate in the case of ToxBoxes). In other words, methyl parathion is usually a true positive in mutagenicity prediction, whereas sodium nitroguaiacolate is usually a false positive.

The main question considered here is whether a judicious choice of analogues and their associated mutagenicity data would lead an assessor to trust the prediction for the true positive, but be doubtful of the prediction for the false positive. A selection of analogues for methyl parathion and sodium nitroguaiacolate is given in Tables 12 and 13, respectively. These analogues were chosen merely to illustrate the process, and are not necessarily the most appropriate analogues for the assessment of the two pesticides.

For methyl parathion, four analogues were identified (Table 12), including a non-mutagen (fenitrothion) and three mutagens (fenitrooxon, hydroxymethylfenitrothion 4-aminofenitrothion).

For nitroguaiacolate, five analogues were identified (Table 13), including two mutagens (o-nitroanisole, p-nitroanisole) and three non-mutagens (m-nitroanisole, 3-methyl-4-nitrophenol, m-nitrophenetole).

2.3.1 Derek for Windows predictions

In the case of methyl parathion, DfW generates correct predictions for the three mutagenic analogues, but incorrectly predicts the non-mutagenic analogue to be a mutagen. Thus, the predicted mutagenicity of methyl parathion might be considered reliable on the basis of weight-of-evidence (3 out of 4 correct), and this would lead to the right conclusion. In the case of nitroguaiacolate, the predictions for the two mutagenic analogues were correct whereas the predictions for two of the three non-mutagens were incorrect (unlike Toxtree, Derek correctly predicts the non-mutagenicity of 3-methyl-4-nitrophenol). On the basis of marginal weight-of-evidence (3 out of 5 correct), the predicted mutagenicity of nitroguaiacolate might be considered reliable, but this would lead to the wrong conclusion.

2.3.2 CAESAR predictions

In the case of methyl parathion, CAESAR generates correct predictions for the three mutagens, and also correctly identifies the non-mutagen. Thus, the predicted mutagenicity of methyl parathion would probably be considered reliable (all four predictions correct). In the case of nitroguaiacolate, the predictions for the two mutagens were correct but the predictions for the three non-mutagens were incorrect (including 1-ethoxy-3-nitrobenzene which is in the training set). Thus the prediction for nitroguaiacolate would probably be considered unreliable on the basis of weight-of-evidence (3 out of 5 incorrect) and the fact the model does not correctly predict a known non-mutagen in its training set. In other words, this interpretation of CAESAR predictions would lead to the right conclusions for both chemicals.

To illustrate the use of analogue data in assessing the reliability of prediction, the argumentation provided above is based entirely on the weight-of-evidence of reliable predictions. This led to mixed results – sometimes the correct conclusion was drawn, and sometimes the incorrect one, but overall the conclusions were the right ones.

2.3.3 ToxBoxes predictions

The application of ToxBoxes is interesting since it appears to perform very well on the basis of a global statistical analysis - positive and negative predictivities of 93% (Worth *et al.*, 2010). However, the apparently high predictivity of the model could be due to its large training set, which contains an unknown overlap with any test set. When ToxBoxes is applied to methyl parathion and nitroguaiacolate, it gives indeterminate predictions, which means that further information would be sought. The generation of indeterminate predictions is not necessarily a weakness in a model. On the contrary, it could be seen as a strong point, since it would avoid reliance on an incorrect prediction.

2.3.4 Lazar predictions

Methyl parathion is in the Lazar training set, so in this case, the assessment based on experimental data would be used. In the case of nitroguaiacolate, the predictions for the three non-mutagens were incorrect, whereas the predictions the two mutagens were correct. Again, the prediction for nitroguaiacolate might be doubted purely on the basis of weight-of-evidence.

2.3.5 TOPKAT predictions

In the case of methyl parathion, TOPKAT generates correct predictions for two of the three mutagenic analogues, but generates a false positive prediction for the non-mutagen. On this basis, the predicted mutagenicity of methyl parathion would probably be considered unreliable (2 out of 4 predictions correct), even though this would be the wrong conclusion. In the case of nitroguaiacolate, one of the two mutagens was correctly identified, and two of the three non-mutagens were correctly identified. Thus the prediction for nitroguiacolate might be considered reliable on the basis of weight-of-evidence (3 out of 5 correct) although this would lead to the wrong conclusion. In other words, this interpretation of TOPKAT predictions would lead to the wrong conclusions for both pesticides.

2.3.6 HazardExpert predictions

In the case of methyl parathion, HazardExpert generates correct predictions for the three mutagen analogues, but incorrectly predicts the non-mutagen analogue to be a mutagen. Thus, the predicted mutagenicity of methyl parathion might be considered reliable on the basis of weight-of-evidence (3 out of 4 correct). In the case of nitroguaiacolate, the predictions of the two mutagenic analogues were correct but the predictions for the three non-mutagenic analogues were incorrect. Thus, the prediction for nitroguaiacolate might be considered unreliable on the basis of weight-of-evidence (3 out of 5 incorrect). This use of HazardExpert would therefore lead to the right conclusions for the methyl parathion and nitroguaiacolate.

2.3.7 Toxtree predictions

In the case of methyl parathion, comparison of the known and predicted mutagenicities reveals that Toxtree makes the correct predictions for the three mutagens, but incorrectly predicts the non-mutagen to be a mutagen. On the basis of a weight-of-evidence (3 out of 4 predictions correct), this might be considered sufficient supporting evidence to rely on the mutagenicity prediction for methyl parathion. The false positive prediction for fenitrothion, as opposed to fenitrooxon for example, might be related to the fact that the latter is the phosphoryl derivative (bearing the P=O functionality) of the former (bearing the P=S functionality). It is known that phosphoryl derivatives are more toxic than the corresponding thiophosphoryl compounds because they bind more strongly to (and act as more potent inhibitors of) certain enzymes, such as acetyl cholinesterase. However, the Toxtree rulebase does not capture this difference in chemistry.

In the case of nitroguaiacolate, the Toxtree predictions for the three non-mutagens were incorrect, whereas the predictions the two mutagens were correct. Thus, in this case, the prediction for nitroguaiacolate might be doubted purely on the basis of weight-of-evidence (3 out of 5 predictions incorrect), and this would be the right conclusion.

Overall, comparison of the Toxtree predictions with the data for the above-mentioned analogues indicates that Toxtree tends to overpredict mutagenicity, which is consistent with the global statistics - a positive predictivity of 73% and a negative predictivity of 82% (Worth *et al.*, 2010). This means that in general greater confidence should be assigned to a negative prediction than a positive prediction. In terms of the underlying (Benigni-Bossa) rulebase, this indicates that the structural alerts act as a coarse-grain filter that tend to overpredict because they are unable to capture subtle differences in molecular substructures and/or properties.

2.3.8 Comments on the assessment of (Q)SAR predictions

In practice, the argumentation would be supplemented with additional information. Available information on the underlying toxicological mechanisms of action could also be considered in the assessment. For example, chemicals that are structural analogues but are known to act by a different

mechanism of action may not be suitable choices, since in these cases the relationship between chemical structure and toxicity might not be valid. Unfortunately, such detailed information is rarely available. Furthermore, if the raw experimental data are available, these could be analysed in more depth. In this example, the conclusions of the assessor (mutagenic or non-mutagenic) are adopted without question, but it is possible that a more in-depth analysis might result in differences of opinion based on the raw data. Clearly, this level of analysis requires toxicological expertise and is more time-consuming.

