RIVM Report 601779001/2007

The EU (Q)SAR Experience Project: reporting formats

Templates for documenting (Q)SAR results under REACH

E. Rorije
E. Hulzebos
B.C. Hakkert

Contact:
Emiel Rorije
Expertise Centre for Substances (SEC)
emiel.rorije@rivm.nl

This investigation has been performed by order and for the account of the Ministry of Housing, Spatial Planning and the Environment (VROM) of the Netherlands, Directorate General for Environment Protection (DGM), Directorate for Chemicals, Waste, Radiation Protection (SAS), within the framework of M/601779 Risk Assessment Tools

RIVM, P.O. Box 1, 3720 BA Bilthoven, the Netherlands Tel +31 30 274 91 11 www.rivm.nl

Rapport in het kort

Het EU (Q)SAR Experience Project: rapportage formats
Sjablonen voor het documenteren van (Q)SAR resultaten voor REACH

De zojuist in werking getreden Europese wetgeving voor chemische stoffen (REACH) propageert alternatieven om het aantal dierproeven te verminderen. (Q)SAR is zo'n alternatief en staat voor kwalitatieve en kwantitatieve structuur-activiteitsrelatie. (Q)SAR's leggen een verband tussen de chemische structuur van de stof en een toxische eigenschap ervan, bijvoorbeeld huidirritatie. Met behulp van deze theoretische modellen is het mogelijk om schadelijke effecten van chemische stoffen voor mens en dier te voorspellen zonder dierproeven te hoeven doen. Het RIVM draagt bij aan het verminderen van het aantal dierproeven door de resultaten van deze modellen, de zogeheten (Q)SAR's, toepasbaar te maken voor beleid.

Het RIVM heeft formats ontwikkeld om de resultaten van (Q)SAR's op drie niveaus te beschrijven: het model, het voorspelde effect van een specifieke stof, en de vertaling van dat effect naar beleid. Deze aanvullende informatie is nodig om de geldigheid en betrouwbaarheid van een voorspelling goed te kunnen beoordelen. De formats blijven in ontwikkeling, maar zijn in concept al opgenomen in de REACH-richtlijnen. Ze kunnen bovendien met enige aanpassingen ingezet worden om resultaten van andere alternatieven voor dierproeven, zoals die genoemd worden in Bijlage XI van de REACH wettekst, transparant te documenteren. Genoemde alternatieven zijn onder andere de "read-across"-aanpak en groepering van stoffen ("category"-aanpak), waarbij de toxiciteit van een nieuwe stof gelijk wordt gesteld aan één of meerdere bekende stoffen .

Bovenstaande activiteiten zijn het resultaat van het Europese (Q)SAR Experience Project, een initiatief uit 2004 dat het RIVM vooruitlopend op REACH heeft opgezet. Europese beleidsmakers en stoffenbeoordelaars doen hierin kennis en ervaring op met (Q)SARs, en geven aanbevelingen voor het gebruik ervan in beleid.

Trefwoorden: (Q)SAR; bewijskracht; groepering van stoffen; "read-across"-aanpak; REACH Bijlage XI.

Abstract

The EU (Q)SAR Experience Project: reporting formats
Templates for documenting (Q)SAR results under REACH

The European chemicals regulation REACH, which just entered into force, strongly advocates the use of alternatives for animal testing. (Q)SAR is such an alternative and stands for Quantitative Structure-Activity Relationship. (Q)SARs try to correlate the chemical structure of a substance to a toxicological property of that substance, for example skin irritation. The use of these theoretical models makes it possible to predict toxic effects of chemical substances without performing animal tests. RIVM contributes to reducing the number of animal tests by making the results of these models, so called (Q)SARs, suitable for regulatory use.

In a European cooperation RIVM developed formats for reporting (Q)SAR results for REACH on three levels: description of the model, the predicted effect for a specific substance, and the interpretation of that effect for regulatory use. This extensive information is needed to be able to judge the validity and reliability of a prediction. These formats continue to develop, but have already been incorporated into the REACH guidance. Furthermore, with simple adaptations they can also serve as reporting formats for other alternatives for animal testing mentioned in Annex XI of the REACH regulation. Examples of such alternatives are the read-across and category approaches, where the unknown toxicity of a new substance is presumed to be equal to one or several similar compounds with known toxic effects.

The above mentioned activities are the result of the European (Q)SAR Experience Project, initiated by RIVM in 2004 as a preparation for new regulations under REACH. In this project European regulators and policy makers increase their knowledge of (Q)SARs, gain experience with existing QSAR models, and give recommendations and guidance on the use of these models.

Key words: (Q)SAR; read-across approach; category approach; Weight-of-Evidence;
REACH Annex XI

Contents

Summary

One of the major outcomes of the activities employed within the EU (Q)SAR Experience Project 2004-2006 is reported. Within the EU (Q)SAR Experience Project the initial goal was to allow EU regulators to gather hands-on and eyes-on experience with Quantitative and Qualitative Structure-Activity Relationships ((Q)SARs) and discuss their experiences. The project should lead to improved knowledge of (Q)SAR methods for regulators, and provide input to the development of guidance on the use of (Q)SARs as foreseen in the European chemicals regulation REACH. Exchange of experiences was established by individually evaluating a number of substances with different models, for different endpoints, and reporting the results in the EU (Q)SAR Working Group meetings. During the project it became clear that there was a need for standardized ways to communicate (Q)SAR model predictions. More importantly it was recognized that information on the model, on the prediction and on the use of the prediction for a specific regulatory purpose needed separate treatment. Therefore a three level reporting approach was proposed by RIVM and consequently developed, discussed and improved upon within the (Q)SAR Experience Project. The discussions on reporting (Q)SAR results, and the information needs as foreseen by regulators in order to be able to assess non-testing data in general, have led to the three (Q)SAR reporting formats presented in this report.

These templates (with slight adjustments) should also be able to prove their value for reporting other non-standard testing results, as mentioned in the Annex XI of the REACH regulation (*in vitro* methods, category approach and read across approach). The reporting formats are still under development, but have already been incorporated in the REACH Implementation Projects (RIP) 3.3 guidance documents (Reach Implementation Project 3.3: Information Requirements).

1 Introduction

In chemicals risk assessment, there are several large-scale regulatory programs such as the OECD HPVC program, the Canadian DSL program, the US EPA/OPPT New Chemicals Program and the European Chemicals legislation (REACH). In certain regulatory settings the use of non-testing data is more established (most notably the OPPT New Chemical Program) than in others. However, it is expected that in the near future, alternatives for in vivo-testing such as *in silico* and *in vitro* methods will become much more frequently used in risk assessment. Both industry (as responsible entities and/or registrants) and regulators will need to deal with the question how the results of these alternative methods should be interpreted, how these results are reported and how they can be evaluated (and weighted).

Within the (Q)SAR Experience Project (see paragraph 1.2), reporting formats were suggested to exchange experience between regulators on the use and interpretation of (Q)SAR models in risk assessment. During the project it became clear that reporting on the use and outcome of alternative methods should be placed in a wider context. If the results of alternative methods are not reported consistently, it will become very difficult for regulators to evaluate if the methods used are valid for a specific risk assessment purpose, if they have been applied correctly and if they have been interpreted correctly. Therefore, it was felt as a joint interest for both industry and regulatory bodies to develop a system for reporting alternative methods, including (Q)SARs, such that they can be easily interpreted and evaluated in the risk assessment procedure. Such a system would also provide the means to properly and uniformly document the results of applying QSAR models. The reporting scheme consists of three separate reporting levels:

1) the QSAR Model Reporting Format (QMRF), describing the model in general;
2) the QSAR Prediction Reporting Format (QPRF), describing in detail the prediction for one specific substance, and
3) the Weight-of-Evidence Reporting Format (WERF), summarizing and weighting all data for a specific regulatory endpoint, and drawing a conclusion.

1.1 Scope of this report

This report is meant to report the outcome of the activities employed within the (Q)SAR Experience Project 2004-2006; namely the development of a system of (Q)SAR reporting formats. The emphasis of the report is on the presentation of the currently proposed reporting formats, not so much on the specific activities within the (Q)SAR Experience Project, i.e. the process of gaining experience with the application of (Q)SARs and subsequently reporting (Q)SAR results. A more detailed report of the activities in the first phase of the project is given in Appendix 7 of this report. It should be noted that the whole process of gaining experience with and applying (Q)SARs to existing substances, the need to report the results of the experience exercises, and discussing these results in the EU (Q)SAR Working Group were the immediate cause for the creation of a system of (Q)SAR Reporting Formats. Furthermore these activities have been essential in the discussion on what information would be (at the very least) necessary for regulators in order to be able to judge a (Q)SAR result, especially when a specific model is not available to the regulator. This situation is to be expected in the near future under REACH, where all available information, including non testing information, will have to be used to fulfil the data requirements.

1.2 Background – the (Q)SAR Experience Project

In 2004, a so-called (Q)SAR experience project was initiated by RIVM, the Netherlands. The activities of this project were conducted under the umbrella of the EU QSAR Working Group. The intention of the project was to provide EU regulatory authorities with hands-on experience in the generation of (Q)SAR predictions for chemical substances, and (eyes-on) experience with the evaluation of (Q)SAR generated predictions, as such predictions were foreseen to be part of a substance dossier under REACH.

Following a meeting held in Den Dolder (NL) on 26-27 October, 2004, a project proposal was developed jointly by the European Chemicals Bureau (ECB), RIVM and the Danish EPA. Phase 1 of this (Q)SAR Experience project was carried out in 2005. This first phase basically consisted of an extension of the Danish (Q)SAR effort for 184 SIDS substances, using other models than those used by Denmark. These predictions were then compared to experimental values included in the SIDS database, as well as estimates generated by the Danish EPA in an OECD-related (Q)SAR activity. This exercise was thought to provide a feeling for the quality and performance of the models applied. Furthermore, the same (Q)SARs were applied to small sets of ten substances where the outcome was discussed in detail, in order to create awareness of issues that can play a role in the prediction of a specific substance. This exercise was performed for three endpoints: biodegradation, mutagenicity and (acute) fish toxicity

The results of this first phase of the (Q)SAR experience project have been reported separately and can be found in Appendix 7.

The exercise generated a large number of valuable take-home messages, eye-openers and discussions on the adequacy of specific models for specific endpoints. It was concluded in a follow up meeting that a lot of the discussion and differences in perceived usefulness of the (Q)SAR in this first phase came from the different levels of interpretation that was used or assumed in presenting the (Q)SAR Results.

A one day meeting with a smaller focus group consisting of ECB, RIVM, Danish EPA, and Health Canada was held in Ispra, IT in January 2006 where the concept of a three level approach to reporting (Q)SAR information was brought forward by RIVM, worked out in the group and presented by RIVM to the EU (Q)SAR Working Group the next day. It was agreed at that EU (Q)SAR WG that a number of real life examples would be worked out in this three level reporting approach, and distributed to the group for discussion at the next WG meeting in October 2006.

Four different substances and three endpoints were selected: Biodegradation (dibenzyltoluene), Skin Irritation (4,4'-methylenebis(2,6-dimethylphenyl isocyanate and 4,4'-diisobutylethylidenediphenol) and Skin Sensitization (cinnamaldehyde) were chosen and worked out by RIVM in cooperation with ECB (sensitization) and INERIS France (Biodegradation) in advance of the October EU (Q)SAR WG meeting. The examples and the problems and inconveniencies encountered were discussed at the EU (Q)SAR WG meeting in October 2006, and a report of the discussion is given in Appendix 1. The worked out examples as they were distributed before the meeting are given in Appendix 5.

The discussion led to a number of improvements and adaptation of the reporting formats. The model format for the general properties of a (Q)SAR model (the (Q)SAR Model Reporting Format, or QMRF) as worked out by ECB - and now forming the basis for the ECB Inventory of (Q)SAR Models [http://ecb.jrc.it/(Q)SAR/(Q)SAR-tools/(Q)SAR_tools_qrf.php] – was already put forward on the internet in the second half of 2006 for beta testing by users contributing (Q)SAR models to the ECB inventory. In this format, among others, the OECD principles on validation of (Q)SARs are addressed. The points raised in the WG discussion (see Appendix 1) were taken into account together with the feedback from the internet beta testing and used for adaptation of the format, leading to a definitive version which is now available from the ECB website.

The format for reporting a (Q)SAR result for a specific substance, the (Q)SAR Prediction Reporting Format or QPRF, was worked out by RIVM, and taking into account the points brought forward in the October WG meeting, a definitive version of the format is now put forward by ECB for formal beta testing via the internet, similar to the beta-testing exercise performed by ECB in the second half of 2006 for the QMRFs. The latest format of the QPRF (2007) is given in Appendix 3, both in empty form and with a help text on what information is expected in the various fields.

During the discussion at the WG meeting in October there was at that moment no agreement on the necessity of a Weight-of-Evidence Reporting Format, where the information from different predictions would be summarized and combined in order to come to a conclusion for a specific endpoint within a regulatory framework (i.e. classification and labelling). It was concluded that the discussion on the necessity and information content of the WERF would be postponed to a later stage, when more experience with the QMRF and QPRF has been gained. It is foreseen that especially in view of the need for integrated assessment of substances under REACH such a WERF is needed, again in order to fulfill the need for robust documentation (as stated in REACH Annex XI) of the information used to come to a conclusion. Therefore in Appendix 4 a sample format (the Weight-of-Evidence Reporting Format, or WERF) for summarizing all information within a Weight-of-Evidence approach relevant to a specific endpoint and a specific regulatory framework is proposed.

2 Reporting formats

The development of these formats started in the context of the (Q)SAR Experience Project, coordinated by RIVM(NL), which was subsequently subsumed into the activities of the (Q)SAR Working Group. The (Q)SAR reporting formats that developed out of the (Q)SAR Experience Project activities will be described in the following paragraphs. These formats are included in the RIP 3.3 cross-cutting guidance on (Q)SARs.

2.1 The need for transparent documentation on (Q)SARs

The need to have a formalized platform to exchange (Q)SAR results became apparent during the (Q)SAR Experience Project exercises. Furthermore, according to Annex XI of the REACH regulation, one of the conditions for using (Q)SARs instead of test data - but also for other non-testing data and non standardized test data - is that ‘adequate and reliable documentation of the applied method is provided’. The text of REACH Annex XI – General Rules for Adaptation of the Standard Testing Regime, requires that:

- the scientific validity of the model has been established;
- the substance falls within the applicability domain of the model;
- results are adequate for the purpose of classification and labelling and/or risk assessment, and
- adequate and reliable documentation of the applied method is provided.

The Agency in collaboration with the commission, Member States and interested parties shall develop and provide guidance in assessing which (Q)SARs will meet these conditions and provide examples.

At present, an extensive summary of ‘adequately and reliably’ documented (Q)SARs is not available. Therefore, the ECB in consultation with the EU (Q)SAR Working Group, has started building an inventory of evaluated (Q)SARs, which should help to identify (Q)SAR models suitable for the regulatory purposes of REACH. This inventory will be made freely available from the ECB website (http://www.ecb.jrc.it/(Q)SAR). The presence of a QSAR in this inventory does not implicate a recommendation of this model over any other available methods, but should provide the proper documentation to make it possible to assess the scientific validity of a model. In the wider international context, the content of the ECB Inventory could also be used in the (Q)SAR Application Toolbox, a project currently being led by the OECD. The (Q)SAR Application Toolbox is intended to be a set of tools supporting the use of (Q)SAR models in different regulatory frameworks by providing estimates for commonly used endpoints together with guidance on the interpretation of estimated data.

The requirement for adequate and reliable documentation of (Q)SARs has led to discussions on what information is required for (Q)SARs and how this information should be structured. Reporting requirements for non-testing methods should not limit the use of (Q)SAR approaches or impose what methods should be used – they are meant to provide all relevant information so that informed choices can be made regarding the use of (Q)SARs. The ECB Inventory of (Q)SAR models should make sure that the same information (on the model description level) is available to Industry registrants, the MS authorities, and the European Chemicals Agency.

It was however felt by the (Q)SAR Experience participants that a (Q)SAR model inventory would only supply ‘adequate and reliable documentation’ for a part of the information required by legislators to interpret a (Q)SAR result. In the terminology of Annex XI this would describe the ‘scientific validity status' of the model. A scientifically valid model can however be applied to substances that it was never intended to be used for. Information on the "applicability" of the model to the specific substance of interest is therefore also required. Furthermore the assessment of how "adequate" a certain model or model prediction is for the legislative purpose (i.e. risk assessment or classification and labelling) should also be performed, and documented, in addition to the previous issues of scientific validity status and the model applicability. The REACH text naturally led to three different reporting levels, for which three different reporting formats were subsequently proposed by RIVM. The idea of reporting the required information to comply with Annex XI of the REACH legislation on three different levels is summarized on the poster presented in Appendix 6 of this report.

2.2 Three levels of reporting (Q)SAR Information

The reporting formats as proposed during the (Q)SAR Experience meeting, January 2006, have three levels:

(Q)SAR Model Reporting Format (QMRF)

Description of a specific method or model, based on the OECD criteria for validation of (Q)SAR models (but not necessarily limited to (Q)SAR models, this could also include *in vitro* methods, read across or category approaches). This format should serve among others as documentation of the scientific validity status of the model.

(Q)SAR Prediction Reporting Format (QPRF)

Reporting the prediction and conclusion for a specific substance and endpoint, for one single method or model. This format should yield proper documentation of the model prediction, but more importantly should address all issues related to the applicability of the model to the specific substance of interest.

Weight-of-Evidence (WoE) Reporting Format (WERF)

Summary of the results of all alternative methods (potentially also including experimental results), leading to a final conclusion for one specific endpoint within a regulatory framework, based on the combined body of evidence. (e.g. all data and predictions on bioaccumulation properties of substance X – evaluated for use within the EU PBT assessment). In this format the final assessment of the adequacy of the model prediction for a specific (regulatory) purpose is reported.