3. Summary and Conclusioms

In this report, the current regulatory framework for documenting and assessing the validity of (Q)SAR models, and the adequacy of (Q)SAR predictions, is described, and some of the scientific issues encountered when applying the framework are illustrated via case studies on the genotoxicity prediction of selected pesticides. Two examples are given that illustrate that these issues are not trivial: a) the interpretation of different diagnostics of model fit and predictivity, which requires a reasonable level of statistical expertise; and b) the selection of suitable analogues to fill a data gap by read-across or to substantiate a QSAR prediction, which requires a expertise in chemistry and biology (toxicology) and access to a range of computational tools and databases.

The framework developed for REACH should be sufficiently flexible to be applied in the risk assessment all types of chemicals and products, irrespective of the legislation under which they are regulated (industrial chemicals, biocides, PPP etc). However, more detailed guidance is needed on how to evaluate models and interpret their predictions in specific regulatory contexts (e.g. establishing TTC values for low level pesticide metabolites).

The QSAR model reporting formats (QMRF and QPRF) were developed on the basis of consensus with the main stakeholders (industry and authorities). They capture a level of resolution that is a compromise between scientific rigour and practicality. However, it is not always clear how much detail should be included under the different headings, and what kind of information is pertinent for models developed by different methodological approaches - in addition to traditional QSAR modelling based on (multiple) linear regression, there is an increasing use of "novel" model-building methods such as Support Vector Machines (e.g. Ferrari & Gini, 2010), artificial neural networks, instance-based learning (e.g. Helma, 2006, Raevsky *et al.*, 2010) and consensus modelling (e.g. Hewitt *et al.*, 2007). If models based on such methods are to gain acceptance, they need to be understandable to the assessor, and described with a sufficient level of transparency to form the basis for regulatory decision making. Furthermore, and more generally, in the regulatory assessment of chemicals, there should be greater emphasis on the development of models that are suitable for specific regulatory purposes, rather than trying to understand whether existing models (developed for other reasons) might be useful.

It is noteworthy that the current reporting formats focus on the result of the modelling process in the case of the QMRF (i.e. the model validation characteristics) and its application to a chemical of interest in the case of the QPRF (i.e. the characteristics of the prediction), rather than on the modelling and prediction process itself. It has been argued that the acceptability of (Q)SAR predictions could be improved through the development of principles and standards for Good Computer Modelling Practice (GCMP), analogous to the principles of Good Laboratory Practice (Judson, 2010).

The usefulness of a model, and in particular the adequacy of a model prediction, can only be considered in the context of the specific application, including the regulatory purpose, in which the prediction is being used (e.g. in a weight-of-evidence assessment with experimental data) and the consequence of being wrong. In the context of pesticide risk assessment, it is expected that QSAR analysis will be applied in the context of a Threshold of Toxicological Concern (TTC) decision scheme (CRD, 2010). The TTC is a generic human exposure level for chemicals below which there is low probability of risk to human health, assuming lifetime exposure. The principle of TTC is built on the premise that a safe level of exposure can be identified for chemicals present at low concentrations in the diet, even for those with unknown toxicity, on the basis of their chemical structure (Barlow, 2005).

In other words, QSARs are not considered here as standalone methods to directly fill data gaps in hazard assessment. Instead, they are being used to identify particular health concerns that may warrant specific thresholds of toxicological concern. In the TTC scheme by Kroes *et al*. (2004), and in the subsequent modifications by Munro *et al*. (2008) and Felter *et al*. (2009), there are three Cramer classes (I - low, II- moderate, and III-high) for different levels of non-cancer life-time risk,

corresponding to threshold doses of 1800, 540 and 90 μg/day/person, respectively. In the case of chemicals containing a structural alert for potential genotoxicity, a lower TTC of 0.15 μg/day is applied. These schemes refer to structural alerts, although the prediction of potential genotoxicity by a QSAR is presumably equivalent. However, it is unclear what is meant by an alert or QSAR for potential genotoxicity – should this be any genotoxic effect (e.g. Ames mutagenicity) or should it be limited to (*in vivo*) effects that are strong enough to warrant regulatory classification? Another open question is where structural alerts and QSARs for carcinogenicity fit in such TTC schemes. Presumably, such models would also be used to trigger a TTC of 0.15 μg/day, especially since most models for potential carcinogenicity are effectively modelling DNA reactivity (like models for potential genotoxicity). However, some models (e.g. Toxtree Benigni-Bossa) make predictions of non-genotoxic carcinogenicity. Non-genotoxic carcinogens which also have the potential to bioaccumulate are typically excluded from TTC schemes. Furthermore, high potency carcinogens (e.g. aflatoxin-like, azoxy and N-nitroso compounds) are also excluded, not as a matter of principle, but because there has been insufficient analysis of their potency distributions on the basis of existing TTC databases.

The crucial question in the application of QSAR is whether any model, or combination of models, is "good enough" for the regulatory purpose (in this case the identification of potential genotoxins). This cannot be answered in the absence of clearly defined performance criteria, and these should be set by the risk assessor and risk manager. For the purpose of pesticide risk assessment in the context of a TTC scheme, the most important criterion is expected to be minimisation of the false negative rate. A global statistical analysis has shown that this can indeed be minimised by combining the use of two or more models. In the case of Ames mutagenicity prediction, the generation of false negatives was found to range from 7-34% (Worth *et al.*, 2010).

At present, there is also little guidance on how to supplement (Q)SAR and read-across predictions with information on biokinetics and mechanisms of action, in order to develop weight-of-evidence based arguments for the replacement of (animal) testing. It is noticeable that there are no internationally adopted reporting formats specifically for metabolic simulation tools or Physiologically Based Biokinetic (PBBK) models, although in the latter case, good modelling practices are under development (Loizou *et al.*, 2008). Therefore, further efforts are needed to develop strategies for integrating different kinds of experimental and non-testing information and guidance is needed on how to report this in a transparent and consistent manner.

In conclusion, there is still a considerable need to gain and share experience in the practical application of *in silico* prediction tools, with a view to improving the scientific robustness, transparency and consistency of chemical risk assessments in which data gaps are filled by computational modelling.

4. Acknowledgements and Disclaimer

This work is based partly on a project carried out under the terms of a grant awarded by the European Food Safety Authority (EFSA) to the European Commission's Joint Research Centre (JRC), Institute for Health & Consumer Protection, Ispra, Italy (Contract number: SLA/EFSA-JRC/2009-01). The full report of the project is freely available from the EFSA website (JRC, 2010).

Any conclusions and opinions expressed in this document are those of the authors as individual scientists and do not constitute an official position by the JRC or the European Commission.

The authors are grateful to Sharon Munn (JRC) for reviewing and providing useful comments on this report.

5. References

Aptula AO, Jeliazkova NG, Schultz TW & Cronin MTD (2005). The Better Predictive Model: High q2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set? QSAR & Combinatorial Science 24(3), 385-396.

Barlow S (2005). Threshold of toxicological concern: a tool for assessing substances of unknown toxicity present at low levels in the diet. ILSI Concise Monograph Series, ISBN 1-57881-188-0, ILSI Press, Washington DC and Brussels.

Benigni R, Bossa C, Jeliazkova N, Netzeva T & Worth A (2008). The Benigni / Bossa rulebase for mutagenicity and carcinogenicity – a module of Toxtree. EUR 23241 EN. http://ecb.jrc.ec.europa.eu/qsar/publications/

Benigni R, Bossa C & Worth A (2010). Structural analysis and predictive value of the rodent in vivo micronucleus assay results. Mutagenesis 25, 335-341.

Boggia R, Forina M, Fossa P & Mosti L (1997). Chemometric Study and Validation Strategies in the Structure-Activity Relationships of New Cardiotonic Agents. Quantitative Structure-Activity Relationships 16(3), 201-213.

Cook RD & Nachtsheim CJ (1980). A Comparison of Algorithms for Constructing Exact D-Optimal Designs. Technometrics 22(3), 315-324.

Cramer RDI, Bunce JD, Patterson DE & Frank IE (1988). Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. Quantitative Structure-Activity Relationships 7(1), 18-25.

CRD (2010) Applicability of thresholds of toxicological concern in the dietary risk assessment of metabolites, degradation and reaction products of pesticides. Report from the UK Chemicals Regulation Directorate (CRD) to the European Food Safety Authority (EFSA). Available from http://www.efsa.europa.eu/en/scdocs/scdoc/44e.htm

Cronin MTD, Bigot F, Enoch SJ, Madden JC, Roberts DW & Schwöbel J (2009). The *in chemico–in silico* interface: challenges for integrating experimental and computational chemistry to identify toxicity? Alternatives to Laboratory Animals 37, 513-521.

Daszykowski M, Walczak B & Massart DL (2002). Representative subset selection. Analytica Chimica Acta 468(1), 91-103.

Dearden JC, Cronin MTD & Kaiser KLE (2009). How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR and QSAR in Environmental Research 20, 241-266.

ECHA (2008). Guidance on Information Requirements and Chemical Safety Assessment. Chapter R6. European Chemicals Agency, Helsinki, Finland. Available at: http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_en.htm?time=1252064523#r6

ECHA (2010). Practical guide 5. How to report (Q)SAR. European Chemicals Agency, Helsinki, Finland.

Felter S, Lane RW, Latulippe ME, Llewellyn GC, Olin SS, Scimeca JA & Trautman TD (2009). Refining the threshold of toxicological concern (TTC) for risk prioritization of trace chemicals in food. Food & Chemical Toxicology 47(9), 2236-2245.

Ferrari T & Gini G (2010). An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chemistry Central*, **4**, S2. Open access at: http://www.journal.chemistrycentral.com/content/4/S1/S2

Gramatica P (2007). Principles of QSAR models validation: internal and external. QSAR & Combinatorial Science 26(5), 694-701.

Hawkins DM (2003). The Problem of Overfitting. Journal of Chemical Information and Computer Sciences 44(1), 1-12.

Helma C (2006). Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. Molecular Diversity 10, 147-158.

Hewitt M, Cronin, MTD Madden JC, Rowe PH, Johnson C, Obi A & Enoch SJ (2007). Consensus QSAR models: do the benefits outweigh the complexity? Journal of Chemical Information and Modeling 47, 1460-1468.

Jeliazkova N, Jaworska J & Worth AP (2010). Open source tools for read across and category formation, in *In Silico Toxicology. Principles and Applications.* MTD Cronin & J Madden (Eds). Royal Society of Chemistry, Cambridge, UK, pp 408-445

JRC (2010). Applicability of QSAR analysis to the evaluation of the toxicological relevance of metabolites and degradates of pesticide active substances for dietary risk assessment. Report from the European Commission's Joint Research Centre (JRC) to the European Food Safety Authority (EFSA). Available from http://www.efsa.europa.eu/en/scdocs/scdoc/50e.htm

Judson PN (2010). Towards establishing good practice in the use of computer prediction. The Quality Assurance Journal, in press. Early online: doi: 10.1002/qaj.457

Kennard RW & Stone LA (1969). Computer Aided Design of Experiments. Technometrics 11(1), 137-148.

Loizou G, Spendiff M, Barton HA, Bessems J, Bois FY, Bouvier d'Yvoire M, Buist H, Clewell III HJ, Meek B, Gundert-Remy U, Goerlitz G & Schmitt W (2008). Development of good modelling practice for physiologically based pharmacokinetic models for use in risk assessment: the first steps. Regulatory Toxicology and Pharmacology 50 (3), 400-411.

Kohonen T (1998). The self-organizing map. Neurocomputing 21(1-3), 1-6.

Kroes R, Renwick AG, Cheeseman M, Kleiner J, Mangelsdorf I, Piersma A, Schilter B, Schlatter J, van Schothorst F, Vos JG & Würtzen G (2004). Structure-based thresholds of toxicological concern (TTC): guidance for application to substances present at low levels in the diet. Food & Chemical Toxicology 42(1), 65-83.

Lindgren F, Hansen B, Karcher W, Sjöström M & Eriksson L (1996). Model validation by permutation tests: Applications to variable selection. Journal of Chemometrics 10(5-6), 521-532.

Maunz A & Helma C (2008). Prediction of chemical toxicity with local support vector regression and activity-specific kernels. SAR and QSAR in Environmental Research **19**(5-6), 413-431.

Mostrag-Szlichtyng A & Worth AP (2010). Review of QSAR Models and Software Tools for predicting Biokinetic Properties. JRC Technical Report EUR 24377 EN. Publications Office of the European Union, Luxembourg. Available at: http://ecb.jrc.ec.europa.eu/qsar/publications/

Munro IC, Renwick AG & Danielewska-Nikiel B (2008). The Threshold of Toxicological Concern (TTC) in risk assessment. Toxicology Letters 180(2), 151-156.

OECD (2005). Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. Series on Testing and Assessment Number 34. ENV/JM/MONO(2005)14. Organisation for Economic Cooperation and Development, Paris, France. Available at: http://www.oecd.org/

OECD (2007). Guidance Document on the Validation of (Quantitative) Structure Activity Relationship [(Q)SAR] Models. OECD Series on Testing and Assessment No. 69. ENV/JM/MONO(2007)2. Organisation for Economic Cooperation and Development, Paris, France. Available at: http://www.oecd.org/

Osten DW (1988). Selection of optimal regression models via cross-validation. Journal of Chemometrics 2(1), 39-48.

Raevsky OA, Grigor'ev VJ, Modina EA & Worth AP (2010). Prediction of acute toxicity to mice by the Arithmetic Mean Toxicity (AMT) modelling approach. SAR and QSAR in Environmental Research 21, 265-275.

Sanderson DM & Earnshaw CG (1991). Computer prediction of possible toxic action from chemical structure; the DEREK system. Human and Experimental Toxicology 10, 261-273.

Serafimova R, Fuart Gatnik M & Worth A (2010). Review of QSAR Models and Software Tools for predicting Genotoxicity and Carcinogenicity. JRC Technical Report EUR 24427 EN. Publications Office of the European Union, Luxembourg. Available at: http://ecb.jrc.ec.europa.eu/qsar/publications/

Snee RD (1977). Validation of Regression Models: Methods and Examples. Technometrics 19(4), 415-428.

Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Öberg 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 and Modeling 48(9), 1733-1746.

Todeschini R & Consonni V (2009). Molecular Descriptors for Chemoinformatics. In *Series of Methods and Principles in Medicinal Chemistry* (Mannhold R., Kubinyi H, Timmerman H, Eds), Volume 41, Wiley-VCH.

Todeschini R, Consonni V & Maiocchi A (1999). The K correlation index: theory development and its application in chemometrics. Chemometrics and Intelligent Laboratory Systems 46(1), 13-29.

Topliss JG & Edwards RP (1979). Chance factors in studies of quantitative structure-activity relationships. Journal of Medicinal Chemistry 22(10), 1238-1244.

Wehrens R, Putter H & Buydens LMC (2000). The bootstrap: a tutorial. Chemometrics and Intelligent Laboratory Systems 54, 35-52.

Worth AP (2010). The role of QSAR methodology in the regulatory assessment of chemicals, Chapter 13 in *Recent Advances in QSAR Studies: Methods and Applications.* T Puzyn, J Leszczynski & MTD Cronin (Eds). Springer, Heidelberg, Germany, pp. 367-382.