The Reporting Formats should be regarded as a communication tool to enable an efficient and transparent exchange of (Q)SAR information between Industry and MS authorities. Ideally, these reports would be attached to the registration dossier. The (Q)SAR Working Group has extensively discussed the need for, and the content of, the three types of QRF. In its third meeting, October 12-13, 2006, the consensus of the Working Group was that well-defined formats are needed to describe models (QMRFs) and individual model predictions (QPRFs). The necessity and usefulness of a defined format to summarize the overall assessment (WERF) was questioned by some participants. It was argued that the documentation of the overall assessment might require more flexibility than can be easily accommodated in a fixed reporting format. Others stated that especially under REACH where the integrated use of all kinds of existing data, including non-testing data, is requested, a clear format reporting the validity of each study and prediction is needed. A format providing the necessary issues to be discussed in the evaluation and summarizing the conclusions drawn from such diverse data would then be a prerequisite. Therefore, the need for such a format, and its general structure, should be reconsidered when more experience is gained in the regulatory use of (Q)SARs and their documentation by means of QMRFs and QPRFs.

During the development and discussion of the reporting formats it was frequently remarked that an identical strategy of documenting results can and should also be applied to other alternative data like i.e. read across, category approaches, *in vitro* test results, or non GLP and/or non-guideline test results.

The reporting formats will be discussed separately in the following paragraphs. The QMRF, QPRF and WERF can be found in Appendices 2, 3 and 4, and filled out examples of both QPRF and WERF can be found in Appendix 5. The latest version of the QMRF and a collection of submitted QMRFs (The ECB Inventory of (Q)SAR models) can also be found at the ECB website, http://ecb.jrc.it/(Q)SAR/(Q)SAR-tools/(Q)SAR_tools_qrf.php.

2.3 The (Q)SAR Model Reporting Format (QMRF)

The QMRF provides the framework for compiling robust summaries of (Q)SAR models and their corresponding validation studies. It should describe the model performance in general, its predictive performance (from internal as well as external validation studies) and compliance with (OECD) guidelines for the validation of (Q)SAR models. The structure of this format has been designed to include all essential documentation that can be used to evaluate the concordance of the (Q)SAR model with the OECD principles. ECB started compiling an inventory of QMRFs to gain experience with this specific Reporting Format. The (Q)SAR Experience exercise as performed in 2006, and discussed in the October 2006 meeting, provided the necessary hands-on experience with some worked out examples of QMRFs. By requiring the participants to give an indication of the reliability of a prediction for a single substance (in the (Q)SAR Prediction Reporting Format, see next paragraph), they were forced to use the information as given in the QMRFs, and subsequently evaluate whether this information is sufficient to draw a conclusion on the reliability of the outcome of the model for that specific example. This lead to a good discussion of the necessary information needed in the QMRF based on actual application of the Format to specific examples.

By centrally compiling an inventory of ((Q)SAR) model descriptions, it will suffice (in the future) to refer to this central inventory, instead of having to report on the method used for every dossier entry. In this way it is analogous to what the inventory of OECD Testing Guidelines is for experimental results.

The QMRF will contain the general descriptive information of the model, using the following nine headings, with sub-questions:

1.	(Q)SAR identifier	
2.	General Information	
3.	Defining the endpoint	– OECD Principle 1
4.	Defining the algorithm	– OECD Principle 2
5.	Defining the applicability domain	– OECD Principle 3
6.	Defining goodness-of-fit and robustness	– OECD Principle 4
7.	Defining predictivity	– OECD Principle 4
8.	Providing a mechanistic interpretation	– OECD Principle 5
9.	Miscellaneous information	
10.	Summary for the ECB Inventory	

Furthermore the QMRF should provide all information on the validity status of the needed to fulfil the five principles as drawn up in the OECD guidance on the validation of (Q)SARs. The five main issues of the OECD guidance on (Q)SAR validation are:

1) Defined Endpoint. The intent of principle 1 (a (Q)SAR should be associated with a defined endpoint) is to ensure clarity 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. It is therefore important to identify the experimental system that is being modelled by the (Q)SAR.
2) Unambiguous Algorithm. Principle 2 (a (Q)SAR should be associated with an unambiguous algorithm) should give transparency in the model algorithm used to generate the predictions. It should address the issue of reproducibility of the predictions.
3) Defined Applicability Domain. A (Q)SAR should be associated with a defined domain of applicability. This can be defined in terms of chemical and/or physico-chemical domain (descriptor space) and/or in terms of the response (biological domain).
4) Statistical Validation. The fourth principle (a model should be associated with appropriate measures of goodness-of-fit, robustness and predictivity) expresses the need to perform statistical validation to establish the performance of the model.
5) Mechanistic Interpretation. According to principle 5, a (Q)SAR should be associated with a mechanistic interpretation, if possible. This should give a physicochemical/ chemical/ biological meaning to the descriptors after the modelling.

More information can be found in the OECD guidance on the validation of (Q)SARs.

The July 2007 version (version 1.2) of the QMRF as developed and beta-tested by ECB is given in Appendix 2. For all the headings and sub questions explanation is provided on the information expected there. For further discussion of all the issues for which information is required on the QMRF level the reader is addressed to the ECB website, http://ecb.jrc.it/(Q)SAR/(Q)SAR-tools/(Q)SAR_tools_qrf.php, where sample QMRFs and guidance on how to fill out the form, and a number of worked out examples can be found.

2.4 The (Q)SAR Prediction Reporting Format (QPRF)

This is the level for reporting a prediction of an individual model or the result of a particular method, for one specific substance. The format states the basic information coming from the model, and an indication of the reliability for that specific prediction. This reliability is influenced by the general predictive performance of a model (as reported in the QMRF) but should on this level be reported in terms of the specific reliability of the model for the specific substance of interest. Even if a model has a high predictive performance in general (e.g. high r2, good results in external validation etc.), a prediction for a certain substance can still be highly questionable, for example because of domain of applicability issues.

Information on where the substance is in the domain of applicability of the model is therefore the most important part of the QPRF, apart from the actual result of the model, and any other factors that might influence the reliability of the prediction for one specific substance.

A proposed scheme for ranking the reliability of the predictions, analogous to the Klimisch approach [Klimisch, 1997], has been discussed by the (Q)SAR Working Group. Klimisch codes are given to rank the quality of a given data point, using the numbers 1, 2, 3 and 4. These indicate roughly the following:

1: reliable without restrictions
2: reliable with restrictions (it should be indicated what restrictrions)
3: unreliable
4: not assignable

This reliability coding was applied in the (Q)SAR Experience exercise on the QPRF level, i.e. for each separate prediction an individual reliability judgment was requested / generated. The view of a part of this group was that such a scheme could be misleading for non-testing data mainly because QSAR models might not always cover the complete endpoint of interest for the regulatory purpose. Non-testing data is generally used *in combination* with other information in a Weight-of-Evidence approach (WoE approach), and successive parts of the toxicological endpoint of interest (e.g. uptake and potential effect) can be estimated by separate models. A stand-alone, absolute, reliability ranking on the level of the QPRF, was therefore suggested to be removed. However, when all relevant data (including experimental data, *in vitro* data, or e.g. information from human exposure) is brought together in the WERF, it becomes feasible to apply a reliability ranking of the overall results, relative to each other, and taking into account the evidence from different sources. This weighting is however not identical to a Klimisch code. The reliability of a single prediction (or experimental result) will as yet not be determined individually (at the QPRF level) by means of a Klimisch code, but only when taking into account all other relevant data (i.e. at the WERF level). It should be noted however that a clear (textual) reasoning on the quality of the prediction should be provided at the QPRF level.

A short description of the headings as used in the RIVM proposed QPRF is given in the following:

GENERAL

Prediction for Substance. Should identify the substance for which the prediction is done. Name, CAS-nr, structure and possible descriptor data used as input in the model should be provided.

Model Name, Version and date of prediction. In this place the model that was used to generate the prediction should be identified as unambiguously as possible.

(Q)SAR Model Reporting Format (QMRF). Refer to an entry in the ECB Inventory of (Q)SARs (http://ecb.jrc.it/(Q)SAR/(Q)SAR_tools/(Q)SAR_tools_qrf.php) whenever possible. Otherwise refer to the QMRF document with the general model description, .accompanying this QPRF

Endpoint description. Here a description of the exact endpoint that the model is predicting/reproducing should be given. This is not necessarily identical to a regulatory relevant endpoint! The assessment of how adequate the model prediction is for the specific regulatory endpoint of interest is not done here, but should be performed in the Weight-of-Evidence Reporting Format (WERF). Information describing the endpoint used for the model can be taken from the QMRF.

PREDICTION

Model outcome. The exact (raw) outcome as produced by the model, before interpretation, is reported here. For example a value like 0.98 when dealing with a quantitative model, or the identification of a number of substructures identified in the substance.

INFORMATION RELEVANT for the ASSESSMENT of the PREDICTION

Model Algorithm/Result interpretation. This should give an explanation of the algorithm used by the model, and the interpretation that needs to be applied to the model outcome.

Is the substance within the Domain of Applicability of the model? Sub questions that (preferably) should be answered here:

1. Is the chemical of interest within the scope of the model, according to the defined applicability domain of the model?
   a) Descriptor domain: do the descriptor values of the chemical fall within defined ranges?
   b) Structural fragment domain: does the chemical contain fragments that are not represented in the model training set?
   c) Mechanistic domain: does the chemical of interest act according to the same mode or mechanism of action as other chemicals for which the model is applicable?
   d) Metabolic domain: does the chemical of interest undergo transformation or metabolism, and how does this affect reliance on the prediction for the parent compound?
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?

Alerts/fragments identified and/or rules applicable to the substance? Identify which part(s) of the structure contribute to (the interpretation of) the model result. Mention applicable rules (i.e. on skin penetration) that fortify or disqualify the model outcome.

Indicate structural analogues identified by the model? Mention analogues (and their experimental data) identified by the model, and/or substances from the model training set that are close structural analogues (This step is not supposed to replace the extensive structural analogue search outside of the model training set data which is proposed in the Stepwise approach to the use of non-testing data, Cross-cutting guidance on the use of (Q)SARs). It should give a feeling of how appropriate the model is for predicting the substance of interest, by indicating the structurally closest substances that were part of the training set of the model.

Is the substance part of training set? Yes/no, and if yes, indicate the experimental value used in the training set for this substance

Other information regarding prediction reliability? Indicate all factors not discussed above that influence the reliability of this specific prediction. As a minimum the list of prediction specific issues as identified under “Miscellaneous information” in the QMRF should be addressed here.

CONCLUSION

Result. This is the place to report the interpreted model prediction, where the prediction value is translated into its meaning for the toxicological endpoint. (for example: a Biowin5 prediction of 0.98 is described in its interpreted form: Readily Biodegradable in OECD301C, modified MITI-I test)

Reasoning. This should give the reasoning on reliability of the result, summarizing all factors influencing reliability for specific prediction as discussed above. No qualification is expected that concludes that the prediction can of cannot be used. That conclusion is drawn on the next level (WERF). The (interpreted) result and the rationale will subsequently be used in the WERF (see next paragraph) to come to a regulatory conclusion (which is not necessarily equal to the toxicological result that is reported in the QPRF).

2.5 The Weight-of-Evidence Reporting Format (WERF)

The Weight-of-Evidence Reporting Format) is the top level reporting format that provides essential information and conclusions for one specific substance, endpoint and regulatory framework. The conclusions part from the QPRF describing the application of the underlying method or model are extracted, so the reasoning on the reliability of different results (when multiple model results are available) can be compared, and the application of a Weight-of-Evidence to the results becomes transparent. As some of the information is dependent on the regulatory framework in question this Reporting Format should also address relevant cut-off criteria, screening criteria, thresholds, classification and labelling issues. These (cut-off criteria, thresholds etc.) might be different for different regulatory needs (i.e. the Bioaccumulation criterion for PBT assessment is much higher (BCF > 2000 l/kg) than for C&L of a substance with R53; potential long term effects to the environment (BCF > 100 l/kg).

Below the headings / issues addressed in the RIVM proposed WERF (also see the sample WERF format in Appendix 4 and the example WERF for specific substances in Appendix 5) will be discussed shortly to give an idea of the information needed for evaluation of a toxicological endpoint at the WERF level:

SUBSTANCE

WERF for substance: The name of the substance, and/or other identifiers like CAS or EINECS number, should be given here. The identifier should be the same as used in the QPRF.

ENDPOINT

Regulatory endpoint. A description of the regulatory framework for which the conclusion will be used is given here. One could imagine data being judged differently in different regulatory frameworks (different threshold values for example). Also the required reliability of a specific prediction or test outcome can be different for different regulatory frameworks.

DATA – (Q)SARs, category approach, *in vivo* and *in vitro* test data

(Q)SAR Model name. Identify the model for which a result is reported. This should be equal throughout the levels (QMRF, QPRF and WERF).

Result. Here the result (as presented in the QPRF) for a specific model prediction is reported.

Reasoning. Here the reasoning from the QPRF for a specific model prediction should be presented, dealing with the reliability of the result. At this point also the interpretation (meaning) of the model result for the specific regulatory endpoint under evaluation should be included.

The same information (name of test, result, and reasoning on the reliability of the test result) can be presented repeatedly, for different ((Q)SAR) models, category approach/read across results, and additionally for any *in vitro* and *in vivo* test results. See the example of an empty format in the Appendix 4 for a suggestion on how to present data from various models. Finally the WERF contains a conclusion taking into account all the presented data:

CONCLUSION

Weighted summary of the presented data - Result. Present here a conclusion for the specific regulatory endpoint under evaluation. For example: for a EU PBT assessment this substance is thought to be persistent in the environment, taking into account all degradation data and model predictions.

Weighted summary of the presented data - Reasoning. Here the reasoning for the result should be given, indicating how the presented data is weighted and/or why a specific data point is preferred or dismissed from the evaluation. Whether this is done quantitatively with a numeric ranking, or weighting, or qualitatively using only text is open for discussion. A numeric ranking can be complicated since that ranking will change when a new, additional data point is introduced. Therefore a textual reasoning why one data point is thought more influential is thought sufficient and more practical at this moment.

Need for further testing? Here a test proposal can be indicated. Based on the evaluated evidence the conclusion could be that no conclusion can (yet) be reached. When more information is needed, here suggestions for additional data can be introduced. Either more ((Q)SAR) model data, or specific input into a model (e.g. physico-chemical data) could be sufficient. Or a need for (further) testing, either *in vitro* or *in vivo* can be identified.

The Weight-of-Evidence Reporting Format is thus not a (Q)SAR specific reporting format, but is suggested as a means to transparently report all the relevant data (including experimental data, *in vitro* data, category approach and (Q)SAR data) used in the Weight-of-Evidence approach.

Whether the WERF should have the form of a (predefined) format or could be a free text was also subject of discussion in the (Q)SAR Experience Project (see Appendix 1). The main message from this three level approach is that the QPRF (and the QMRF) should provide adequate information on the reliability of the prediction, to be able to do a Weight-of-Evidence analysis where all relevant data is taken into account. To that end the WERF format served an important role in the (Q)SAR Experience project, as working with examples and trying to reach a conclusion (in the WERF) turned out the only meaningful way to discuss the adequacy of the information provided for a specific (Q)SAR prediction (in the QPRF and QMRF formats).

3 Conclusions and further activities

The (Q)SAR Experience Project as initiated in 2004 by RIVM has led to the development of a transparent system for the reporting of (Q)SAR results. This project was aimed at gaining hands-on and eyes-on (Q)SAR experience for regulators, and should provide input to the development of guidance on the use of (Q)SARs in a later stage. It was due to the actual hands-on experience exercises performed within this project that the need for more elaborate reporting schemes was identified. In the first exercises a lot of the discussion was due to differences in interpretation of a result, because people did not recognize the different levels of information and interpretation that were present. Some examples: A (Q)SAR model can be very robust and valid, but the prediction for substance X is not valid, as it lies outside of the applicability domain of the model. On another level, a prediction can be perfectly valid, but what the model predicts (the endpoint) is not valid for use in a specific regulatory endpoint. After the identification of the need to report results on different levels, and therefore develop separate reporting formats for these different levels, it also proved to be highly beneficial for the discussion of the contents of these formats to actually apply them to real life substances ((Q)SAR experience). The development of this system of reporting (Q)SAR results is not over yet.

Discussion on the format of the QMRF as developed by the ECB is now finalized after an extensive beta testing and evaluation. The contents of the format have been agreed upon now. The attention is now turned to the actual filling of the ECB Inventory of (Q)SAR Models by generating and reviewing QMRFs. This is not intended to become a formal endorsement of validity or acceptance of a particular model but simply a quality check of harmonized documentation of (some) of the models. This will most probably also serve as repository of (Q)SAR models for implementation into the OECD (Q)SAR Application Toolbox.

The QPRF will be put forward for further beta testing and evaluation through the ECB website [http://ecb.jrc.it/(Q)SAR/(Q)SAR-tools/(Q)SAR_tools_qrf.php]. This procedure should be similar to the procedure followed for the QMRF in the second half of 2006. In the second half of 2007 a definitive version of the QPRF should then be agreed upon.

The WERF concept is not developed further at this moment. RIVM is of the opinion that some form of WERF (either as a (more restrictive) format, or as a free text discussion that should address at least certain issues) is desirable, and necessary in the light of Integrated Testing Strategies and the Weight-of-Evidence approach which is suggested in the RIP documents. Such a Weight-of-Evidence approach of course does not have to be limited to (Q)SAR results alone, and should ideally take into account all evidence available, i.e. other non-testing data, *in vitro* testing data, non-guideline testing data and guideline test results. First there should be a further discussion on desirability / workability of the WERF within the EU (Q)SAR WG.