Worth A, Lapenna S, Lo Piparo E, Mostrag-Szlichtyng A & Serafimova R (2010). The Applicability of Software Tools for Genotoxicity and Carcinogenicity Prediction: Case Studies relevant to the Assessment of Pesticides. JRC Technical Report EUR 24640 EN. Publications Office of the European Union, Luxembourg. Available at: http://ecb.jrc.ec.europa.eu/qsar/publications/

Table 1 Explanation of the OECD (Q)SAR validation principles

A (Q)SAR model should be …	Explanation
1) associated with a defined endpoint.	Aims to ensure transparency in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. The endpoint refers to any physicochemical property, biological effect (human health or ecological) environmental fate parameter that can be measured and therefore modelled.
2) expressed in the form of an unambiguous algorithm.	Aims to ensure transparency in the description of the model algorithm.
3) associated with a defined domain of applicability	Recognises that (Q)SARs are reductionist models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions. The principle expresses the need to justify that a given model is being used within the boundary of its limitations when making a given prediction.
4) appropriate measures of goodness-of–fit, robustness and predictivity.	Expresses the need to provide two types of information: a) the internal performance of a model (as represented by goodness-of-fit and robustness), determined by using a training set; and b) the predictivity of a model, determined by using an appropriate test set.
5) associated with a mechanistic interpretation, wherever possible.	Aims to ensure that there is an assessment of the mechanistic associations between the descriptors used in a model and the endpoint being predicted, and that any association is documented. Where a mechanistic interpretation is possible, it can add strength to the confidence in the model already established on the basis of Principles 1-4. The wording of this principle, while seemingly redundant in its use of "where possible" emphasises that is not always possible to provide a mechanistic interpretation of a given (Q)SAR.

Table 2. Common errors in the development of QSARs (after Dearden *et al.*, 2008)

No	Type of error	Related OECD principle(s)
1	Failure to take account of data heterogeneity	1
2	Use of inappropriate endpoint data	1
3	Use of collinear descriptors	2, 4 and 5
4	Use of incomprehensible descriptors	2 and 5
5	Error in descriptor values	2
6	Poor transferability of model	2
7	Inadequate/undefined applicability domain	3
8	Unacknowledged omission of data points	3
9	Use of inadequate data	3
10	Replication of compounds in dataset	3
11	Too narrow a range of endpoint values	3
12	Over-fitting of data	4
13	Use of excessive numbers of descriptors	4
14	Lack of/inadequate statistics	4
15	Incorrect calculation	4
16	Lack of descriptor auto-scaling	4
17	Misuse/misinterpretation of statistics	4
18	No consideration of distribution of residuals	4
19	Inadequate training/test set selection	4
20	Inadequate validation	4
21	Lack of mechanistic interpretation	5

Table 3. Goodness-of-fit parameters for regression models

Statistic	Definition	Interpretation
R^2 (Multiple determination coefficient)	The variance of the response that is explained by the regression model.	* R^2=0 – the lack of linear relationship between the response and predictors variables * R^2= 1 – the perfect linear fit * R^2>0.5 – the variance explained by the model is higher than the unexplained variance. The value of R^2 increases with additional predictors, even if they do not contribute to the explained variance in the response.
R^2_{adj} (Adjusted multiple determination coefficient)	The variance of model's response that is explained by the regression for respective degrees of freedom.	The value of R^2_{adj} decreases if a predictor is added to the equation but does not contribute to the explanation of the variance.
s (Standard error of estimate)	The dispersion of the observed values about the regression line.	The smaller the value of s, the higher the reliability of prediction. Standard error of estimate smaller than the experimental error of the biological data is indicative of an overfitted model.
Q^2(Y) (Cross-validation coefficient)	Explained variance in prediction.	Indicates the model with the highest predictive ability. Q^2(Y) does not increase after a certain degree of model complexity and indicates the zone with a balance between predictive power and reasonable fit. The difference between R^2(Y) and Q^2(Y) should not exceed 0.3. A larger difference indicates an overfitted model, the presence of irrelevant predictors or outliers in the data.
F-value	The ratio between explained and unexplained model variance for a given number of degrees of freedom.	The regression equation/coefficients are statistically significant if calculated F-value/t-value is greater than a tabulated value for the chosen level of significance (typically 95%) and the corresponding degrees of freedom of F/t (p). Higher F-/t-values of a regression equation/coefficient correspond to greater statistical significance. Significance of the equation at the 95% level means that there is 5 % probability that the dependence found is obtained due to chance correlations between the variables. Significance of a coefficient at the 95% level means that there is a 5 % probability that the coefficient of a given variable is not significantly different from zero.
t-test	The ratio between estimated regression coefficient and its standard deviation for a given number of degrees of freedom.	

Table 4. Checklist of questions to help establish the adequacy of a (Q)SAR prediction

No	Question	Interpretation
1	Is the predicted endpoint clearly defined?	If the endpoint is not clearly defined, the use of the prediction will be open to different interpretations, and thus of questionable value.
2	If the predicted endpoint is clearly defined ("yes" to Q1), does it represent a direct information requirement under the legislation of interest, or is it related to one of the information requirements?	If the predicted endpoint corresponds directly with an information requirement, it may be possible to use the prediction instead of experimental data. Alternatively, if the predicted endpoint is indirectly related to an information requirement, it may be useful as supporting information.
3	If the model is statistically based (as opposed to knowledge-based), is the model training set fully available?	If the model training set of a statistically-based model is not fully available (e.g. because the data are proprietary), it will be impossible for another practitioner to independently reproduce the model, which may reduce confidence in the model estimates. However, this may not be an issue if the model is coded into a software tool. This does not apply to knowledge-based models, which are based on human knowledge and do not have a clearly identified training set.
4	Is the method used to develop the model documented or referenced (e.g. in a scientific paper or QMRF)	If the details of model development are not documented, it will be impossible for another practitioner to independently develop and confirm the model, which may reduce confidence in the model estimates. Even if the method is documented, it will require a QSAR specialist to determine whether the documentation is sufficiently detailed to reproduce the model.
5	Is information available (in terms of statistical properties) concerning the performance of the model, including its goodness-of-fit, predictivity, robustness and error of prediction (uncertainty)?	The statistical properties of a model can provide evidence of its usefulness in a given context (e.g. need to minimise false negatives) and can also be used to assess whether the model has been overfitted (see question 7).
6	If the model is statistically based (as opposed to knowledge-based), does examination of the available statistics indicate that the model may have been overfitted?	The overfitting of statistically based models is undesirable because it can result in unpredictable errors. This consideration does not apply to knowledge-based models. Overfitted statistical models typically show worse predictivity (outside their training sets) than their internal validation statistics imply. Several simple diagnostics exist, for example: a) the model estimation error (uncertainty of prediction) should not be significantly less than the known experimental error. b) the ratio of datapoints (chemicals) to variables (descriptors) should be at least 5:1.

No	Question	Interpretation
7	Does the model training set contain the chemical of interest?	If the model training set contains the chemical of interest, then a prediction is not needed because some experimental data is available for direct use.
8	Does the model make reliable predictions for analogues of the chemical structure of interest?	The generation of reliable predictions for analogues of the chemical of interest increases confidence in the prediction. In the case of a software tool, it should be indicated whether the software automatically identifies analogues and their associated data within the model training set. In the case of a literature model, it should be considered whether suitable analogues can be identified in the training set (if available).
9	Is the model prediction substantiated with argumentation based on the applicability domain of the model?	Confidence in a prediction is increased if information is available concerning the applicability domain of the model, and thus whether the model is applicable to the chemical of interest. The applicability domain can include physicochemical and structural space, as well as mechanistic and metabolic considerations.
10	Can the model prediction be easily reproduced?	Not all model predictions can be easily reproduced, depending on the complexity and transparency of the model development process, and the availability of a user-friendly software tool implementing the model. If the model is a simple SAR (structural alert) it should be possible to apply it by visual inspection. However, some differences of expert interpretation may arise. If the model is a QSAR in the form of a transparent mathematical formula, it will be possible to apply it in a spreadsheet (e.g. Excel). If the model is implemented in the form of a freely or commercially available software tool, it is possible for different users to verify the same result (even if the model development process is not transparent), thereby increasing confidence in the prediction.

Table 5. Application of the checklist of questions to Derek for Windows

No	Question (Q) and answer (A)
Q1	Is the predicted endpoint clearly defined?
A1	Yes. Derek provides predictions for mutagenicity, genotoxicity, chromosome damage and carcinogenicity (among other endpoints). The user can select the species and the endpoint of interest. Derek v.11 contains 87 alerts for mutagenicity, 4 for genotoxicity, 74 for chromosome damage and 56 for carcinogenicity.
Q2	If the predicted endpoint is clearly defined ("yes" to Q1), does it represent a direct information requirement under the legislation of interest, or is it related to one of the information requirements?
A2	Yes, genotoxicity test data are required under most types of chemicals legislation (e.g. industrial chemicals, pesticides, biocides)
Q3	If the model is statistically based (as opposed to knowledge-based), is the model training set fully available?
A3	Not applicable (knowledge-based predictions). Derek rules are not built from the automated analysis of training sets. The alerts are created by experts who compare the toxicity of compounds from various different sources (proprietary, public domain and freely available) and knowledge of the mechanism of action where known. The software gives additional information in the alert description (comments, references, examples).
Q4	Is the method used to develop the model documented or referenced (e.g. in a scientific paper or QMRF)
A4	Yes. A QMRF for the genotoxicity module is available in the JRC QSAR model database. There are also numerous papers which describe the development and use of Derek for different endpoints. The documentation which is supplementary to the software is very readable. Users receive support from the support centre of Lhasa Ltd. The original publication is: Sanderson M & Earnshaw CG (1991).
Q5	Is information available (in terms of statistical properties) concerning the performance of the model, including its goodness-of-fit, predictivity, robustness and error of prediction (uncertainty)?
A5	Not applicable.
Q6	If the model is statistically based (as opposed to knowledge-based), does examination of the available statistics indicate that the model may have been overfitted?
A6	Not applicable.
Q7	Does the model training set contain the chemical of interest?
A7	Not applicable.
Q8	Does the model make reliable predictions for analogues of the chemical structure of interest?
A8	The software does not identify a set of most similar analogues, but the case of most alerts, examples with experimental data are provided.
Q9	Is the model prediction substantiated with argumentation based on the applicability domain of the model?
A9	There is no defined applicability domain but all predictions are well supported with explanations, references, and examples. If boundary conditions exist for an alert, these are clearly noted.
Q10	Can the model prediction be easily reproduced?
A10	Yes, the software is commercially available and is easy to use.

Table 6. Application of the checklist of questions to CAESAR

No	Question (Q) and answer (A)
Q1	Is the predicted endpoint clearly defined?
A1	Yes, the endpoint is Ames (S. Typhimurium) mutagenicity
Q2	If the predicted endpoint is clearly defined ("yes" to Q1), does it represent a direct information requirement under the legislation of interest, or is it related to one of the information requirements?
A2	Yes, genotoxicity test data are required under most types of chemicals legislation (e.g. industrial chemicals, pesticides, biocides)
Q3	If the model is statistically based (as opposed to knowledge-based), is the model training set fully available?
A3	Yes, the training and test set will be soon available from the JRC QSAR model database
Q4	Is the method used to develop the model documented or referenced (e.g. in a scientific paper or QMRF)
A4	Yes, a QMRF is in preparation, based on the following publications: Ferrari T, Gini G & Benfenati E (2009). Support vector machines in the prediction of mutagenicity of chemical compounds. Proc NAFIPS 2009, June 14-17, Cincinnati, USA, p 1-6. Ferrari T & Gini G (2010). Developing a new computational intelligence approach to predict the mutagenicity of chemical compounds. Computational Intelligence, in press. Ferrari T & Gini G (2010). A new multistep model to predict mutagenicity from statistic analysis and relevant structural alerts. Central Chemistry 4, Suppl 1, S2.
Q5	Is information available (in terms of statistical properties) concerning the performance of the model, including its goodness-of-fit, predictivity, robustness and error of prediction (uncertainty)?
A5	Yes. Information on the accuracy (82.1%), sensitivity (90.6%) and specificity (71.4%) are provided in the QMRF.
Q6	If the model is statistically based (as opposed to knowledge-based), does examination of the available statistics indicate that the model may have been overfitted?
A6	The model is statistically based but should not be overfitted because the ratio of chemicals (3380) to descriptors (42) is 80.5.
Q7	Does the model training set contain the chemical of interest?
A7	The model training set includes some pesticides including parathion-methyl but not sodium nitroguaiacolate.

Q8	Does the model make reliable predictions for analogues of the chemical structure of interest?
A8	Yes, the Caesar software gives the chance to examine, for each compound submitted, the six most similar compounds found in the model training set. For these compounds the experimental value for the selected endpoint is shown, together with the prediction made by the model. The similarity measure employed by the Caesar software takes into account functional group similarity, constitutional similarity, ring similarity and fingerprint similarity. For parathion methyl (correctly predicted by the software), the similar structures obtained are: parathion methyl (input structure contained in the training set), aminofenitrothion, 1-ethenoxy-4-nitro-benzene, fenitrooxon, o-nitroanisole, N-hydroxy-N-(4-nitrophenyl)acetamide. All of them are predicted correctly by the software. For nitroguaiacolate (wrongly predicted by the software) the similar structures obtained are: o-nitroanisole, 1-ethoxy-3-nitro-benzene, 2,5-dinitrophenol, p-nitrosoanisole, 2-methoxy-1,3,5-trinitro-benzene, 1-ethenoxy-4-nitro-benzene. All of them are predicted correctly by the software.
Q9	Is the model prediction substantiated with argumentation based on the applicability domain of the model?
A9	Yes, Caesar addresses the applicability domain in several ways, namely by: a) checking whether the compound of interest falls in the descriptors space – if the compound is out of domain, this is noted in the output; b) providing a similarity score (1=identity) for the structure-based comparison with analogues; c) visual representation of the most similar compounds; d) by revealing the known and predicted the known and predicted toxicities for the analogues, thereby indicating the prediction error. Thus Caesar provides an assessment based on both the input (descriptor) space and the output (toxicological endpoint) space.
Q10	Can the model prediction be easily reproduced?
A10	Yes, the software is accessible in the form of a freely accessible web platform (http://www.caesar-project.eu) and is easy to use, even for non-specialists.