A number of follow up actions related to (Q)SAR reporting formats can be identified:

Within the OECD (Q)SAR Toolbox initiative the reporting formats (QPRF and QMRF) have been indicated as determining which information should be provided by the Toolbox to generate acceptable (Q)SAR or read-across predictions. Ultimately the Toolbox is also foreseen to directly generate QMRFs and/or QPRFs as needed.

Within IUCLID 5 the possibility is presented to use predetermined (text) templates which guide the user on the information required/expected when data is entered in specific sections. The QPRF specifically could be easily converted into a IUCLID5 template for reporting (Q)SAR results. Whether this will be implemented in the IUCLID5 version for roll-out in 2007, or whether these templates have to be inserted by the user is not yet determined.

Development of similar reporting formats for read-across/category approaches is currently initiated, and the (Q)SAR Reporting Formats are serving as starting points for this development.

By actually working with the formats, and applying them in order to create an advice, they will be further refined. A possible next step for the EU (Q)SAR WG is to start using the formats in the generation of (Q)SAR advice for substances to be discussed in EU TCNES, EU C&L and EU PBT Working Groups.

The reporting formats are meant as guidance on how to adequately report results obtained from qualitative of quantitative structure-activity relationship models ((Q)SARs), by providing a format or template with headings that indicate the issues that should be addressed. The actual form of such a format or template (i.e. a word document or an excel worksheet, or whether it should be a table format or a more free text type of report) is not meant to be compulsory prescribed by these formats. The should rather be seen as check-lists with issues that need to be discussed and documented, in order to make sure that QSAR predictions can be accepted as evidence under REACH. In this way QSAR model data are no different than experimental data, where e.g. a robust study summary is also expected to address a number of issues in order to make the result acceptable in the regulatory process.

References

Klimisch HJ, Andreae M, Tillmann U (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regulatory Toxicology and Pharmacolygy* **25**, 1-5.

OECD (2004). The report from the expert group on (quantitative) structure-activity relationships ((Q)SARs) on the principles for the validation of (Q)SARs. OECD Environment Health and Safety Publications. Series on Testing and Assessment No. 49. ENV/JM/MONO(2004)24. Organisation for Economic Cooperation and Development, Environment Directorate, Paris, France. http://appli1.oecd.org/olis/2004doc.nsf/43bb6130e5e86e5fc12569fa005d004c/e4553fbf1c1fdf7bc1256f6d003ab596/$FILE/JT00176183.PDF

REACH (2006). Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006, concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Agency, amending Directive 1999/45/EC and Repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93.67/EEC, 93/105/EC and 2000/21/EC.

Technical Guidance Document to Industry on the Information Requirements for REACH (2007) Part 1 (of 4) General Issues. Final draft for review by RIP 3.3-2 PMG members only – not for wider distribution. Chapter 6 Other approaches for evaluating intrinsic properties of chemicals, section 6.1 Guidance on (Q)SARs, pages 72-143. http://ecb.jrc.it/documents/REACH/RIP_FINAL_REPORTS/RIP_3.3_INFO_REQUIREMENTS/FINAL_DRAFT_GUIDANCE/RIP3.3_TGD_FINAL_2007-05-02_Part1.pdf

Worth AP, Bassan A, Gallegos A, Netzeva TI, Patlewicz G, Pavan M, Tsakovska I and Vracko M (2005). The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance. JRC Report EUR 21866 EN. European Chemicals Bureau, Joint Research Centre, Europena Commission, Ispra, Italy. http://ecb.jrc.it/DOCUMENTS/QSAR/QSAR_characterisation_EUR_21866_EN.pdf

Appendix 1. Minutes of (Q)SAR Experience Project

Discussion on Reporting Formats

EU TCNES/(Q)SAR Working Group, Varese, October 12, 2006

In the 2nd TCNES / (Q)SAR WG meeting (January 2006) RIVM proposed a three level approach for (Q)SAR reporting formats ((Q)SAR Model Reporting Format -QMRF, (Q)SAR Prediction Reporting Format - QPRF and Weight-of-Evidence Reporting Format - WERF). A month in advance of the 3rd meeting four worked out examples (2x skin irritation, 1x biodegradation, 1x skin sensitization) were distributed to the participants by RIVM in order to stimulate discussion on the (Q)SAR reporting formats. This discussion is used to determine whether the information provided in the formats (QPRF and QMRF) is sufficient to draw a conclusion that is considered relevant for regulatory decision making (as indicated in WERF). An in depth discussion or consensus on the (regulatory) conclusions (WERF) for the individual substances was NOT the goal of the exercise.

Four central questions were distributed before, and consequently discussed at the meeting:

1. Is the information in the (Q)SAR Model Reporting Format (QMRF) sufficient to assess the reliability of a specific prediction in the QPRF? Would a future ECB Inventory of QMRFs be sufficient in order for a registrant to simply refer to that inventory when using a method? This feed-back would be helpful for ECB on its beta-testing of the QMRFs.
2. Is the information in the QPRF sufficient to judge the reliability/usefulness of a given prediction for its use in a regulatory setting as indicated in the WERF.
3. Is the Weight-of-Evidence Reporting Format sufficient to provide an assessment for a given regulatory purpose? If not, please try to specify what is lacking. Is the information provided in the QMRF and QPRF informative enough to serve more than one regulatory purpose (C&L, RA, priority setting, within different regulatory frameworks)?
4. Is the use of reliability codes (in WERF and QPRF) useful? The definition of and need for reliability codes is open for discussion.

During the meeting of the EU TCNES / (Q)SAR WG a whole afternoon of discussion was chaired by RIVM/the Netherlands on the topic of (Q)SAR Reporting Formats. The following reflects the points taken from both the discussion as well as written comments received in advance of the meeting.

Discussion Point 1 – (Q)SAR Model Reporting Format (QMRF) and ECB Inventory of (Q)SARs

This format was attached to a draft version of an OECD document on validation of (Q)SARs (ref.). They also form the basis of the ECB Inventory of (Q)SAR models, and can be downloaded form the ECB website in excel format. (http://ecb.jrc.it/(Q)SAR/(Q)SAR_tools/(Q)SAR_tools_qrf.php). This format should report the general description of the model, its basis, the statistics, including information on (possible) external validation exercises.

Comments received during the meeting:

- A number of participants were of opinion that the model reporting format could reflect and follow the OECD principles on the validation of (Q)SARs more. If the format follows the guidelines more closely it will be easier to conclude whether or not the model complies to the guidelines. It was explained that the development of the Model Reporting Format had already started before the adoption of the OECD principles. An alternative QMRFormat was proposed by DK, as a suggestion how an adapted QMRF could reflect the OECD principles more accurately.

- There are more issue relevant for reporting QMRF than OECD principles (i.e. version number, specific issue to be addressed for the model, etc).
- The headings specifying which model is referred to should explicitly ask for a version number of the model used, and which computer program (and/or version) was used in which the model was incorporated. More specifically a separate field indicating if a model is part of some kind of software was indicated to be very practical. This should subsequently also contain information on the name of the software, its status (commercial, free, …) and where it can be obtained.
- It would be helpful to have an additional issue/point at the end of the format where specific issues that are considered important/relevant for the evaluation of a specific model prediction are indicated, and which should be included in the QPRFormat for a specific model. The BIOWIN specific questions that were raised in the biodegradation example (the BIOWIN QPRF) can serve as an example. Instead of incorporating the (model specific) question “What is frequency of appearance of the identified fragments in the training set” in the QPRF, this question could be indicated in the QMRF as one of the prediction specific issues that need to be addressed in the QPRF under the heading of “other information regarding the prediction reliability”.
- Some felt it would be useful to have some indications in the end on the total reliability of the model, whereas others were of the opinion that this is context depend, and both the context as well as the information in the QMRF provide information for such an context dependent assessment.
- Some participants felt that more detailed information on the applicability domain would be helpful, i.e. distribution of the training set data (and possibly also validation set data) over the parameter domain, whereas others indicated this is often not available and will therefore not always be reported.
- More information on the test protocol used to generate the training set data would be important in helping to assess the quality and the applicability of the model. The model endpoint should be described in as much detail as possible, including information on the test protocols used for the experimental data that is being modelled. It was recognized that this information will not always be available in such detail.
- The format could provide information on the availability of the training (and possibly validation-) set data, possibly indicating where the raw data can be retrieved.

A separate discussion was conducted on the questions of who is going to fill the Inventory of (Q)SARS and whether there will be an evaluation / quality assurance of the formats that are entered in the Inventory.

- OECD mentions that it will discuss this issue at the next OECD steering group meeting of the ad hoc (Q)SAR WG. It is possibly a task of the steering group to evaluate the filled in QMRFs. This does not mean validation, but a check on completeness and adequacy of the filled in QMRFs before they are entered into an Inventory.
- A remark is made that probably QMRFs will not only be sent to OECD, but also to ECB and/or ECA in the future. How will these (parallel) streams of information be harmonized / evaluated?
- It is suggested that it might become a kind of core task of this group (the current TCNES / (Q)SAR WG) to assess and comment on the QMRFs that are being submitted (to ECB/ECA) and that it could become a returning point on the agenda to agree on new model descriptions being added to the inventory.

In general it was concluded that the model format is considered useful and serves its purpose - the general description of the model -, but there is also still room for some improvements.

Discussion point 2 – the (Q)SAR Prediction Reporting Format (QPRF)

A (very early) draft of the QPRF was included in the (draft) document on cross-cutting guidance on the use of (Q)SARs, prepared by ECB for use in the RIP3.3. EWGs. This format should report the result of a prediction by a specific model for one specific substance. Attention should be given in this format to those issues that influence the reliability of a prediction for this specific substance.

Comments received during the meeting:

- Headings in the QPRF should be more defined. They should "force" the user more to give the desired answer to the question. For example the heading "Domain of Applicability" should read "Is the substance within the domain of applicability of this model?". The heading "Structural Analogues" should read: "Does the model supply information on structural analogues for the substance?". Information on analogues could then be included, although a separate analysis of existing data for structural analogues should also be included as a separate information source in the overall analysis (in the WERF). The search for analogues is also mentioned as separate step in the stepwise approach for the use of non-testing data in the Cross-cutting guidance on the use of (Q)SARs.
- Section 4.3 of the Cross Cutting guidance on (Q)SARs would provide a good set of (sub)questions that need to be answered on the Applicability Domain issue both in the QPRF as well as in the QMRF.
- It was noticed that the use of more information defining the Applicability Domain also implies that this information will be used in a Weight-of-Evidence approach.
- The heading "other information regarding prediction reliability" is now too undefined to be of much use. Either model specific questions are detailed here (leading to model specific QPRFs) or this heading should cover (at the minimum) issues that are indicated in the QMRF to influence the reliability of a specific prediction. That way (by referring to the QMRF for the specific questions that need to be addressed) the QPRF can stay generic. Also see the fourth bullet point in the discussion on the WERF. The biodegradation example that was distributed before the meeting illustrates this discussion best, as the QPRF for BioWIN incorporates some very model specific questions, i.e. "What is the frequency of appearance in the training set of the identified substructures". Instead of specifying these model specific questions in the QPRF, it was proposed to introduce a field in the QMRF where all questions that should be addressed in the QPRF are specified. In the biodegradation example the QMRF on Biowin should include a text (i.e. under "Miscellaneous information") that states that the issue of the frequency of appearance in the training set of the identified substructures should be addressed in the QPRF.
- A separate heading "Conclusion" should not be part of the QPRF. A model result (the interpretation of the model outcome), and a rationale/reasoning on the reliability of this result (but not a reliability code) should suffice here. Instead of having a separate heading "Conclusion" is is proposed to change this to "Result"

In general, the prediction reporting format is thought to be very helpful in the use and evaluation of (Q)SARS, specifically after the above mentioned issues have been incorporated. The reporting of a specific model result in such a reporting format will help to start working with (Q)SARs in a transparent manner. It was recognised it is important to balance the amount of information requested to such an extent that is remains doable/workable. At this stage the amount of information was considered sufficient refinement could be given in the headings. An adapted QPRF, incorporating the issues discussed, is proposed and appended to these minutes (*Appendix 3 in this report*).

Discussion point 3 –the Weight-of-Evidence Reporting Format (WERF)

For the sake of discussion, and to force the participants to think about the consequences of conclusions on the QPRF level for regulatory decisions, the examples also included a WERF for two specific regulatory endpoints (Persistency in the PBT assessment, and Classification and Labelling). The provided WERF did not contain more that the summary of predictions taken from the QPRF, and a conclusion (based on these assembled data). It was stressed that in “real life” also other information (from non-testing data, *in vitro* data a well as existing experimental data) should be taken into account as well in the Weight-of-Evidence approach.

Comments received during the meeting:

- A number of participant were of the opinion that a WoE Reporting Format is much to prescriptive, and they were consequently against the use of a reporting format at this level.
- Others indicated that it is important to at least start thinking about how to report the Weight-of-Evidence discussion. In all RIP 3.3 EWG products this issue has not been touched upon (yet), and none of the EWGs have been able to address this issue in a proper and transparent manner.

No conclusion on this format was reached, although it was agreed that it is useful for the discussion and for working with the QPRF and QMRF that people realize that these formats serve to reach a (regulatory) conclusion in the end.

Discussion point 4 – The use of reliability codes or scores in the Reporting Formats

In the examples provided the prediction given in the QPRF were provided with a reliability score, similar to a Klimisch code for the reliability of experimental data. Whether or not there is a need for such a score, at which reporting level, and what form it should take was open for discussion.

Comments received during the meeting:

- Most participants were not in favour of using Klimisch codes to indicate the reliability of a (Q)SAR prediction. A number of participants felt that it was impossible to assign such a code in isolation (at the QPRF level) since all other information (presented in the WERF) or the absence thereof will also influence the assignment of a reliability score for a given purpose. Providing a Klimisch code at the QPRF level would then imply that at this stage the regulatory use is already included. If the Klimisch code only refers to the performance of the model than it might lead to confusion at the WERF level.
- After some discussion most participants were of the opinion that a code / score should be omitted at the QPRF level, and the prediction should be accompanied by a rationale including remarks on Applicability Domain, validity etc.
- Some participants were strongly against the use of a reliability score / code, as this would mean loosing information (when compared to a textual rationale). Furthermore a score might imply that all possible factors influencing the reliability of a prediction have been taken into account, although often not all information will have been available to come to this conclusion.
- A remark was made that in OECD programmes people are already using this kind of scoring, also for scoring non-testing data.

It was concluded that at this point the reliability score is left out of the QPRF, but with more experience gained on the use of these formats it will be re-evaluated in the future whether or not a simple score of reliability is wanted, and at what level (maybe the WERF would be a more appropriate level to score reliability of individual data).

The discussion was concluded with the very valid remark / useful recommendation that we should not try to create “perfect” reporting formats at once. It will work better to establish something, start working with it now, and refine the formats by gaining experience.

Emiel Rorije/Betty Hakkert,

RIVM, The Netherlands

October 2006

Appendix 2. (Q)SAR Model Reporting Format – QMRF

(European Chemicals Bureau, Ispra, Italy - Draft Version 1.2)

Please, try to fill in the fields of the QMRF for the model of interest. If the field is not pertinent with the model you are describing, or if you cannot provide the requested information, please answer “no information available”. **The set of information that you provide will be used to facilitate regulatory considerations of (Q)SARs.** For this purpose, the structure of the QMRF is devised to reflect as much as possible the OECD principles for the validation, for regulatory purposes, of (Q)SAR models. You are invited to consult the OECD “Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship Models” that can aid you in filling in a number of fields of the QMRF.

1. QSAR identifier

1.1 **QSAR identifier (title):** *Provide a short and indicative title for the model including relevant keyword. Some possible keywords are: endpoint modelled (as specified in field 3.2, recommended), name of the model, name of the modeller, and name of the software coding the model. Examples: “BIOWIN for Biodegradation”; “TOPKAT Developmental Toxicity Potential Aliphatic Model”.*

1.2 **Other related models:** *If appropriate, identify any model that is related to the model described in the present QMRF. Example: “TOPKAT Developmental Toxicity Potential Heteroaromatic Model and TOPKAT Developmental Toxicity Potential Carboaromatic Model” (these two models are related to the primary model “TOPKAT Developmental Toxicity Potential Aliphatic Model”).*

1.3 **Software coding the model:** *If appropriate, specify the name and the version of the software that implements the model. Examples: “BIOWIN v. 4.2 (EPI Suite)”; “TOPKAT v. 6.2”.*

2. General information

2.1 **Date of QMRF:** *Report the date of QMRF drafting (day/month/year). Example: “5 November 2006”.*

2.2 **QMRF author(s) and contact details:** *Indicate the name and the contact details of the author(s) of the QMRF (first version of the QMRF).*

2.3 **Date of QMRF update(s):** *Indicate the date (day/month/year) of any update of the QMRF. The QMRF can be updated for a number of reasons such as additions of new information (e.g. addition of new validation studies in section 7) and corrections of information.*

2.4 **QMRF update(s):** *Indicate the name and the contact details of the author(s) of the updates QMRF (see field 2.3) and list which sections and fields have been modified.*

2.5 **Model developer(s) and contact details:** *Indicate the name of model developer(s)/author(s), and the corresponding contact details; possibly report the contact details of the corresponding author.*

2.6 **Date of model development and/or publication:** *Report the year of release/publication of the model described in the current QMRF.*

2.7 **Reference(s) to main scientific papers and/or software package:** *List the main bibliographic references (if any) to original paper(s) explaining the model development and/or software implementation. Any other reference such as references to original experimental data and related models can be reported in field 9.2 “Bibliography”.*

2.8 **Availability of information about the model:** *Indicate whether the model is proprietary or non-proprietary and specify (if possible) what kind of information about the model cannot be disclosed or are not available (e.g., training and external validation sets, source code, and algorithm). Example: "The model is non-proprietary but the training and test sets are not available"; "The model is proprietary and the algorithm and the data sets are confidential".*

2.9 **Availability of another QMRF for exactly the same model:** *Indicate if you are aware or suspect that another QMRF is available for the current model you are describing. If possible, identify this other QMRF.*

3. Defining the endpoint – OECD Principle 1

PRINCIPLE 1: "A DEFINED ENDPOINT". ENDPOINT refers to any physicochemical, biological, or environmental effect that can be measured and therefore modelled. The intent of PRINCIPLE 1 (a (Q)SAR should be associated with a defined endpoint) is to ensure clarity 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. It is therefore important to identify the experimental system that is being modelled by the (Q)SAR.

3.1 **Species:** *Indicate the species for the endpoint being modelled.*

3.2 **Endpoint:** *Choose the endpoint (physicochemical, biological, or environmental effect) from the pre-defined classification. If the pre-defined classification does not include the endpoint of interest, select "Other" and report the endpoint in the subsequent field 3.3.*

3.3 **Comment on the endpoint:** *Include in this field any other information to define the endpoint being modelled. Specify the endpoint further if relevant, e.g. according to test organism such as species, strain, sex, age or life stage; according to test duration and protocol; according to the detailed nature of endpoint etc. You can also define here the endpoint of interest in case this is not listed in the pre-defined classification (see field 3.2) or you can add information about a second endpoint modelled by the same model. Example: Nitrate radical degradation rate constant: kNO3.*

3.4 **Endpoint units:** *Specify the units of the endpoint measured.*

3.5 **Dependent variable:** *Specify the relationship between the dependent variable being modelled and the endpoint measured since the two quantities may be different. Example: For modelling purposes all rate constants (i.e. Nitrate radical degradation rate constant kNO3) were transformed to logarithmic units and multiplied by -1 to obtain positive values. The dependent variable is: -log(kNO3).*

3.6 **Experimental protocol:** *Make any useful reference to a specific experimental protocol (or protocols) followed in the collection and evaluation of the experimental data sets.*

3.7 **Endpoint data quality and variability:** *Provide available information about the test data selection and evaluation and include a description of the data quality used to develop the model. This includes provision of information about the variability of the test data, i.e. repeatability (variability over time) and reproducibility (variability between laboratories) and sources of error (confounding factors which may influence testing results).*

4. Defining the algorithm – OECD Principle 2

PRINCIPLE 2: "AN UNAMBIGUOUS ALGORITHM". The (Q)SAR estimate of an endpoint is the result of applying an ALGORITHM to a set of structural parameters which describe the chemical structure. The intent of PRINCIPLE 2 (a (Q)SAR should be associated with a unambiguous algorithm) is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. In this context, algorithm refers to any mathematical equation, decision rule or output from a formalised modelling approach.

4.1 **Type of model:** *Describe the type of model (e.g., SAR, QSAR, Expert System, Neural Network, etc.).*

4.2 **Explicit algorithm:** *Report the algorithm (only the algorithm) for generating predictions from the descriptors; more text information about the algorithm can be reported in the following fields of this section or as supporting information (see field 9.3). If the algorithm is too long and complicated and thus cannot be reported here, include in this field a reference to a paper or a document where the algorithm is described in detail. This material can be attached as supporting information.*

4.3 **Descriptors in the model:** *Identify the number and the name or identifier of the descriptors included in the model. In this context, descriptors refers to e.g. physicochemical parameters, structural fragments etc*

4.4 **Descriptor selection:** *Indicate the number and the type (name) of descriptors / decision rules initially screened, and explain the method used to select the descriptors and develop the model from them.*

4.5 **Algorithm and descriptor generation:** *Explain the approach used to derive the algorithm and the method (approach) used to generate each descriptor.*

4.6 **Software name and version for descriptor generation:** *Specify the name and the version of the software used to generate the descriptors. If relevant, report the specific settings chosen in the software to generate a descriptor.*

4.7 **Descriptors/Chemicals ratio:** *Report the following ratio: number of descriptors to number of chemicals (chemicals from the training set), if applicable (if not, explain why).*

5. Defining the applicability domain – OECD Principle 3

PRINCIPLE 3: "A DEFINED DOMAIN OF APPLICABILITY". APPLICABILITY DOMAIN refers to the response and chemical structure space in which the model makes predictions with a given reliability. Ideally the applicability domain should express the structural, physicochemical and response space of the model. The CHEMICAL STRUCTURE (x variable) space can be expressed by information on physicochemical properties and/or structural fragments. The RESPONSE (y variable) can be any physicochemical, biological or environmental effect that is being predicted. According to PRINCIPLE 3 a (Q)SAR should be associated with a defined domain of applicability. Section 5 can be repeated (e.g., 5.a, 5.b, 5.c, etc) as many time as necessary if more than one method has been used to assess the applicability domain.

5.1 **Description of the applicability domain of the model:** *Describe the response and chemical structure and/or descriptor space in which the model makes predictions with a given reliability. Discuss if relevant whether: a) fixed or probabilistic boundaries define the applicability domain; b) structural features, a descriptor or a response space defines the applicability domain; c) in the case of SAR, there exists a description of the limits on its applicability (inclusion and/or exclusion rules regarding the chemical classes to which the substructure is applicable); d) in the case of SAR, there exist rules describing the modularity effects of the substructure's molecular environment; e) in the case of QSAR, there exist inclusion and/or exclusion rules that define the descriptor variable ranges for which the QSAR is applicable; f) in the case of QSAR, there exist inclusion and/or exclusion rules that define the response variable ranges for which the QSAR is applicable; g) there exists a (graphical) expression of how the descriptor values of the chemicals in the training set are distributed in relation to the endpoint values predicted by the model.*

5.2 **Method used to assess the applicability domain:** *Describe the method used to assess the applicability domain of the model.*

5.3 **Software name and version for applicability domain assessment:** *Specify the name and the version of the software used to apply the applicability domain method, where applicable. If relevant, report the specific settings chosen in the software to apply the method.*

5.4 **Limits of applicability:** *Describe for example the inclusion and/or exclusion rules (fixed or probabilistic boundaries, structural features, descriptor space, response space) that define the applicability domain.*

6. Defining goodness-of-fit and robustness – OECD Principle 4

PRINCIPLE 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY". PRINCIPLE 4 expresses the need to perform validation to establish the performance of the model. GOODNESS-OF-FIT and ROBUSTNESS refer to the internal model performance.

6.1 **Availability of the training set:** *Indicate whether the training set is somehow available (e.g., published in a paper, embedded in the software implementing the model, stored in a database) and appended to the current QMRF as supporting information (field 9.3). If it is not available, explain why. Example: "It is available and attached" "It is available but not attached"; "It is not available because the data set is proprietary"; "The data set could not be retrieved".*

6.2 **Available information for the training set***:* *Indicate whether the following information for the training set is reported as supporting information (see field 9.3): a) Chemical names (common names and/or IUPAC names); b) CAS numbers; c) SMILES; d) InChI codes; e) MOL files; f) Structural formula; g) Any other structural information.*

6.3 **Data for each descriptor variable for the training set***:* *Indicate whether the descriptor values of the training set are available and are attached as supporting information (see field 9.3).*

6.4 **Data for the dependent variable (response) for the training set:** *Indicate whether dependent variable values of the training set are available and attached as supporting information (see field 9.3).*

6.5 **Other information about the training set:** *Indicate any other relevant information about the training set (e.g, number and type of compounds in the training set (e.g. for models predicting positive and negative results the number of positives and the number of negatives in the training set)).*

6.6 **Pre-processing of data before modelling:** *Indicate whether raw data have been processed before modelling (e.g. averaging of replicate values); if yes, report whether both raw data and processed data are given.*

6.7 **Statistics for goodness-of-fit:** *Report here goodness-of-fit statistics (r^2, r^2 adjusted, standard error, sensitivity, specificity, false negatives, false positives, predictive values etc).*

6.8 **Robustness – Statistics obtained by leave-one-out cross-validation:** *Report here the corresponding statistics.*

6.9 **Robustness – Statistics obtained by leave-many-out cross-validation:** *Report here the corresponding statistics, the strategy for splitting the data set (e.g. random , stratified), the percentage of left out compounds and the number of cross-validations.*

6.10 **Robustness – Statistics obtained by Y-scrambling:** *Report here the corresponding statistics and the number of iterations.*

6.11 **Robustness – Statistics obtained by bootstrap:** *Report here the corresponding statistics and the number of iterations.*

6.12 **Robustness – Statistics obtained by other methods:** *Report here the corresponding statistics.*

7. Defining predictivity – OECD Principle 4

PRINCIPLE 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY". PRINCIPLE 4 expresses the need to perform validation to establish the performance of the model. PREDICTIVITY refers to the external model validation. Section 7 can be repeated (e.g., 7.a, 7.b, 7.c, etc) as many time as necessary if more validation studies needs to be reported in the QMRF.

7.1 **Availability of the external validation set:** *Indicate whether an external validation set is available and appended to the current QMRF as supporting information (field 9.3). If it is not available, explain why.*

7.2 **Available information for the external validation set:** *Indicate whether the following information for the external validation set is reported as supporting information (see field 9.3): a) Chemical names (common names and/or IUPAC names); b) CAS numbers; c) SMILES; d) InChI codes; e) MOL files; f) Structural formula; g) Any other structural information.*

7.3 **Data for each descriptor variable for the external validation set***: Indicate whether descriptor values of the external validation set are somehow available and attached as supporting information (see field 9.3).*

7.4 **Data for the dependent variable for the external validation set:** *Indicate whether dependent variable values of the external validation set are somehow available and attached as supporting information (see field 9.3)..*

7.5 **Other information about the external validation set:** *Indicate any other relevant information about the validation set. Example: "External validation set with 56 compounds appended".*

7.6 **Experimental design of test set:** *Indicate any experimental design for getting the test set (e.g. by randomly setting aside chemicals before modelling, by literature search after modelling, by prospective experimental testing after modelling, etc.).*

7.7 **Predictivity – Statistics obtained by external validation:** *Report here the corresponding statistics. In the case of classification models, include false positive and negative rates.*

7.8 **Predictivity – Assessment of the external validation set:** *Discuss whether the external validation set is sufficiently large and representative of the applicability domain.*

Describe for example the descriptor and response range or space for the validation test set as compared with that for the training set. Here the descriptor values of the chemicals predicted by the model (training set) should be compared with the descriptor value range of the test set. In addition the distribution of the response values of the chemicals in the training set should be compared to the distribution of the response values of the test set.

7.9 **Comments on the external validation of the model:** *Add any other useful comments about the external validation procedure.*

8. Providing a mechanistic interpretation – OECD Principle 5

PRINCIPLE 5: "A MECHANISTIC INTERPRETATION, IF POSSIBLE". According to PRINCIPLE 5, a (Q)SAR should be associated with a mechanistic interpretation, if possible.

8.1 **Mechanistic basis of the model:** *Provide information on the mechanistic basis of the model (if possible). In the case of SAR, you may want to describe (if possible) the molecular features that underlie the properties of the molecules containing the substructure (e.g. a description of how sub-structural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region). In the case of QSAR, you may give (if possible) a physicochemical interpretation of the descriptors used (consistent with a known mechanism of biological action). If it is not possible to provide a mechanistic interpretation, try to explain why.*

8.2 ***A priori* or *a posteriori* mechanistic interpretation:** *Indicate whether the mechanistic basis of the model was determined a priori (i.e. before modelling, by ensuring that the initial set of* *training structures and/or descriptors were selected to fit pre-defined mechanism of action) or a posteriori (i.e. after modelling, by interpretation of the final set of training structures and or descriptors).*

8.3 **Other information about the mechanistic interpretation:** *Report any other useful information about the (purported) mechanistic interpretation described in the previous fields (8.1 and 8.2) such as any reference supporting the mechanistic basis.*

9. Miscellaneous information

9.1 **Comments:** *Add here other relevant and useful comments (e.g. other related models, known applications of the model) that may facilitate regulatory considerations on the model described. Include if relevant experience obtained by use of model prediction for various types of regulatory decisions (incl. references as appropriate).*

9.2 **Bibliography**: *Report useful references other than those directly associated with the model development (references describing the model development are reported in field 2.5).*

9.3 **Supporting information**: *Indicate whether supporting information is attached (e.g. external documents) to this QMRF and specify its content and possibly its utility.*

10. Summary for the ECB Inventory

The summary section is specific for the ECB Inventory. If the model is submitted to ECB for inclusion in the ECB Inventory of QSAR models, then this summary is compiled by ECB after QMRF submission. The QMRF author does not have to fill in any of the fields of the summary section.

10.1 **QMRF number:** *A unique number (numeric identifier) is assigned to any QMRF that is published in the ECB inventory. The number encodes the following information: model described in the QMRF (as derived from field 4.2), software implementing the model (as derived from field 1.3), version of the QMRF for the same model and the same software (as derived from the information included in field 2.4) and author of the QMRF (as derived from field 2.2). The number is unique for any QMRF uploaded and stored in the ECB inventory.*

10.2 **Publication date**: *The date (day/month/year) of publication in the ECB inventory is reported here.*

10.3 **Keywords:** *Any relevant keywords associated with the present QMRF are reported here.*

10.4 **Comments:** *Any comments that are relevant for the publication of the QMRF in the ECB Inventory (e.g., comments about updates and about supporting information) are reported here.*

Appendix 3. (Q)SAR Prediction Reporting Format – QPRF

RIVM Proposal for (final) QPRF incorporating points raised during the October 12, 2006 meeting, and the written comments received by e-mail prior to the meeting. This (empty) format is followed by a version with an indication of the required information for each heading/question.

QPRF Version date 07-2007

GENERAL

Prediction for Substance	
Model Name, Version and date of prediction	
(Q)SAR Model Reporting Format (QMRF)	
Endpoint description	

PREDICTION

Model outcome:	

INFORMATION RELEVANT for the ASSESSMENT of the PREDICTION

Model Algorithm / Result interpretation	
Is the substance within the Domain of Applicability of the model?	
Alerts/fragments identified and/or rules applicable to the substance?	
Indicate structural analogues identified by the model?	
Is the substance part of training set?	
Other information regarding prediction reliability?	

CONCLUSION

Result	
Reasoning	

(Q)SAR Prediction Reporting Format (QPRF) with explanation on what information is expected under the specific headings (in italics). It should be noted that not all questions will be relevent for all predictions, and for specific predictions / models it will not be possible to supply the full information for every question on the QPRF.

QPRF Version date 07-2007

GENERAL

Prediction for Substance	*Identify the substance for which the prediction is done. A separate chemical identity format. Name, CAS-nr, Structure*
Model Name, Version and date of prediction	*Identify as unambiguously as possible the model that was used to generate the prediction*
(Q)SAR Model Reporting Format (QMRF)	*Refer to an entry in the ECB Inventory of (Q)SARs (http://ecb.jrc.it/(Q)SAR/(Q)SAR_tools/(Q)SAR_tools_qrf.php) whenever possible. Otherwise refer to the QMRF document accompanying this QPRF with the model description.*
Endpoint description	*Description of the exact endpoint that the model is predicting/reproducing (not necessarily a regulatory relevant endpoint)*

PREDICTION

Model outcome:	*exact outcome as produced by the model, before interpretation. For example a value for a quantitative model.*

INFORMATION RELEVANT for the ASSESSMENT of the PREDICTION

Model Algorithm / Result interpretation	*Explanation of the algorithm used by the model, and/or the interpretation that needs to be applied to the model outcome.*
Is the substance within the Domain of Applicability of the model?	*Subquestions that (preferably) should be answered here:* *1) is the chemical of interest within the scope of the model, according to the defined applicability domain of the model?* *a) descriptor domain (do the descriptor values of the chemical fall within defined ranges?)* *b) structural fragment domain: does the chemical contain fragments that are not represented in the model training set ?* *c) mechanistic domain: does the chemical of interest act according to the same mode or mechanism of action as other chemicals for which the model is applicable?* *d) metabolic domain: does the chemical of interest undergo transformation or metabolism, and how does this affect reliance on the prediction for the parent compound ?* *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?*

Alerts/fragments identified and/or rules applicable to the substance?	*Identify which part(s) of the structure contribute to (the interpretation of) the model result. Mention applicable rules (i.e. on skin penetration) that fortify or disqualify the model outcome.*
Indicate structural analogues identified by the model?	*Mention analogues (and their experimental data) identified by the model, and/or substances from the model training set that are close structural analogues (This step is not supposed to replace the extensive structural analogue search outside of the model training set data which is proposed in the Stepwise approach to the use of non-testing data, Cross-cutting guidance on the use of (Q)SARs*
Is the substance part of training set?	*Yes/no, also indicate the experimental value used in the training set for this substance*
Other information regarding prediction reliability?	*Indicate all factors not discussed above that influence the reliability of this specific prediction. As a minimum the list of prediction specific issues as identified under "Miscellaneous information" in the QMRF should be adressed here.*

CONCLUSION

Result	*interpreted model prediction (i.e. a Biowin5 prediction of 0.98 would here be described in its interpreted form: Readily Biodegradable in the OECD301C mod.MITI test)*
Reasoning	*reasoning on reliability of the result, summarizing all factors influencing reliability for specific prediction as discussed above*

Appendix 4. Weight-of-Evidence Reporting Format – WERF

RIVM Proposal for a reporting format which can be used for the evaluation of the overall body of evidence for the combination of a specific (toxicological) endpoint and a specific (regulatory) purpose or setting. It should be noted that the October 12, 2006 meeting of the EU (Q)SAR Working Group felt that a Reporting Format on this level (the actual regulatory decision based on all (summarized) data) was not needed or too restrictive. The need to actually perform such an evaluation of all available evidence was however clearly identified. This “format” is therefore only used (in the (Q)SAR Experience project) as a means to evaluate the adequacy and completeness of the information in the QMRF and (specifically) the QPRF. The information reported in the QPRF should allow a user to estimate both the reliability of a prediction (or a test outcome) and the applicability of the prediction (or test outcome) to the specific regulatory endpoint for which a conclusion will be drawn. If multiple model predictions, or multiple test results should be reported, one entry (name, result and reasoning) for each model prediction, or each *in vitro* or *in vivo* test result should be presented.