Table 7. Application of the checklist of questions to ToxBoxes

No	Question (Q) and answer (A)
Q1	Is the predicted endpoint clearly defined?
A1	Two models are available: a fragment-based model predicts Ames test results, whereas the genotoxicity hazards module is a knowledge-based expert system containing 27 known "genotoxicophore" fragments collected from literature.
Q2	If the predicted endpoint is clearly defined ("yes" to Q1), does it represent a direct information requirement under the legislation of interest, or is it related to one of the information requirements?
A2	Yes, genotoxicity test data are required under most types of chemicals legislation (e.g. industrial chemicals, pesticides, biocides)
Q3	If the model is statistically based (as opposed to knowledge-based), is the model training set fully available?
A3	The Ames model training set is available through the software. The data consist of Ames test results obtained with various *S. Typhimurium* and *E. Coli* bacteria strains.
Q4	Is the method used to develop the model documented or referenced (e.g. in a scientific paper or QMRF)
A4	No, not adequately. The software manual gives some general information about the model, but not a detailed description of model development. No research papers are cited.
Q5	Is information available (in terms of statistical properties) concerning the performance of the model, including its goodness-of-fit, predictivity, robustness and error of prediction (uncertainty)?
A5	No, information not available.
Q6	If the model is statistically based (as opposed to knowledge-based), does examination of the available statistics indicate that the model may have been overfitted?
A6	The model is statistically based but no information relating to overfitting could be found.
Q7	Does the model training set contain the chemical of interest?
A7	The training set includes parathion-methyl, but not sodium nitroguaiacolate.
Q8	Does the model make reliable predictions for analogues of the chemical structure of interest?
A8	Yes, the ToxBoxes software provides a list of the five most similar structures from its training set. Some general information on the similarity measure is provided – it is calculated using Tanimoto and fingerprint Max2 similarity. For methyl parathion, the five most similar structures identified are: parathion-methyl (the input structure), fenitrothion, parathion, hydroxymethyl-fenitrothion and fenitrooxon. The second compound (a non-mutagen) and the third (with an inconclusive Ames result) are predicted as indeterminate (probability of positive Ames test between 0.3 and 0.7), whereas the fourth and fifth (the known mutagens) are predicted correctly. For nitroguaiacolate five similar structures are identified, including one false positive (FP), three true positives (TP) and one true negative (TN): m-nitroanisole (FP), o-nitroanisole (TP), p-nitroanisole (TP), 2,5-dimethoxy-4-nitroazobenzene (TP), and 3-methyl-4-nitrophenol (TN)
Q9	Is the model prediction substantiated with argumentation based on the applicability domain of the model?
A9	Yes, the software calculates a Reliability Index for every prediction, which takes account of the applicability domain.
Q10	Can the model prediction be easily reproduced?
A10	Yes, the software is very easy to use, even for non-specialists.

Table 8. Application of the checklist of questions to Lazar

No	Question (Q) and answer (A)
Q1	Is the predicted endpoint clearly defined?
A1	Yes. Mutagenicity (Ames test).
Q2	If the predicted endpoint is clearly defined ("yes" to Q1), does it represent a direct information requirement under the legislation of interest (e.g. PPP directive), or is it related to one of the information requirements?
A2	Yes, genotoxicity test data are required under the PPP.
Q3	If the model is statistically based (as opposed to knowledge-based), is the model training set fully available?
A3	Yes. Lazar contains two models for Ames mutagenicity based on different training sets. They are both publicly available: the Kazius/Bursi dataset and Toxbenchmark dataset.
Q4	Is the method used to develop the model documented or referenced (e.g. in a scientific paper or QMRF)
A4	Yes, in: Helma, C. (2006). Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. Molecular Diversity 10, 147-158.
Q5	Is information available (in terms of statistical properties) concerning the performance of the model, including its goodness-of-fit, predictivity, robustness and error of prediction (uncertainty)?
A5	Yes. Leave-one-out (LOO) and external validation tests indicate that Salmonella mutagenicity can be predicted with 85% accuracy for compounds within its applicability domain.
Q6	If the model is statistically based (as opposed to knowledge-based), does examination of the available statistics indicate that the model may have been overfitted?
A6	This is not applicable for a model such as Lazar, which is based on an instance-based learning approach in which the prediction is generated from the nearest neighbours but the chemical of interest is excluded from the learning algorithm (and is therefore never in the training set).
Q7	Does the model training set contain the chemical of interest?
A7	Parathion-methyl is present in both the Kazius/Bursi dataset and Toxbenchmark dataset. However, the learning algorithm always excludes the chemical of interest (in case data are available). Nitroguaiacoloate is not present in either training set.
Q8	Does the model make reliable predictions for analogues of the chemical structure of interest?
A8	The model output can be inspected to identify the nearest neighbours, which may be considered as analogues. However, the relevance of such analogues needs to be checked.
Q9	Is the model prediction substantiated with argumentation based on the applicability domain of the model?
A9	Yes.
Q10	Can the model prediction be easily reproduced?
A10	Yes, the software is freely accessible online and is easy to use.

Table 9. Application of the checklist of questions to TOPKAT

No	Question (Q) and answer (A)
Q1	Is the predicted endpoint clearly defined?
A1	Yes. Mutagenicity (Ames test).
Q2	If the predicted endpoint is clearly defined ("yes" to Q1), does it represent a direct information requirement under the legislation of interest (e.g. PPP directive), or is it related to one of the information requirements?
A2	Yes, genotoxicity test data are required under the PPP.
Q3	If the model is statistically based (as opposed to knowledge-based), is the model training set fully available?
A3	The model is statistically based but the training set is not available.
Q4	Is the method used to develop the model documented or referenced (e.g. in a scientific paper or QMRF)
A4	No. Topkat`s QSAR (QSTR) models are linear equations, which have not been described in detail. The variables in the equations are mentioned: the calculated values of electronic attributes, shape and symmetry indices, and transport-related descriptors such as molecular weight and VlogP. The electronic attributes are expressed in terms of the electrotopological state (E-State) values of specially designed 1-atom and 2-atom fragments of non-hydrogen atoms in different hybridization states. While the model equation is not reported, the individual contributions of all the descriptors used by the TOPKAT model to calculate the total contribution of the query structure are reported.
Q5	Is information available (in terms of statistical properties) concerning the performance of the model, including its goodness-of-fit, predictivity, robustness and error of prediction (uncertainty)?
A5	No.
Q6	If the model is statistically based (as opposed to knowledge-based), does examination of the available statistics indicate that the model may have been overfitted?
A6	Insufficient information to judge.
Q7	Does the model training set contain the chemical of interest?
A7	Yes. The chosen chemicals of interest are: 1. parathion-methyl – computed probability of mutagenicity = 0.989 (true positive mutagen) 2. nitroguiacolate (anion) – computed probability of mutagenicity = 0.767 (true positive mutagen) TOPKAT gives a warning if a query molecule contains a substructure which was not considered during the model development process. TOPKAT performs this by comparing all 1- and 2-atom fragments in the query structure with the list of fragments from the training set of the model. Should the query structure contain an uncovered fragment; i.e., a fragment that is in the query structure but not in the training set, it will caution you as to the acceptability of the assessment. These validation criteria were satisfied for the two chosen chemicals.

Q8	Does the model make reliable predictions for analogues of the chemical structure of interest?
A8	TOPKAT performs similarity searching of the model's database for the query compound, i.e., similarity in descriptor space (type of descriptors and their values), not chemical structure space. For each analogue of the query structure, the Ames Mutagenicity Prediction module v.3.1 was applied, which returned the following results: a) the actual experimental result; b) the TOPKAT prediction; c) whether the compound was in the training set; d) the similarity distance from the query on a scale of 0.0 - 1.0. The smaller the distance, the greater the similarity. With this information, one can determine whether the query structure lies in an information-rich region of the model data space, and f similar compounds are accurately and/or correctly predicted by the model. A TOPKAT analogue search for the chosen chemicals returned: 1. 58 analogues of parathion-methyl, the majority of which were included in the training set and were correctly predicted. 2. 58 analogues of nitroguiacolate, all of which were included in the training set and were correctly predicted.
Q9	Is the model prediction substantiated with argumentation based on the applicability domain of the model?
A9	Yes. TOPKAT estimates the confidence in the prediction by applying the patented Optimal Predictive Space (OPS) validation method. The OPS is TOPKAT's formulation of the model applicability domain - a unique multivariate descriptor space in which a given model is considered applicable. Any prediction generated for a query structure outside of the OPS space is considered unreliable.
Q10	Can the model prediction be easily reproduced?
A10	The software is commercially available and easy to use. However, the algorithm is not sufficiently transparent to be reprogrammed into another software platform.