This (empty) format is followed (on the next page) by a version with an indication of the required information for each heading/question.

WERF Version 2007

SUBSTANCE

WERF for substance:	

ENDPOINT

Regulatory endpoint:	

DATA – (Q)SARs, category approach, *in vivo* and *in vitro* test data

(Q)SAR Model name	Result	
	Reasoning	
In vitro test name	Result	
	Reasoning	
In vivo test name	Result	
	Reasoning	
Other data:	Result	
	Reasoning	

CONCLUSION

Weighted summary of the presented data	Result	
	Reasoning	
Need for further testing?		

Weight-of-Evidence Reporting Format (WERF) with indications what information is expected under the specific headings (in italics).

WERF Version 2007

SUBSTANCE

WERF for substance:	*Name of the substance and/or other identifier like CAS or EINECS nr*

ENDPOINT

Regulatory endpoint:	*Description of the regulatory framework for which the conclusion will be used. One can imagine that data will be judged differently in different regulatory frameworks (different threshold values for example). Also the required reliability of a specific prediction or test outcome can be different for different regulatory frameworks.*

DATA – (Q)SARs, category approach, *in vivo* and *in vitro* test data

(Q)SAR Model name	Result	*Present the result from the QPRF for a specific model prediction*
	Reasoning	*Present the reasoning from the QPRF for a specific model prediction, on the reliability of the result as well as the interpretation (meaning) of the model result for the regulatory endpoint under evaluation*
In vitro test name	Result	*Present an in vitro test result*
	Reasoning	*Elaborate on the reliability of the test result, and the meaning of the test result for the regulatory endpoint under evaluation*
In vivo test name	Result	*Present an in vivo test result*
	Reasoning	*Elaborate on the reliability of the test result, and the meaning of the test result for the regulatory endpoint under evaluation*
Other data:	Result	
	Reasoning	

CONCLUSION

Weighted summary of the presented data	Result	*What is the conclusion for the specific regulatory endpoint under evaluation. For example: this substance is Persistent / Not Persistent when evaluating degradation data and predictions for a PBT assessment.*
	Reasoning	*Provide reasoning for the result, indicating how the presented data is weighted and/or why a specific data point is preferred or dismissed from the evaluation.*
Need for further testing?	*Can you make an assessment based on the data provided? Is more information needed? More ((Q)SAR) model data? Or specific input into a model (e.g. physico-chemical data)? Is there a need for (further) testing, either in vitro or in-vivo?*	

Appendix 5. QPRF and WERF examples for four different substances

4 examples of a WERF are provided, together with the appropriate QPRFs for the models that have been applied. A large part of the formats has been filled out, but partes (notably the conclusions) were intentionally left open or question (in italics) were entered in order to have the participants of the (Q)SAR Experience Project form an opinion on what should be concluded for the different examples. The QMRFs of the respective models (BIOWIN 1-6 for biodegradation, TOPKAT and DEREKfW for both skin irritation and sensitization, the Potts&Guy for skin sensitization and BfR/Gerner rules for skin irritation) were provided as well to the participants of the (Q)SAR experience meeting. A number of these QMRFs can already be encountered at the ECB Inventory of (Q)SAR models (http://ecb.jrc.it/(Q)SAR/(Q)SAR-tools/(Q)SAR_tools_qrf.php).

Since the formats used for the examples were improved upon according to the discussion of the examples at the (Q)SAR WG meeting in October 2006, they are different from the versions in Appendices 3 and 4).

E.g. the examples still contain a field for a Klimisch code of reliability of the prediction, something that was decided against in the (Q)SAR WG meeting.

CONTENTS of the APPENDIX:

Skin Sensitization example for Cinnamaldehyde / Classification&Labelling:

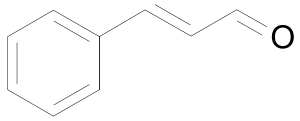
Skin Irritation example for 4,4'-methylenebis(2,6-dimethylphenyl isocyanate) / C&L:

Skin Irritation example for 4,4'-diisobutylethylidenediphenol / C&L:

Biodegradation example for dibenzyltoluene / PBT assessment:

Skin Sensitization example for: Cinnamaldehyde / Classification & Labelling:

IDENTITY

Chemical Name (English)	Cinnamaldehyde
CAS RN	104-55-2
EINECS/ELINCS-nr.	203-213-9
SMILES	C1ccccc1C=CC=O
Structure (2D):	
Molecular Weight	132.16 g/mol
Bruto Formula	C_9H_8O

PHYSICO-CHEMICAL PARAMETERS

Parameter	**Value**	**Unit**	**Source**
Log K_{ow}	2.12 +/-0.36		ACD Labs

N.B.:
Since the formats used for the examples were improved upon according to the discussion of the examples at the (Q)SAR WG meeting in October 2006, they are different from the versions in Appendices 3 and 4.

(Q)SAR Prediction Reporting Format – DEREKfW

Adapted from the example provided by ECB, Italy

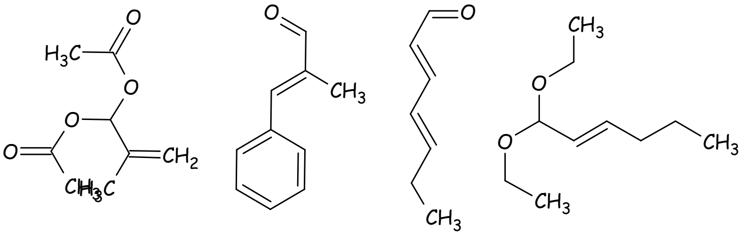
Prediction for Substance	Cinnamaldehyde

Model Name	Derek for Windows Version 9
(Q)SAR Model Reporting Format (QMRF)	QMRF – Skin Sensitization - DEREK.doc
Endpoint description	Skin Sensitization

PREDICTION

Results:	Alert 479 alpha,beta-Unsaturated aldehyde or precursor. Skin sensitisation. Number of matches = 1, Plausible skin sensitiser

INFORMATION RELEVANT for the ASSESSMENT of the PREDICTION

Algorithm / Result interpretation	Identification of substructures (structural alerts).
Domain of applicability	CAD is a small organic chemical that is likely to be within the domain of the DEREKfW sensitisation rulebase
Alerts/fragments identified	See next page for the Alert overview: of 479 alpha,beta-Unsaturated aldehyde or precursor
Rules applicable	Not applicable
Structural analogues identified?	Known skin sensitizers: 
Is the substance part of training set?	The substance is not part of the training set.
model specific information on the prediction reliability	This alert describes the skin sensitisation of alpha,beta-unsaturated aldehydes and precursors which interact with skin proteins via a Michael addition mechanism [Patlewicz et al]. beta-Disubstituted alpha,beta-unsaturated aldehydes are less susceptible to Michael addition and are thought to react via a Schiff base mechanism [Patlewicz et al]. The activity of such compounds is described elsewhere in the knowledge base. Skin sensitisation activity for alpha,beta-unsaturated aldehydes and their precursors has been demonstrated in various assays including the guinea pig maximisation test [Cronin and Basketter] and the mouse local lymph node assay [Patlewicz et al]. Skin sensitisation in humans has also been described [Ford].

CONCLUSION

Result	Plausible skin sensitiser
Result Reasoning	The alert in this case is well described and the chemical of interest has other structural analogues that are of the same homologous with positive sensitising data.
Reliability	High *(ECB assessment of reliability. Would you assign some other reliability, or a numerical (Klimisch) code? Is that more or less helpful?)*
Reliability Reasoning	High - as alert is well characterised and examples very similar. Cinnamic aldehyde is well within the scope and domain of the alert - hence this would be a robust assessment.

DEREK for Windows

Alert overview: 479 alpha,beta-Unsaturated aldehyde or precursor

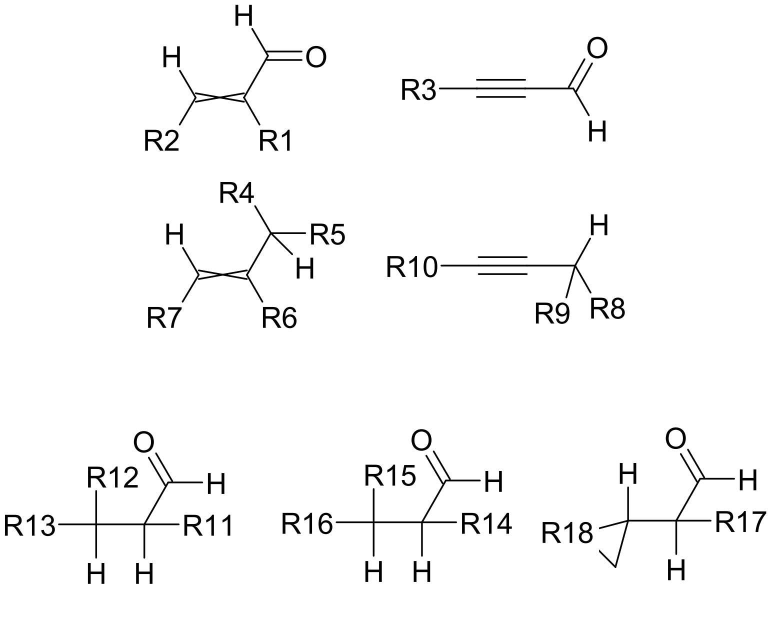
This alert describes the skin sensitisation of alpha,beta-unsaturated aldehydes and precursors which interact with skin proteins via a Michael addition mechanism [Patlewicz et al]. beta-Disubstituted alpha,beta-unsaturated aldehydes are less susceptible to Michael addition and are thought to react via a Schiff base mechanism [Patlewicz et al]. The activity of such compounds is described elsewhere in the knowledge base.

Skin sensitisation activity for alpha,beta-unsaturated aldehydes and their precursors has been demonstrated in various assays including the guinea pig maximisation test [Cronin and Basketter] and the mouse local lymph node assay [Patlewicz et al]. Skin sensitisation in humans has also been described [Ford].

A series of alpha,beta-unsaturated ester precursors have been shown to give a positive response in a modified single injection adjuvant test in the guinea pig [Franot et al 1994a, Franot et al 1994b]. Correspondingly, the precursors of alpha,beta-unsaturated aldehydes are included in the scope of the current alert.

1,3-Benzodioxole precursors and other aromatic analogues are excluded from the alert on the basis that they are relatively stable, hydrolysing readily only at raised temperatures and pH < 1 [Greene and Wuts].

The presence of a skin sensitisation structural alert within a molecule indicates the molecule has the potential to cause skin sensitisation. Whether or not the molecule will be a skin sensitiser will also depend upon its percutaneous absorption. Generally, small lipophilic molecules are more readily absorbed into the skin and are therefore more likely to cause sensitisation.



R1 = H, C, F, Cl, Br, I
R2, R3 = any except OH

R4, R5, R8, R9 = O, S
R6 = H, C, F, Cl, Br, I
R7, R10 = any except OH

R11, R14, R17 = H, C, F, Cl, Br, I
R12 = F, Cl, Br, I
R13 = C, H, F, Cl, Br, I
R15 = OSO_2C, OP, SCN
R16 = H, C
R18 = O, S, N (no additional heteroatoms attached)

(Q)SAR Prediction Reporting Format – TOPKAT

Adapted from the example provided by ECB, Italy

Prediction for Substance	Cinnamaldehyde
Model Name	TopKat v6.2
(Q)SAR Model Reporting Format (QMRF)	QMRF – Skin Sensitization - TOPKAT.doc
Endpoint description	Skin Sensitization

PREDICTION

Results:	Probability of 1 for being a sensitiser (vs. non-sensitiser), and Probability of 1 for being a strong sensitizer

INFORMATION RELEVANT for the ASSESSMENT of the PREDICTION

Algorithm / Result interpretation	Probability between 0-1. Results 0.3-0.7 are indicated to be not valid.
Domain of applicability	CAD falls within the Optimal Prediction Space (OPS) as described by the model.
Alerts/fragments identified	See next page for the Alert overview: of 479 alpha,beta-Unsaturated aldehyde or precursor
Rules applicable	Not applicable
Structural analogues identified?	Other structural analogues identified are phenyl acetic aldehyde [122-78-1], benzaldehyde [100-52-7] and cinnamic alcohol.
Is the substance part of training set?	CAD is within the training set of both models and reported to be a sensitiser in the GPMT
Model specific information on the prediction reliability	Statistics on Leave One Out cross validation for the relevant classes:

Chemical Class	Number of Compounds	Specificity %	Sensitivity %	Indeterminate %
NON SENS vs SENS				
Aliphatics & Single benzenes	252	91	93	3
Aromatics (excl single benzene)	75	81	95	1
WEAK/MOD vs STRONG				
Aliphatics & Single benzenes	158	89	85	6
Aromatics (excl single benzene)	59	90	92	3

CONCLUSION

Result	Probability of 1 for being a sensitiser (vs. non-sensitiser) and Probability of 1 for being a strong sensitiser
Result Reasoning	Cinnamic aldehyde falls within the applicability of both models in TOPKAT. however it is in the training set of both models and reported as a strong sensitiser in the GPMT.
Reliability	High, but the "most similar" analogues are weak in terms of supporting the prediction
Reliability Reasoning	High - as Cinnamic aldehyde is within the domain of the models and is also in the training set of both models. Structural analogues are weak and not thought to be particular similar since the driving factor in the sensitisation behaviour of CAD is thought to be the unsaturated carbonyl system rather than the carbonyl group itself.

(Q)SAR Prediction Reporting Format – Potts&Guy

Adapted from the example provided by ECB, Italy

Prediction for Substance	Cinnamaldehyde

Model Name	Potts&Guy
(Q)SAR Model Reporting Format (QMRF)	QMRF - Skin Penetration - POTTS&GUY.doc
Endpoint description	Skin penetration

PREDICTION

Results:	Kp = 9.53 x 10-3 cm/hr

ALL INFORMATION RELEVANT for RELIABILITY of the PREDICTION

Algorithm / Result interpretation	Molecular Weight and Ocatonol/Water partition coefficient are used to estimate a Kp value. Skin penetration is favourable if Kp > ??
Domain of applicability	The log P and MW values determined for CAD are within the ranges of descriptor values for the Potts and Guy model.
Alerts/fragments identified	See next page for the Alert overview: of 479 alpha,beta-Unsaturated aldehyde or precursor
Rules applicable	Not applicable
Structural analogues identified?	Not applicable.
Is the substance part of training set?	Not applicable.
model specific information on the prediction reliability	-

CONCLUSION

Result	skin permeability is strongly favoured
Result Reasoning	Cinnamic aldehyde falls within the applicability domain of the model.
Reliability	High *(ECB assignment of reliability. Would you assign a different reliability, or a numerical (Klimisch) code? Is that more or less helpful?)*
Reliability Reasoning	High - as Cinnamic aldehyde is within the domain of the model.

rivm

WERF (Weigh of Evidence Reporting Format) – Classification & Labelling

Adapted from the example provided by ECB, Italy

SUBSTANCE

WERF for substance:	Cinnamaldehyde

ENDPOINT

Regulatory endpoint:	EU Classification and Labelling for dangerous substances and preparations: http://ecb.jrc.it/Legislation/1967L0548EC.htm

DATA – (Q)SARs, category approach, *in vivo* and *in vitro* test data)

DEREK for Windows v.9	Result	Plausible skin sensitiser
	Reliability	High
	Reasoning	The alert in this case is well described and the chemical of interest has other structural analogues that are of the same homologous with positive sensitising data
TOPKAT v6.2	Result	Probability of 1 for being a sensitiser (vs. non-sensitiser) and a probability of 1 for being a strong sensitiser
	Reliability	High
	Reasoning	Cinnamic aldehyde falls within the applicability of both models in TOPKAT. However it is in the training set of both models and reported as a strong sensitiser in the GPMT.
Pott & Guy skin penetration model	Result	skin permeability is strongly favoured
	Reliability	High
	Reasoning	Cinnamic aldehyde falls within the applicability domain of the model.
Other Information	Result	Supporting information for the WoE conclusion (free text) is given on the next page
Available *in vitro* data	Result	No data available
	Reliability	
	Reasoning	
Available *in vivo* data	Result	No data available
	Reliability	
	Reasoning	

CONCLUSION

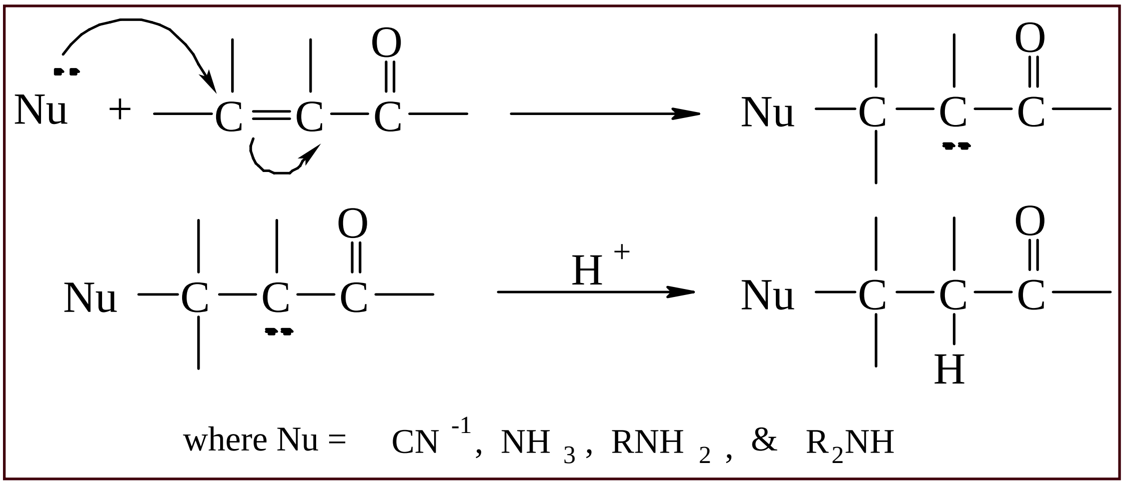
Weighted summary of the presented data	Result	*Should the substance be classified for Skin Sensitization (R43) or not?*
	Reliability	*Give a reliability*
	Reasoning	*Provide a reasoning*
Need for further testing	*Is further information and/or testing needed?* *> Physico-chemical or related to model input* *> In vitro testing* *> In vivo testing*	

Supporting information for the WoE conclusion

Cinnamic aldehyde falls within the applicability of both models in TOPKAT and is reported as a strong sensitiser in the GPMT. In DEREKfW, CAD is within the scope and domain of the alert and is predicted to be a plausible skin sensitiser by DEREKfW. Other information in the literature report shows CAD to be a moderate skin sensitiser in the LLNA. The skin permeability was predicted as Kp = 9.53 x 10-3 cm/hr. This value suggests that skin permeability is strongly favoured.