Table 10. Application of the checklist of questions to HazardExpert

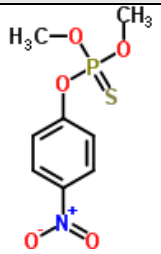
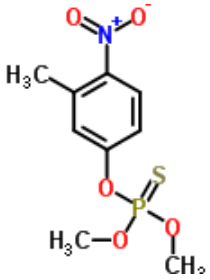
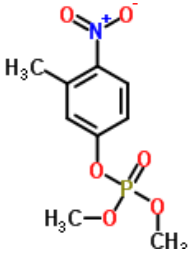
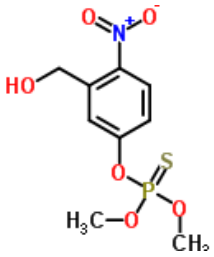
No	Question (Q) and answer (A)
Q1	Is the predicted endpoint clearly defined?
A1	No, not in a specific way. The software predicts a range of endpoints (oncogenicity, mutagenicity, teratogenicity, etc.) contributing to the toxicity of an organic molecule. The user defines the species (soil invertebrates, fish, birds or mammals), route of administration (oral, inhalation/ intrabrachial in case of fish), duration of exposure (single, repeated or permanent) and dosage (low, medium or high). However, for more detailed information (e.g. if "oncogenicity" is synonymous to "carcinogenicity") on the endpoints it is necessary to contact the developer, as the manual does not provide a reference to a relevant scientific paper.
Q2	If the predicted endpoint is clearly defined ("yes" to Q1), does it represent a direct information requirement under the legislation of interest (e.g. PPP directive), or is it related to one of the information requirements?
A2	Yes, genotoxicity test data are required under the PPP.
Q3	If the model is statistically based (as opposed to knowledge-based), is the model training set fully available?
A3	Not applicable (knowledge-based model). The predictions are based on a fully available knowledge database, containing the list of 105 pre-defined toxic fragments, which cannot be modified. The knowledge is based on a report by the US EPA and scientific information collected by CompuDrug Chemistry Ltd. HazardExpert also allows the users to create their own knowledge bases with modified rules.
Q4	Is the method used to develop the model documented or referenced (e.g. in a scientific paper or QMRF)
A4	The methodology is not well documented. The software manual neither describes it, nor indicates any relevant scientific papers. Basically, HazardExpert works by searching the query structure for known toxicophores which are held in the toxic fragments knowledge base. The identification of a toxicophore leads to estimates of the toxicity endpoints by triggering rules in the knowledge-bases. The rules describe toxic segments and their effects on various biological systems, and are based on the toxicological knowledge and expert judgement, supported by QSAR models and fuzzy logic (which simulates the effects of different exposure conditions) (Dearden *et al*., 1997).
Q5	Is information available (in terms of statistical properties) concerning the performance of the model, including its goodness-of-fit, predictivity, robustness and error of prediction (uncertainty)?
A5	No, not for genotoxicity
Q6	If the model is statistically based (as opposed to knowledge-based), does examination of the available statistics indicate that the model may have been overfitted?
A6	Not applicable – knowledge-based.
Q7	Does the model training set contain the chemical of interest?
A7	Not applicable – knowledge-based.
Q8	Does the model make reliable predictions for analogues of the chemical structure of interest?
A8	The software does not provide any information on analogues to assist the user.
Q9	Is the model prediction substantiated with argumentation based on the applicability domain of the model?
A9	Not applicable.
Q10	Can the model prediction be easily reproduced?
A10	Yes, the software is easy to use, even for non-specialists.

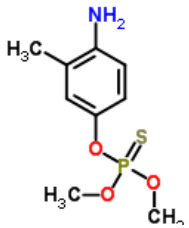
Table 11. Application of the checklist of questions to Toxtree

No	Question (Q) and answer (A)
Q1	Is the predicted endpoint clearly defined?
A1	No. The software makes generic predictions of genotoxicity using two rulebases: a) the Benigni/Bossa rulebase is a decision tree for estimating both the genotoxic and non-genotoxic carcinogenicity potentials of chemicals, based on the rules published in the EC report "The Benigni/Bossa rulebase for mutagenicity and carcinogenicity – a module of Toxtree", by Benigni et al (2008).[53] The present list of SAs (over 30) refers mainly to genotoxic carcinogenicity, and also includes a number of SAs for potential non-genotoxic carcinogens. In addition to the SAs, the module performs discriminant QSAR analyses for i) the mutagenic activity in the Salmonella typhimurium TA100 strain (Ames test) of aromatic amines (QSAR6) and alpha,beta-unsaturated aldehydes (QSAR13), and ii) the carcinogenicity in rodents of aromatic amines (QSAR8). The underlying mechanism(s) for the triggering of an alert for genotoxic carcinogenicity are not clearly defined, as these may include different possible mechanisms of genotoxicity (e.g. Ames mutagenicity, chromosomal aberration, chromosomal instability, etc.) which may be linked to carcinogenicity. b) the in vivo micronucleus assay plug-in, for the prediction of the outcome of this assay in rodents, is based on the SAs published in the EC report "Development of structural alerts for the in vivo micronucleus assay in rodents", in Benigni *et al.* (2009) and Benigni *et al*. (2010). Thus, except for the QSAR model estimates of the Ames test and the SA-based predictions for the in vivo micronucleus assay, the predictions of (non)genotoxic carcinogenicity are intended in a broad sense, rather than representing specific mechanisms of genotoxic damage.
Q2	If the predicted endpoint is clearly defined ("yes" to Q1), does it represent a direct information requirement under the legislation of interest, or is it related to one of the information requirements?
A2	Yes, genotoxicity test data are required under most types of chemicals legislation (e.g. industrial chemicals, pesticides, biocides)
Q3	If the model is statistically based (as opposed to knowledge-based), is the model training set fully available?
A3	Yes, in the case of the the QSAR models included in the Benigni/Bossa rulebase, the training sets are described in Benigni *et al.* (2008).
Q4	Is the method used to develop the model documented or referenced (e.g. in a scientific paper or QMRF)
A4	Yes, described in Benigni *et al*. (2008, 2009).
Q5	Is information available (in terms of statistical properties) concerning the performance of the model, including its goodness-of-fit, predictivity, robustness and error of prediction (uncertainty)?
A5	Yes. Measures of performance include r2 values and external predictivity values (Benigni *et al*., 2009). In addition, the included SAs have been tested against the ISSCAN carcinogenicity and mutagenicity database, resulting in 70% and 78% accuracy, respectively.
Q6	If the model is statistically based (as opposed to knowledge-based), does examination of the available statistics indicate that the model may have been overfitted?
A6	There is no evidence of overfitting. The included QSAR models have been tested using external test sets, which resulted in high predictive accuracy (Benigni *et al*., 2009).

Q7	Does the model training set contain the chemical of interest?
A7	No, parathion-methyl (true positive mutagen) and sodium nitroguaiacolate (false positive mutagen) are not included in the QSAR training sets.
Q8	Does the model make reliable predictions for analogues of the chemical structure of interest?
A8	The software does not provide information on analogues.
Q9	Is the model prediction substantiated with argumentation based on the applicability domain of the model?
A9	Yes. The QSAR models are applicable to aromatic amines and alpha,beta-unsaturated aldehydes, with a few specified exceptions. The SAs are considered to be generally applicable.
Q10	Can the model prediction be easily reproduced?
A10	Yes, the software is freely available from the JRC and easy to use, even for non-specialists. Software and supporting documents are available at: http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/index.php?c=TOXTREE The software is based on qualitative SARs (structural alerts) and (in the case of the Benigni/Bossa module) QSARs in form of transparent mathematical formulas. These algorithms can therefore be reprogrammed into a different software platform.