A critical requirement for skin sensitisation is that the chemical must form a stable (usually covalent) association with protein to form an immunogenic complex. Thus the reactivity and partition properties of a chemical are determining factors in whether that chemical will be a sensitiser or not. In this case, CAD has the potential to react covalently with a nucleophile via a Michael addition or Schiff base reaction. The former is preferred (as shown in the supporting figure). This is supported by the model predictions made and by the experimental *in vivo* test data that is already available. In addition, the penetration profile estimated suggests that this compound will readily penetrate.

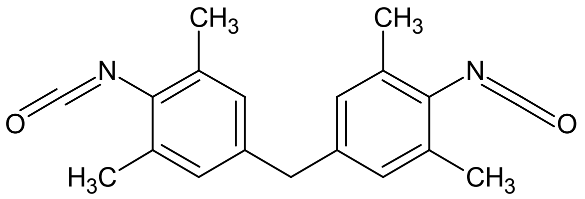
Reaction pathway for Michael addition compounds



where Nu = CN^{-1}, NH_3, RNH_2, & R_2NH

Skin Irritation example for: 4,4'-methylenebis(2,6-dimethylphenyl isocyanate) / Classification & Labelling

IDENTITY

Chemical Name (English)	4,4'-methylenebis(2,6-dimethylphenyl isocyanate)
CAS RN	101657-77-6
EINECS/ELINCS-nr.	
SMILES	O=C=Nc1c(C)cc(cc1C)Cc2cc(C)c(c(C)c2)N=C=O
Structure (2D):	
Molecular Weight	306.36 g/mol
Bruto Formula	$C_{19}H_{18}N_2O_2$

PHYSICO-CHEMICAL PARAMETERS

Parameter	Value	Unit	Source
Melting point	135 107	°C °C	(estimate) confidential test
Water Solubility	5.3 6.5	mg/l mg/l	(estimate) confidential test
Log Kow	7.4 7.6		(estimate) confidential test
Surface tension	37.8	mN/m	est. Chemsketch 8
Lipid solubility	3.87	??	Confidential test
Hydrolysis	Unknown		
pH in water solubility test	Unknown		

N.B.

Since the formats used for this example were improved upon according to the discussion of the examples at the (Q)SAR WG meeting in October 2006, they are different from the versions in Annex 3 and 4.

(Q)SAR Prediction Reporting Format – DEREKfW

Prediction for substance	4,4'-methylenebis(2,6-dimethylphenyl isocyanate)

MODEL

Model Name	DEREKfW8.0
(Q)SAR Model Reporting Format (QMRF)	QMRF – DEREK – Skin irritation.doc
Endpoint description	Skin Irritation (mammalian). Not necessarily strong enough to lead to classification (alert dependent)

PREDICTION

Result	Irritant to skin (mammals). Skin penetration favorable for skin

INFORMATION RELEVANT to the ASSESSMENT of the PREDICTION

Algorithm / Result interpretation	There is no algorithm, only a qualitative evaluation of structural alerts (leading to skin irritation) and parameters for skin penetration (favouring or hindering the potential skin irritation caused by the structural alert).
Domain of applicability	Organic substances that contain at least one alert. The substance contains the isocyanate structural alert for skin irritation, therefore the chemical is in the applicability domain
Alerts/fragments identified	R1-N=C=O, R1= carbon atom (2X)
Rules applicable	Skin penetration is favoured by relatively lipophilic molecules (log Kow = 1-4) of low molecular weight (<500). Log Kp: -2.036 Calc. by the Potts & Guy equation. Log P: 3.596 Calc. by the Moriguchi estimation MW: 306.37 g/mol
Structural analogues	The structural alert is illustrated with 4 analogues which are however smaller than the submitted chemical. See DEREKfW result. Known irritants given by DEREKfW: Methyl isocyanate, Ethyl isocyanate, Phenyl isocyanate, Toluene diisocyanate
Is the substance part of the training set?	No
Model specific information on the prediction reliability	The presence of two isocyanate alerts in one structure strengthens the prediction of skin irritation potential. log P (=log Kow) used is significantly different from the experimental value and from other estimations (ClogP and KOWWIN (Q)SARs).

CONCLUSION

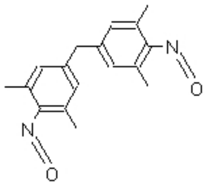
Result	Skin irritant
Reliability	*Assess reliability (using Klimisch codes or other ranking?)*
Reasoning	The presence of an alert for skin irritation (2X) indicates potential skin irritation. The alert is thought to be valid, the substance is within the structural domain of the alert. The evaluation of the potential for skin penetration is uses a suspect log Kow estimation. When the experimental value is used, skin penetration of the substance is NOT favorable. The interpretation of the combination of the effect of the structural alert and the influence of skin penetration is left to the end user, no definite prediction is given by the algorithm.

DEREK for Windows report

Version: 8.0.1
Species: human
mammal
SuperEndpoints: Irritation

Compound Name:
Log Kp: -2.036 Calculated by the Potts & Guy equation
Log P: 3.596 Calculated by the Moriguchi estimation
Molecular Weight: 306.365 Calculated by LPS

Submitted Compound:



List of alerts found:

211 Isocyanate. Irritation (of the skin, eye and respiratory tract). Number of matches = 2
Alert overview: 211 Isocyanate

R1—N=C=O

Known irritants which fire the alert include:
Methyl isocyanate
Ethyl isocyanate
Phenyl isocyanate
Toluene diisocyanate

R1 = C

Isocyanates are highly reactive substances and generally irritating to the skin, eyes and respiratory tract. Hydrolysis and reaction with biologically important molecules, including proteins, occurs rapidly. Irritation to the respiratory tract may occur at low concentrations. E.g. exposure of humans to 2ppm methyl isocyanate for 1-5 minutes produced tears and irritation to the nose and throat. Diisocyanates are generally stronger irritants than monoisocyanates. A polymeric isocyanate, polymethylene polyphenyl isocyanate, has been classified as irritating to the skin, eyes and respiratory tract.
N.B. A structural alert for irritancy indicates some potential for this effect. Additionally, except for highly reactive corrosive substances, the skin and eye irritation potential of a chemical is very dependent on physicochemical properties which influences the concentrations at and exposure to component tissues. Skin penetration is favoured by relatively lipophilic molecules (Log P(octanol/water)= 1-4) of low molecular weight (<500). For many classes of chemicals (e.g. aliphatic amines) eye irritation is greatest for the more water soluble compounds which readily dissolve in the aqueous tear film on the cornea and conjunctiva. Liquid substances (cf.solids) have good tissue contact and are more likely to be irritating, particularly to the skin. Highly reactive corrosive chemicals may penetrate tissue as a result of corrosive damage with a lower dependence on solubility characteristics.

QPRF ((Q)SAR Prediction Reporting Format) – GERNER Rules

Prediction for substance	4,4'-methylenebis(2,6-dimethylphenyl isocyanate)

MODEL

Model Name	BfrR (Gerner) physico-chemical exclusion rules for skin irritation
(Q)SAR Model Reporting Format (QMRF)	QMRF – Skin irritation – BfR (Gerner) rules.doc
Endpoint description	NOT Classifying for EU C&L as R38 (irritant to skin) and/or R34/R35 (corrosive to skin)

PREDICTION

Result	NOT R38 (irritant to skin) or R34/35 (corrosive to skin)

INFORMATION RELEVANT to the ASSESSMENT of the PREDICTION

Algorithm / Result interpretation	General algorithm of the exclusion rules: IF (rule) THEN substance is NOT R38 and/or R34/45
Domain of applicability	No organometallic compounds, Purity of the substance should be >95%. Specific rules for a subclass (class CN, substances containing only C, H, O and N atoms) exist. The substance is thought to be in the applicability domain.
Alerts/fragments identified	Not applicable

Rules applicable:	Class	Rule	Result	Goodness of fit
	CN	mol.weight > 290 g/mol	NOT R34/35	338/338
	CN	log Kow > 4.5	NOT R34/35	119/119
	CN	aqueous solubility < 0.1 mg/l	NOT R38	104/104
	CN	log Kow > 5.5	NOT R38	85/85

Structural analogues	Not given – no means available to search the training set for structural analogues.
Is the substance part of the training set?	Not known, but not probable as the training set are EU new chemicals
Model specific information on the prediction reliability	See the goodness of fit statistics under “Rules applicable”

CONCLUSION

Result	NOT R38 (irritant to skin) or R34/35 (corrosive to skin)
Reliability	Please assess reliability (using Klimisch codes?)
Reasoning	The aqueous solubility rule for the CN class gave one false negative in the external validation set (borderline substance). However in combination with the three other applicable rules the quality of the prediction is thought to be sufficient. The rules based on molecular weight and log Kow don’t have exceptions in the training set, and did not give any false negatives in the external validation set.

WERF (Weight-of-Evidence Reporting Format) – Classification & Labelling

SUBSTANCE

WERF for substance:	4,4'-methylenebis(2,6-dimethylphenyl cyanate)

ENDPOINT

Regulatory endpoint:	EU Classification and Labelling for dangerous substances and preparations: http://ecb.jrc.it/Legislation/1967L0548EC.htm

DATA – (Q)SARs, category approach, *in vivo* and *in vitro* test data

Does the intended use of the chemical give any indication for corrosive properties?	Result	Yes, reactive chemicals – skin corrosion or irritation might be likely
	Reliability	
	Reasoning	No data is available on the use of this substance but isocyanates are known to spontaneously react with water, forming a primary amine (known alert for skin irritancy) and carbondioxide.
Is the pH of the substance indicative of corrosive properties (2>pH>11.5)?	Result	No data available. Skin corrosion not likely
	Reliability	
	Reasoning	No strongly acidic or basic functionality is present, also not after reaction with water.
Is the substance an organic hydroperoxide?	Result	Not applicable
	Reliability	
	Reasoning	Substance is not an organic hydroperoxide
Is the substance an organic peroxide?	Result	Not applicable
	Reliability	
	Reasoning	Substance is not an organic peroxide
Does the substance contain impurities (> 0.1%) that are known skin irritants or corrosives?	Result	No – No classification needed for impurities
	Reliability	
	Reasoning	Impurities are not considered in this exercise
Results of the Gerner exclusion rules for skin irritation:	Result	Not a skin irritant (NOT R38), and not a skin corrosive (NOT R34/35)
	Reliability	Use reliability from the QPRF
	Reasoning	The combination of four applicable rules is thought to give sufficient evidence of the absence of skin irritation potential.
Results of the DEREKfW 8.0 prediction for skin irritation:	Result	Skin irritant (mammalian)
	Reliability	Use reliability from the QPRF
	Reasoning	The isocyanide alert (2X) indicates potential skin irritation. The evaluation of the potential for skin penetration by DEREKfW uses a suspect log Kow estimation. When the experimental value is used, the evaluation would be that skin penetration of the substance is NOT favorable.

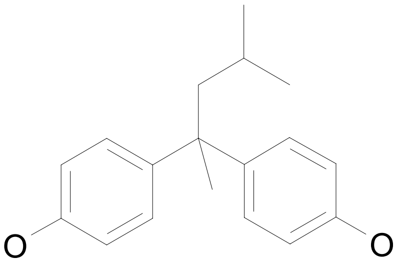
Available *in vitro* data	Result	No data available
	Reliability	
	Reasoning	
Available *in vivo* data	Result	No data available
	Reliability	
	Reasoning	

CONCLUSION

Weighted summary of the above presented data	Result	*Should the substance be classified for skin irritation or corrosion (R38 or R34/35)?*
	Reliability	*Please assess reliability (using Klimisch codes or other ranking?)*
	Reasoning	*Please give a reasoning of why a certain conclusion is reached, and why a certain reliability is given to this conclusion.*
Need for further testing	*Is further information needed to assess skin irritating properties? Information on physico-chemical parameters? Dermal absorption? In vitro testing? In vivo testing?*	

Skin Irritation example for: 4,4'-diisobutyl-ethylidenediphenol / Classification & Labelling

IDENTITY

Chemical Name (English)	4,4'-diisobutylethylidenediphenol
CAS RN	6807-17-6
EINECS/ELINCS-nr.	not in ESIS
SMILES	CC(c1ccc(cc1)O)(c2ccc(cc2)O)CC(C)C
Structure (2D):	
Molecular Weight	270.37 g/mol
Bruto Formula	$C_{18}H_{22}O_2$

PHYSICO-CHEMICAL PARAMETERS

Parameter	Value	Unit	Source
Melting point	148	°C	Estimate
Water Solubility	3.4	mg/l	Test
Log Kow	5.04		Estimate
Surface tension	62.4	mN/m	Test - confidential
Hydrolysis DT50	Unknown	Days	
pH (in water solubility test)	Unknown		
Lipid solubility	Unknown	mg/kg	
Others…			

N.B.:
Since the formats used for the examples were improved upon according to the discussion of the examples at the (Q)SAR WG meeting in October 2006, they are different from the versions in Annex 3 and 4.

QPRF ((Q)SAR Prediction Reporting Format) – DEREKfW

Prediction for substance	4,4’-diisobutylethylidenediphenol

MODEL

Model Name	DEREKfW8.0
(Q)SAR Model Reporting Format (QMRF)	QMRF – DEREK – Skin irritation.doc
Endpoint description	Skin Irritation (mammalian). Not necessarily strong enough to lead to classification (alert dependent)

PREDICTION

Result	No prediction

INFORMATION RELEVANT to the ASSESSMENT of the PREDICTION

Algorithm / Result interpretation	No structural alerts identified for this structure
Domain of applicability	Organic substances that contain at least one alert. The substance does not contain any structural alert for skin irritation, therefore the chemical is thought to be outside the applicability domain
Alerts/fragments identified	No alerts identified
Rules applicable	Not applicable
Structural analogues	Not applicable
Is the substance part of the training set?	Unknown / Not applicable
Model specific information on the prediction reliability	-

CONCLUSION

Result	No prediction
Reliability	4 (Klimisch code)
Reasoning	DEREKfW cannot give a prediction of skin irritation potential when no known structural alert is encountered in the substance. The substance is effectively outside the applicability domain.

QPRF ((Q)SAR Prediction Reporting Format) – GERNER Rules

Prediction for substance	4,4’-diisobutylethylidenediphenol

MODEL

Model Name	BfrR (Gerner) physico-chemical exclusion rules for skin irritation
(Q)SAR Model Reporting Format (QMRF)	QMRF – Skin irritation – BfR (Gerner) rules.doc
Endpoint description	NOT Classifying for EU C&L as R38 (irritant to skin) and/or R34/R35 (corrosive to skin)

PREDICTION

Result	NOT R38 (irritant to skin) or R34/35 (corrosive to skin)

INFORMATION RELEVANT to the ASSESSMENT of the PREDICTION

<table>
<tr><td>Algorithm / Result interpretation</td><td colspan="4">General algorithm of the exclusion rules:
IF (rule) THEN substance is NOT R38 and/or R34/45</td></tr>
<tr><td>Domain of applicability</td><td colspan="4">No organometallic compounds, Purity of the substance >95%. The substance is thought to be in the applicability domain.</td></tr>
<tr><td>Alerts/fragments identified</td><td colspan="4">Not applicable</td></tr>
<tr><td rowspan="3">Rules applicable:</td><td>Class</td><td>Rule</td><td>Result</td><td>Goodness of fit</td></tr>
<tr><td>C</td><td>melting point >55°C</td><td>NOT R34/35/38</td><td>128/130</td></tr>
<tr><td>C</td><td>surface tension >62 mN/m</td><td>NOT R34/35/38</td><td>94/95</td></tr>
<tr><td>Structural analogues</td><td colspan="4">Not given – no possibility to search training set for structural analogues.</td></tr>
<tr><td>Is the substance part of the training set?</td><td colspan="4">Not known, but not probable as the training set are EU new chemicals</td></tr>
<tr><td>Model specific information on the prediction reliability</td><td colspan="4">See the goodness-of-fit statistics under “Rules applicable”. However, rules based on melting point and surface tension did not cover 100% of all substances (false negatives allowed for in the training set). In an external evaluation and validation of the set of rules the melting point rules were judged to be insufficiently reliable, and the surface tension rules were evaluated to have a very limited data basis, and rules based on surface tension were not validated to the extent that the other rules have been.</td></tr>
</table>

CONCLUSION

Result	NOT R38 (irritant to skin) or R34/35 (corrosive to skin)
Reliability	Please assess reliability (using Klimisch codes or other ranking?)
Reasoning	The two rules on which absence of skin irritatin potential is based are considered weak [see QMRF]. Melting Point cut off values are not covering 100% of the skin irritants; two false negatives based on melting point rules were encountered in the model external validation. Surface Tension has a limited data basis, and has not been validated to the extent that other parameters have been.

WERF (Weigh of Evidence Reporting Format) Classification & Labelling

SUBSTANCE

WERF for substance:	4,4'-diisobutylethylidenediphenol,

ENDPOINT

Regulatory endpoint:	EU Classification and Labelling for dangerous substances and preparations: http://ecb.jrc.it/Legislation/1967L0548EC.htm

DATA – (Q)SARs, category approach, *in vivo* and *in vitro* test data

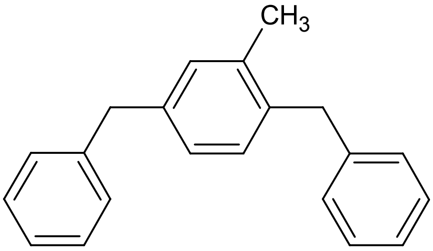
A substance with pH < 2 or pH > 11.5 should be considered corrosive [EU C&L guideline].	Result	Unknown
	Reliability	
	Reasoning	No data available. Extreme pH seems unlikely based on structure
Organic hydroperoxides should be assigned the risk phrase R34 [EU C&L guideline]	Result	Not applicable
	Reliability	
	Reasoning	Substance is not an organic hydroperoxide
Organic peroxides should be assigned the risk phrase R38 [EU C&L guideline]	Result	Not applicable
	Reliability	
	Reasoning	Substance is not an organic peroxide
Does the substance contain impurities (> 0.1%) that are known skin irritants or corrosives?	Result	Not applicable
	Reliability	
	Reasoning	No classification needed for impurities
Results of the Gerner exclusion rules for skin irritation:	Result	Not a skin irritant (NOT R38) and not a skin corrosive (NOT R34/35)
	Reliability	Use reliability from the QPRF
	Reasoning	The two applicable descriptors on which absence for skin irritation potential is based are considered weak
Results of the DEREKfW 8.0 prediction for skin irritation:	Result	No prediction
	Reliability	4 (Klimisch code)
	Reasoning	DEREKfW cannot give a prediction of skin irritation potential when no known structural alert is encountered in the substance. The substance is outside the applicability domain.
Other Models:	Result	Not available
	Reliability	
	Reasoning	
Available *in vitro* data	Result	No data available
	Reliability	
	Reasoning	
Available *in vivo* data	Result	No data available
	Reliability	
	Reasoning	

CONCLUSION

Weighted summary of the presented data	Result	*Should the substance be classified for skin irritation or corrosion (R38 or R34/35)?*
	Reliability	*Please assess reliability (using Klimisch codes or other ranking?)*
	Reasoning	pH, chemical class and puritiy of the substance do not require classification. No structural alert is identified (according to DEREKfW) which would lead to classification. Physico-chemical properties of the substance indicate absence of skin irritation potential (BfR/Gerner rules) but the applicable rules are considered weak.
Need for further testing?	*Is further information needed to assess the skin irritation properties?* *> Physico-chemical or related to model input* *> In vitro testing* *> In vivo testing*	

Biodegradation example for: dibenzyltoluene / PBT Assessment

IDENTITY

Chemical Name (English)	Dibenzyltoluene
CAS RN	26898-17-9
EINECS/ELINCS-nr.	248-097-0
SMILES	Cc1cc(ccc1Cc2ccccc2)Cc3ccccc3
Structure (2D):	 SMILES & STRUCTURE provided by INERIS, Fr.
Molecular Weight	272.38 g/mol
Bruto Formula	C21H20

PHYSICO-CHEMICAL PARAMETERS

Parameter	Value	Unit	Source
-	-	-	-

N.B.:
Since the formats used for the examples were improved upon according to the discussion of the examples at the (Q)SAR WG meeting in October 2006, they are different from the versions in Annex 3 and 4.

QPRF ((Q)SAR Prediction Reporting Format) – BIOWIN 1&2

Adapted from the example provided by INERIS, France

Prediction for Substance	Dibenzyltoluene
Model Name	BIOWIN 1+2 (v4.02)
Endpoint description	Rapid aerobic biodegradation probability
(Q)SAR Model Reporting Format (QMRF)	QMRF – Biodegradation – Biowin1-6.doc

PREDICTION

Biowin 1 Biowin 2	1.0381 0.9887

INFORMATION RELEVANT for the ASSESSMENT of the PREDICTION

Algorithm / Result interpretation	If the result of the probability is greater or equal to 0.5 then the substance biodegrades fast. If the result of the probability is less than 0.5 then the substance does not biodegrade fast
Domain of applicability	The substances is an organic non-ionizing substance, MW = 272.39, 2 molecular fragments are identified. The substance is thought to be in the applicability domain of the model
Alerts/fragments identified	3 x Alkyl substituent on aromatic ring 2 x unsubstituted phenyl group (C6H5-)
Rules applicable	Not applicable
Structural analogues identified?	No analogues are indicated by the model.
Is the substance part of training set?	The substance is not part of the training set.
model specific information on the prediction reliability[1]:	The result is not in the range of 0.4-0.6 where predictions are thought to be less reliable. All fragments of the molecule are used to derive the estimation. The frequency of appearance of fragments in the training set for the “Alkyl substituent on aromatic ring” is 36 (or 36/295 = 12.2%), and for “unsubstituted phenyl group (C6H5-)” it is 25 (or 25/295 = 8.5%).

CONCLUSION

Result	Biodegrades fast
Reliability	2 (INERIS). *Give your idea of a reliability of this prediction. Please also use and comment on the flow chart developed by INERIS, France, provided on the last page.*
Reasoning	The result is not in the range 0.4 – 0.6, the molecular fragments used for the prediction describe the whole molecule, and their frequency of appearance in the training set is acceptable. Therefore, the result can be considered reliable. As the result is an estimation obtained by a model, it is considered as reliable with restrictions.

[1] Frequency of appearance of fragments used in the training set should be > 10 (3.4%) for reliable predictions. This is a default value that should be discussed and agreed before being used.

QPRF ((Q)SAR Prediction Reporting Format) – BIOWIN 3&4

Adapted from the example provided by INERIS, France

Prediction for Substance	Dibenzyltoluene
Model Name	BIOWIN 3+4 (v4.02)
Endpoint description	Expert estimated half life for ultimate (Biowin3) and primary (Biowin4) biodegradation
(Q)SAR Model Reporting Format (QMRF)	QMRF – Biodegradation – Biowin1-6.doc

PREDICTION

Biowin 3 Biowin 4	2.4167 3.2589

INFORMATION RELEVANT for the ASSESSMENT of the PREDICTION

Algorithm / Result interpretation	The result provided is a predicted rating which can be converted to time required to achieve ultimate or primary biodegradation as follows: 5.00 → hours; 4.00 → days; 3.00 → weeks; 2.00 → months; 1.00 → longer
Domain of applicability	The substances is an organic non-ionizing substance, MW = 272.39, 2 molecular fragments are identified. The substance is thought to be in the applicability domain of the model
Alerts/fragments identified	3 x Alkyl substituent on aromatic ring 2 x unsubstituted phenyl group (C6H5-)
Rules applicable	Not applicable
Structural analogues identified?	No analogues are indicated by the model.
Is the substance part of training set?	The substance is not part of the training set.
model specific information on the prediction reliability[2]:	The result is not in the range of 0.4-0.6 where predictions are thought to be less reliable. All fragments of the molecule are used to derive the estimation. The frequency of appearance of fragments in the training set for the “Alkyl substituent on aromatic ring” is 36 (36 (or 36/200 = 18%), and for “unsubstituted phenyl group (C6H5-)” it is 25 (or 22/200 = 11%).

CONCLUSION

Result	Ultimate biodegradation in weeks to months
Reliability	2 (INERIS). *Give your idea of a reliability of this prediction. Please also use and comment on the flow chart developed by INERIS, France.*
Reasoning	The result is not in the range 0.4 – 0.6, the molecular fragments used for the prediction describe the whole molecule, and their frequency of appearance in the training set is acceptable. Therefore, the result can be considered reliable. As the result is an estimation obtained by a model, it is considered as reliable with restrictions.

[2] Frequency of appearance of fragments used in the training set should be > 10 (3.4%) for reliable predictions. This is a default value that should be discussed and agreed before being used.

QPRF ((Q)SAR Prediction Reporting Format) – BIOWIN 5&6

Adapted from the example provided by INERIS, France

Prediction for Substance	Dibenzyltoluene
Model Name	BIOWIN 5+6 (v4.02)
Endpoint description	Ready biodegradation probability
(Q)SAR Model Reporting Format (QMRF)	QMRF – Biodegradation – Biowin1-6.doc

PREDICTION

Biowin 5	-0.0613
Biowin 6	0.0237

INFORMATION RELEVANT for the ASSESSMENT of the PREDICTION

Algorithm / Result interpretation	If the result of the probability is greater or equal to 0.5 then the substance is readily biodegradable (RB). If the result of the probability is less than 0.5 then the substance is not readily biodegradable (NRB)
Domain of applicability	The substances is an organic non-ionizing substance, MW = 272.39, 2 molecular fragments are identified. The substance is thought to be in the applicability domain of the model
Alerts/fragments identified	1 x aromatic -CH3 2 x aromatic –CH2 13 x aromatic -H
Rules applicable	Not applicable
Structural analogues identified?	No analogues are indicated by the model.
Is the substance part of training set?	The substance is not part of the training set.
model specific information on the prediction reliability[3]:	The result is not in the range of 0.4-0.6 where predictions are thought to be less reliable. All fragments of the molecule are used to derive the estimation. The frequency of appearance of fragments in the training set or the aromatic –CH3 49 (or 49/589 = 8.3%), for the aromatic –CH2 it is 32 (or 32/589 = 5.4%), and for the aromatic –H it is 302 (or 302/589 = 51.3%).

CONCLUSION

Result	The substance is not readily biodegradable
Reliability	2 (INERIS). Give your idea of a reliability of this prediction. Please also use and comment on the flow chart developed by INERIS, France, provided on the last page.
Reasoning	The result is not in the range 0.4 – 0.6, the molecular fragments used for the prediction describe the whole molecule, their frequency of appearance in the training set is acceptable. Therefore, the result can be considered reliable. As the result is an estimation obtained by a model, it is considered as reliable with restrictions.

[3] Frequency of appearance of fragments used in the training set should be > 10 (3.4%) for reliable predictions. This is a default value that should be discussed and agreed before being used.

Weight-of-Evidence Reporting Format - PBT Assessment

SUBSTANCE

WERF for substance:	dibenzyltoluene

ENDPOINT

Regulatory endpoint:	PBT assessment from EU TGD on Risk Assessment. The use of three out of the six models is suggested as follows: Substance = P IF (Biowin 2 OR Biowin 6 < 0.5) AND (Biowin 3 < 2.2)

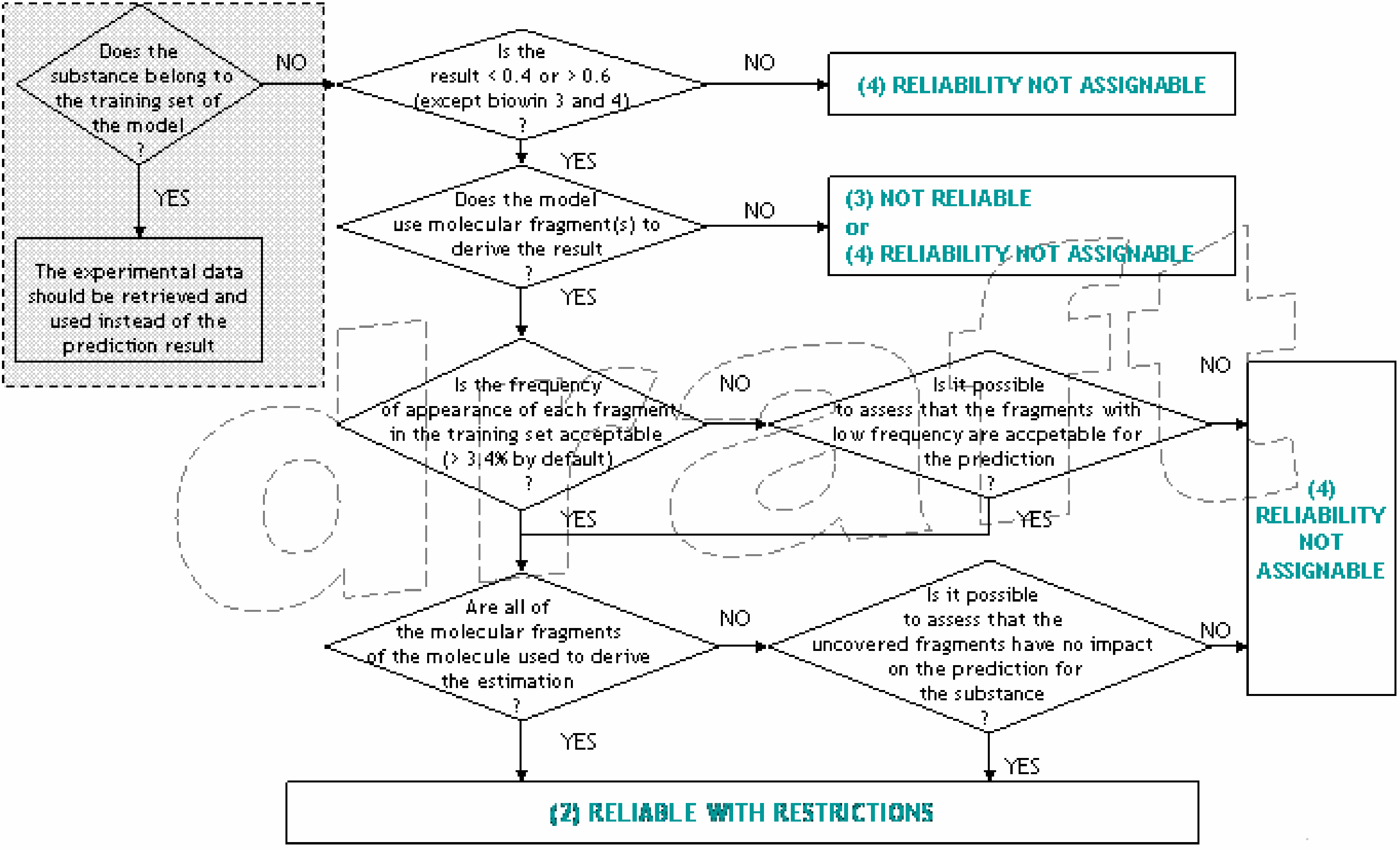
DATA – (Q)SARs, category approach, *in vivo* and *in vitro* test data)

Biowin 1 & 2	Result	Biodegrades Fast
	Reliability	Use reliability from the QPRF
	Reasoning	All reliability criteria of the model are met (frequency of fragment, coverage of molecule, prediction domain > 0.6, models 1&2 are in agreement)
Biowin 3 & 4	Result	2.42: Ultimate biodegradation in weeks to months
	Reliability	Use reliability from the QPRF
	Reasoning	All reliability criteria of the model are met (frequency of fragment, coverage of molecule, prediction domain > 0.6)
Biowin 5 & 6	Result	Not Readily Biodegradable
	Reliability	Use reliability from the QPRF
	Reasoning	All reliability criteria of the model are met (frequency of fragment, coverage of molecule, prediction domain > 0.6, models 5&6 are in agreement)
Other Models:	Result	Not available
	Reliability	
	Reasoning	
Available experimental data	Result	No data available
	Reliability	
	Reasoning	

CONCLUSION

Weighted summary of the presented data	Result	*Is this substance Persistent or Not Persistent?*
	Reliability	*Indicate the reliabity of the result*
	Reasoning	*Provide a reasoning for the result*
Need for further testing?	*Can you make an assessment on the P criterion for PBT based on the model information provided? More information needed? More model data? More input (e.g. physico-chemical data? What further testing is needed? If biodegradation tests are required what is needed to come to a definitive conclusion (ready biodegradability test, inherent testing, simulation testing)?*	

Reliability scores assignement to prediction results of BIOWIN models



Does the substance belong to the training set of the model ?

YES: The experimental data should be retrieved and used instead of the prediction result

NO: Is the result < 0.4 or > 0.6 (except biowin 3 and 4) ?

NO: (4) RELIABILITY NOT ASSIGNABLE

YES: Does the model use molecular fragment(s) to derive the result ?

NO: (3) NOT RELIABLE or (4) RELIABILITY NOT ASSIGNABLE

YES: Is the frequency of appearance of each fragment in the training set acceptable (> 3.4% by default) ?

NO: Is it possible to assess that the fragments with low frequency are accpetable for the prediction ?

NO: (4) RELIABILITY NOT ASSIGNABLE

YES: Are all of the molecular fragments of the molecule used to derive the estimation ?

NO: Is it possible to assess that the uncovered fragments have no impact on the prediction for the substance ?

NO: (4) RELIABILITY NOT ASSIGNABLE

YES: (2) RELIABLE WITH RESTRICTIONS

INERIS

Appendix 6. Poster, SETAC Europe 2007

This poster was presented at SETAC Europe, 21-25 May 2007, Porto, Portugal, and summarizes the concept of using reporting formats for alternative testing data at three different levels, in order to satisfy the documentation requirements as mentioned in Annex XI of the REACH legislation.