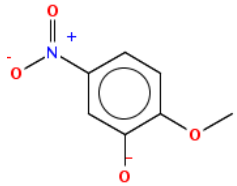
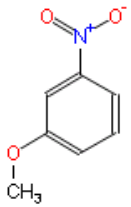
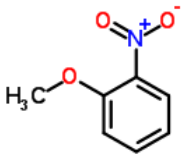
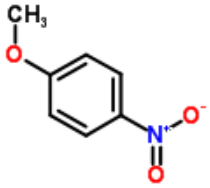
Table 12. Methyl parathion and some of its analogues

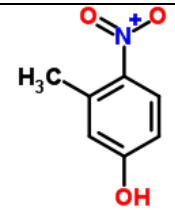
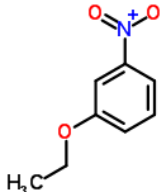
Chemical and experimental mutagenicity	Structure	Derek	CAESAR	ToxBoxes	Lazar	TOPKAT	HazardExpert	Toxtree
Methyl parathion Dimethoxy-(4-nitrophenoxy)-thioxo-phosphorane CAS 298-00-0 S=P(Oc1ccc(cc1)[N+]([O-])=O)(OC)OC **Mutagen**		Mutagen (TP)	Mutagen (TP) In TS	IND (p=0.631) In TS	Mutagen (TP) Mutagen (TP) In TS	Mutagen (TP)	Mutagen (TP)	Mutagen (TP)
Fenitrothion Dimethoxy-(3-methyl-4-nitro-phenoxy)-thioxo-phosphorane CAS 128-14-5 S=P(Oc1cc(c(cc1)[N+]([O-])=O)C)(OC)OC **Non-mutagen**		Mutagen (FP)	Non-Mutagen (TN) In TS	IND (p=0.668) In TS	Mutagen (FP) Mutagen (FP)	Mutagen (FP)	Mutagen (FP)	Mutagen (FP)
Fenitrooxon Dimethyl (3-methyl-4-nitro-phenyl) phosphate CAS 2255-17-6 [O-][N+](=O)c1c(cc(OP(=O)(OC)OC)cc1)C **Mutagen**		Mutagen (TP)	Mutagen (TP) In TS	Mutagen (TP) (p=0.889) In TS	Mutagen (TP) Mutagen (TP)	Non-Mutagen (FN)	Mutagen (TP)	Mutagen (TP)
Hydroxymethylfenitrothion Phosphorothioic acid (5-dimethoxyphosphinothioyloxy-2-nitro-phenyl)methanol CAS 59417-73-1 S=P(Oc1cc(c(cc1)[N+]([O-])=O)CO)(OC)OC		Mutagen (TP)	Mutagen (TP)	Mutagen (TP) (p =0.796) In TS	Mutagen (TP) Mutagen (TP)	Mutagen (TP)	Mutagen (TP)	Mutagen (TP)

Mutagen 4-Aminofenitrothion 4-dimethoxyphosphinothioyloxy-2-methyl-aniline CAS 13306-69-9 S=P(Oc1ccc(c(c1)C)N)(OC)OC **Mutagen**		Mutagen (TP)	Mutagen (TP) In TS	Mutagen (TP) (p =0.762)	Mutagen (TP) Mutagen (TP) In TS	Mutagen (TP)	Mutagen (TP)	Mutagen (TP)

p – probability of positive Ames test result; IND – indeterminate; FN – false negative; TN – true negative; TP – true positive; FP – false positive; In TS – compound in model training set

Table 13. Sodium nitroguaiacolate and some of its analogues

Chemical and experimental mutagenicity	Structure	Derek	CAESAR	ToxBoxes	Lazar Kazius / Toxbenchmark	TOPKAT	HazardExpert	Toxtree
Nitroguiacolate 2-methoxy-5-nitro-phenolate CAS 67233-85-6 [Na+].[O-]c1cc(ccc1OC)[N+]([O-])=O **Non mutagen**		Mutagen (FP)	Mutagen (FP)	IND (p=0.643)	Mutagen (FP) Mutagen (FP)	Mutagen (FP)	Mutagen (FP)	Mutagen (FP)
m-Nitroanisole 1-methoxy-3-nitro-benzene CAS 555-03-3 COc1cccc(c1)[N+]([O-])=O **Non mutagen**		Mutagen (FP)	Mutagen (FP) In TS	Mutagen (FP) (p=0.738) In TS	Mutagen (FP) Mutagen (FP) In TS	Mutagen (FP)	Mutagen (FP)	Mutagen (FP)
o-Nitroanisole 1-methoxy-2-nitro-benzene CAS 91-23-6 COc1c(cccc1)[N+](=O)[O-] **Mutagen**		Mutagen (TP)	Mutagen (TP) In TS	Mutagen (TP) (p=0.738) In TS	Mutagen (TP) Mutagen (TP) In TS	Mutagen (TP)	Mutagen (TP)	Mutagen (TP)
p-Nitroanisole 1-methoxy-4-nitro-benzene CAS 100-17-4 COc1ccc(cc1)[N+]([O-])=O **Mutagen**		Mutagen (TP)	Mutagen (TP) In TS	Mutagen (TP) (p=0.955) In TS	Mutagen (TP) Mutagen (TP) In TS	Non-Mutagen (FN)	Mutagen (TP)	Mutagen (TP)

3-Methyl-4-nitrophenol CAS 2581-34-2 O=[N+]([O-])c1c(cc(O)cc1)C **Non-mutagen**		Non-Mutagen (TN)	Mutagen (FP)	Non-Mutagen (TN) (p= 0.193) In TS	Mutagen (FP) Mutagen (FP)	Non-Mutagen (TN)	Mutagen (FP)	Mutagen (FP)
m-Nitrophenetole 1-Ethoxy-3-nitrobenzene CAS 621-52-3 CCOc1cc(ccc1)[N+](=O)[O-] **Non-mutagen**		Mutagen (FP)	Mutagen (FP) In TS	Mutagen (TP) (p= 0.717)	Mutagen (FP) Mutagen (FP) In TS	Non-Mutagen (TN)	Mutagen (FP)	Mutagen (FP)

p – probability of positive Ames test result; IND – indeterminate; FN – false negative; TN – true negative; TP – true positive; FP – false positive; In TS – compound in model training set

European Commission

EUR 24705 EN – Joint Research Centre – Institute for Health and Consumer Protection
Title: A Framework for assessing *in silico* Toxicity Predictions: Case Studies with selected Pesticides
Author(s): Andrew Worth, Silvia Lapenna, Elena Lo Piparo, Aleksandra Mostrag-Szlichtyng and Rositsa Serafimova
Luxembourg: Publications Office of the European Union
2011 – 46 pp. – 21 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN 978-92-79-19081-0
doi:10.2788/29048

Abstract

In the regulatory assessment of chemicals, the use of *in silico* prediction methods such as (quantitative) structure-activity relationship models ([Q]SARs), is increasingly required or encouraged, in order to increase the efficiency and effectiveness of the risk assessment process, and to minimise the reliance on animal testing. The main question for the assessor concerns the usefulness of the prediction approach, which can be broken down into the practical applicability of the method and the adequacy of the predictions. A framework for assessing and documenting (Q)SAR models and their predictions has been established at the European and international levels. Exactly how the framework is applied in practice will depend on the provisions of the specific legislation and the context in which the non-testing data are being used. This report describes the current framework for documenting (Q)SAR models and their predictions, and discuses how it might be built upon to provide more detailed guidance on the use of (Q)SAR predictions in regulatory decision making. The proposed framework is illustrated by using selected pesticide active compounds as examples.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LB-NA-24705-EN-C

Publications Office

ISBN 978-92-79-19081-0