E. Rorije – RIVM/Expertise Centre for Substances (SEC), Bilthoven, The Netherlands
E. Hulzebos – RIVM/SEC
B. Hakkert – RIVM/SEC

Reporting Alternative Testing Information under REACH

Introduction

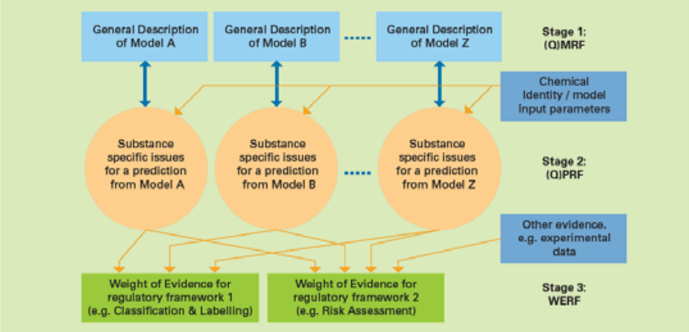
In 2004 RIVM initiated a QSAR Experience project, aimed at regulators gaining experience in the application of (Q)SARs, and communicating results. From the discussions in this project it became clear that interpretation of (Q)SAR results differed greatly, and reasoning was often difficult to follow because 3 main issues (model validity, model applicability and result, and adequacy for a (regulatory) purpose) were confused. This called for a (standardized) reporting format. Within the (Q)SAR Experience Project a 3-stage reporting procedure was proposed by RIVM and worked out together with ECB with support from the EU Working Group on (Q)SARs in 2006. By separating these three issues the proposed formats follow closely the information requirements in the REACH text:

- REACH Annex XI – General Rules for Adaptation of the Standard Testing Regime, section 1.3 on the use of (Q)SARs requires that:
  - The **scientific validity** of the model has been established,
  - A substance falls within the **applicability** domain **of the model**,
  - Results are **adequate for** the **purpose** of classification and labeling and/or risk assessment, and
  - **Adequate and reliable documentation** of the applied method is provided.

Furthermore the Agency in collaboration with the commission, Member States and interested parties shall **develop and provide guidance** in assessing which (Q)SARs will meet these conditions **and provide examples**.

Results

Figure 1: Three stages in the adequate and reliable documentation on alternatives to standard testing. Stage 1: model validity, using the (Q)MRF, Stage 2: model applicability, using the (Q)PRF, and Stage 3: adequacy of the result (for regulatory purpose), using the WERF.


General Description of Model A
General Description of Model B
General Description of Model Z
Stage 1: (Q)MRF
Chemical Identity / model Input parameters
Substance specific issues for a prediction from Model A
Substance specific issues for a prediction from Model B
Substance specific issues for a prediction from Model Z
Stage 2: (Q)PRF
Other evidence, e.g. experimental data
Weight of Evidence for regulatory framework 1 (e.g. Classification & Labelling)
Weight of Evidence for regulatory framework 2 (e.g. Risk Assessment)
Stage 3: WERF

(Q)MRF = (Q)SAR Model Reporting Format
(Q)PRF = (Q)SAR Prediction Reporting Format
WERF = Weight of Evidence Reporting Format

Figure 2. Fulfilling the three information requirements from the REACH Annex XI text, and properly documenting this, should give results of alternatives for standard testing that may be used instead of (standard) testing.

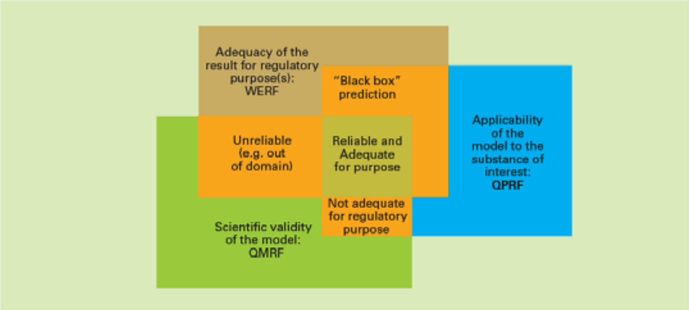

Adequacy of the result for regulatory purpose(s): WERF
"Black box" prediction
Applicability of the model to the substance of interest: QPRF
Unreliable (e.g. out of domain)
Reliable and Adequate for purpose
Not adequate for regulatory purpose
Scientific validity of the model: QMRF

Table: Indication (in the form of headings) of the information required in the three different reporting formats, and the conclusions drawn from that information.

	(Q)MRF	(Q)PRF	WERF
Information Requirements	Model Identifier	Identifier, refer to (Q)MRF	Results, refer to the (Q)PRFs
	Endpoint	Model result	Regulatory purpose
	Algorithm	Interpretation of result	Weighting of each result
	Applicability domain (general definition)	Applicability domain (specific for substance)	Adequacy for regulatory purpose
	In-/external validation	Structural analogues (from training set)	
	Mechanistic interpretation		
Evaluation	No Conclusion	Conclusion on reliability (refer to QMRF) and applicability	Conclusion for regulatory endpoint using all evidence
	No judgment of the model, only information on the validation status	No judgment of adequacy for specific regulatory purpose	Further testing proposal.
Archiving/ availability	Central Repository of models descriptions: i.e.ECB (Q)SAR Inventory	One (Q)PRF for each prediction – in IUCLID	One WERF for each endpoint/regulatory purpose – in the dossier
Compare to:	**(central repository for) OECD testing guidelines**	**Robust Study Summary**	**Single endpoint in Risk Assessment Report**

Discussion and Conclusions

- 3 formats for reporting (Q)SAR results follow closely 3 main issues in REACH Annex XI section 1.3, and form **guidance** to what adequate and reliable documentation should contain.
- Discussion of a (Q)SAR result is simplified, by separating issues on 1) **validity**, 2) **applicability**, and 3) **adequacy** for regulatory purpose.
- Applying the reporting formats and discussing results as part of the QSAR Experience Project was *essential* for evaluating the information content of the formats, and recognizing omissions and problems in the formats.
- Similar issues are relevant, and mentioned in the REACH text for *in vitro* **methods** (Annex XI section 1.4), **grouping of substances and read-across approach** (section 1.5), and also **the use of existing** (non-guideline and/or non GLP) **testing data** (section 1.2). It would therefore be logical to follow a similar approach for reporting validity and applicability of the method(s) ((Q)MRF and the (Q)PRF).
- Results from alternatives for standard testing (not only (Q)SAR) reported using (Q)MRF/(Q)PRF can be easily used in a Weight of Evidence Approach as advocated by REACH, e.g. by using the Weight of Evidence Reporting Format (WERF).

Acknowlegdments

To the European Chemicals Bureau (ECB) – for work on QMRF and the ECB (Q)SAR Inventory (http://ecb.jrc.it/qsar/qsar-tools/qsar_tools_qrf.php), discussion and on QPRF and WERF, and proposing the formats in the REACH Implementation Project guidance documents.

070717 MEV/SEC FvdB

Appendix 7. The (Q)SAR Experience Project 2005

Prepared by H. Verhaar

Henk Verhaar[¥], Betty Hakkert[*], Etje Hulzebos[*], Emiel Rorije[*], Henrik Tyle[†], Andrew Worth[‡], Tatiana Netzeva[‡], Manuela Pavan[‡], Jan Ahlers[§], Guido Jacobs[#], Sylvain Bintein[$], Jos Bessems[&].

[¥]ENVIRON Netherlands B.V., Zeist, the Netherlands
[*]National Institute for Public Health and Environment (RIVM), Bilthoven, the Netherlands
[†]Miljøstyrelsen (Danish EPA), Copenhagen, Denmark
[‡]European Chemicals Bureau, EU Joint Research Centre, Ispra, Italy
[§]UmweltBundesAmt, Berlin, Germany
[#]Wetenschappelijk Instituut Volksgezondheid, Brussel, België
[$]L'Institut National de l'environnement industriel et des risques (INERIS), Frankrijk
[&]TNO Kwaliteit van Leven, Zeist, the Netherlands

Abstract

The impending implementation of REACH, the proposed European regulation for the Registration, Evaluation, Authorisation and restriction of CHemicals, is expected to greatly increase the regulatory use of (Q)SARs. A major barrier to the use of (Q)SARs is lack of experience with and confidence in (Q)SAR among the EU Competent Authorities. In 2004, a so-called (Q)SAR experience project was launched, with participation of representatives of many EU member states. This project is aimed at providing EU regulatory authorities with hands-on experience in the generation of (Q)SAR predictions for chemical substances, and the appreciation of their usefulness. The first phase of this project entailed the generation of (Q)SAR predictions for three relevant hazard end points, viz. ready biodegradability, acute toxicity to fish, and mutagenicity, for 177 so-called SIDS substances. These predictions were then compared to experimental values included in the SIDS database, as well as estimates generated by the Danish EPA in an OECD-related (Q)SAR activity.

For ready biodegradability, predictions were generated with the US-EPA's BIOWIN models. The results were in agreement with the Danish EPA's estimates. It appears that ready biodegradability is an end point that is quite suitable to (Q)SAR prediction; if the end point is changed from ready biodegradability to persistence in the environment, the suitability of a (Q)SAR approach even increases. Apart from the above indicated experience some conclusions from the discussions were: the accuracy of a (Q)SAR can not be higher than that of experimental results; the OECD (Q)SAR principles are a useful tool, but should not be implemented too strictly.

For acute toxicity to fish, several published (Q)SAR models, two expert systems (US-EPA's ECOSAR and Fraunhofer's PropertEst), and one commercial system were used. As is well-known and understood, (Q)SAR predictions work quite well for so-called baseline toxicity, or narcosis-type substances. Important issues are the scope of a (Q)SAR/end point (is an end point species specific or is a fish test/(Q)SAR applicable –within reasonable limits– to all fish); the classification of substances into mode-of-action classes (based on substructures, tests, or expert knowledge; rough or detailed); and what approach to take in providing predictions for non-baseline substances (excess toxicity approach or class-specific (Q)SARs).

For mutagenicity, predictions were generated with a commercial system, viz. DEREKfW, and

the published Ashby & Tennant structural alerts model. Again, the results are in very good agreement with the Danish EPA (Q)SAR results. Major conclusions were that different test and models are based on different end points (Ames test vs carcinogenicity); a better guidance is needed on how to apply (Q)SARs to real unknowns; good models for human toxicology end points are few, and are difficult or expensive to obtain; (Q)SAR should be part of an intelligent testing strategy rather than be seen as an isolated tool; and no evaluation system can be developed without taking into account expert knowledge.

Overall, the (Q)SAR experience project generated an increased understanding and awareness of (Q)SAR among the participants, some of which did not have any prior experience with (Q)SAR at all.

1 Introduction

1.1 REACH and (Q)SAR

The impending implementation of REACH, the proposed European regulation for the Registration, Evaluation, Authorisation and restriction of CHemicals, will greatly increase the regulatory use of (Q)SARs in the foreseeable future. In particular, there is policy to reduce experimental testing on animals by promoting the use of alternative test and computational methods. At present, however, a barrier to the acceptance and widespread use of (Q)SARs is the lack of experience with these approaches among the Competent Authorities that will eventually need to evaluate chemical substances under REACH, taking into account the registration of chemicals based partially on (Q)SAR estimates.

1.2 The Danish EPA (Q)SAR project

Since ca. 1999, the Danish EPA has been selecting (Q)SAR models for the prediction of hazard properties of chemical substances. At present, this has resulted in a selection of ca 50 models. Using these models, predictions of relevant hazard properties have been generated for around 166.000 discrete organic chemicals for around 50 different end points. These predictions are stored in a database. These predictions have been used internally for a.o. screening for POP and PBT substances, or for classification of untested chemicals.

Within the OECD HPV program, Denmark applied their (Q)SAR predictions for three selected end points to 184 discrete organic chemicals (‘SIDS[4]’ chemicals) discussed at SIAM 11 to SIAM 18. The predictions were compared to the experimental data that had been submitted for these SIDS chemicals. The end points chosen were: ready biodegradability, acute toxicity to fish, and mutagenicity in the Ames test. The reason for choosing these end points was the availability of enough experimental data as well as robust models to enable a (statistically) meaningful comparison. It should be kept in mind that the purpose of the Danish study explicitly was not the formal validation of the (Q)SAR models, but rather to get a real-world impression of (Q)SAR performance. Even so, the outcome of the comparison of the (Q)SAR predictions with the internationally accepted test data suggested that these or similar models may have a potential for increased use in the future.

1.3 The (Q)SAR experience project

[4] The Screening Information Data Set (SIDS) is the minimum amount of data that is required for making an initial hazard assessment of HPV chemicals which has been agreed upon by OECD for their High Production Volume Chemicals program. The ‘SIDS’ substances in the Danish EPA database are substances for which these SIDS hazard data are available.

Because of this perceived potential for increased use of (Q)SAR, especially taking into account the emphasis that REACH, in addition to other EU legislation, places on the use of alternative methods for determining the hazard profile of chemical substances, the Dutch Chemical Substances Bureau ('RIVM') foresaw a need to promote experience with and confidence in the use of (Q)SAR models and results. To that end, a project was proposed aimed at providing experience in applying and evaluating (Q)SAR. Following a meeting held in Den Dolder (NL) on 26-27 October, 2004, a project proposal was developed jointly by the European Chemicals Burea ('ECB'), RIVM and Danish EPA. Phase 1 of this (Q)SAR Experience Project was carried out in the period January-May 2005. This first phase basically consisted of an extension of the Danish (Q)SAR effort for 184 SIDS substances, using alternative models to those used by Denmark. To provide 'hands-on' and 'eyes-on' experience to as many participants as possible, all participants were provided with the relevant information on the 184 SIDS substances, and a limited number of models for the three end points mentioned above. Participants were free to individually add additional models to their effort. Participants in the project included the ECB, a broad participation of the CAs, and some additional organisations, formalized as a part of the subgroup on (Q)SARs of the EU Technical Committee on New and Existing Substances ('TC-NES'). The effort was concluded by a plenary meeting in May 2005, in which all results were thoroughly discussed.

2 Materials and Methods

2.1 SIDS data

2.1.1 Work Package 1

In the OECD HPV Chemicals program, data sets for hazard data for a selected set of end points (the so-called 'SIDS' data) are assembled from open and proprietary sources, including EU IUCLID data files. These data sets are evaluated and endorsed at the SIAM meetings, and consolidated into a SIDS Initial Assessment Report ('SIAR').

The Danish EPA effort retrieved SIDS data from the SIAPs, and checked these against the studies' robust summaries, where available. In order to capture any typos and other copying errors, participants in the (Q)SAR experience project rechecked all SIDS data for the three selected end points to the original reports. Moreover, the SIDS data were also compared to available IUCLID data sets, wherever available.

Both the Danish EPA effort and the (Q)SAR experience project compared the available SIDS data to the training set data for the individual models.

2.1.2 Work Package 2

In Work Package 2, predictions for 10 (SIDS) substances were studied in more detail, noting were application of models would present problems to operators, and were and why models may break down. The selected substances are presented in table 1.

One of the primary questions for Work Package 2 was whether it is at all possible to judge the quality of individual substance predictions. To that end, participants were asked to provide reasoned reliability scores for individual predictions.

2.2 Models and Model Choice

The Danish EPA effort used the following (Q)SAR models:

- Biodegradation
  - BIOWIN models 1, 2, 3, 5, 6 from the US-EPA's EPISuite™ package (ref); the training sets for these models is reported in the EPISuite™ help file
  - An custom MultiCASE (ref) model developed in-house by the Danish EPA; the training set consisted of MITI data augmented by additional literature data.

Table 1

Biodegradability		Fish acute toxicity		Mutagenicity	
CAS	**name**	**CAS**	**name**	**CAS**	**name**
78-59-1	isophorone	100-41-4	ethylbenzene	50-00-0	Formaldehyde
88-12-0	1-vinyl-2-pyrrolidone	107-98-2	1-methoxy-2-propanol	75-38-7	1,1-difluoroethylene
99-54-7	1,2-dichloro-4-nitrobenzene	123-86-4	acetic acid butyl ester	98-59-9	4-methylbenzenesulfonyl chloride
107-41-5	Hexylene glycol	79-20-9	acetic acid methyl ester	106-88-7	1,2-epoxybutane
127-19-5	DMAC	96-18-4	1,2,3-trichloro-propane	25321-14-6	dinitrotoluene
288-32-4	Imidazole	108-88-3	methyl-benzene	78-59-1	Isophorone
505-32-8	Isophytol	78-87-5	1,2-dichloro-propane	140-88-5	Ethyl acrylate
839-90-7	Tris(2 hydroxyethyl) isocyanurate	770-35-4	1-phenoxy-2-propanol	1163-19-5	Bis(pentabromodiphenyl)ether
1477-55-0	1,3-bis (aminomethyl) benzene	95-50-1	1,2-dichloro-benzene	123-54-6	2,4-pentadione
6864-37-5	cyclo-hexamine	1490-04-6	5-methyl-2-(1-methylethyl)cyclohexanol	98-92-0	Nicotinamide

- Acute toxicity to fish
  - A linear model for non-polar narcosis based on log Kow, as recommended by the TGD.
  - A non-linear model for multiple chemical classes based on MultiCASE, developed in-house by the Danish EPA; the training set consisted of a custom selection of fathead minnow LC50 data
- Mutagenicity
  - Commercial models included in the TOPKAT and MultiCASE packages.
  - A custom MultiCASE model developed in-house by the Danish EPA; the training set was not given in the report.

For the (Q)SAR experience project, several considerations were brought to bear on model selection, *viz*. ready availability, ease of operation (for experience building), transparency, and comprehensiveness of applicability, as well as the possibility of comparing results with the Danish EPA effort. A particular problem in this respect was the availability of commercial (Q)SAR packages, since no budget was available within the (Q)SAR experience project for acquiring expensive proprietary (Q)SAR packages. Several suppliers of commercial (Q)SAR packages were approached with a request for co-operation. LHASA UK was the only supplier to provide the project with a copy of their DEREK for Windows package, under the constraint that no individual predictions were to be published. DEREK for Windows was used for predicting mutagenicity.

The (Q)SAR experience project used the following (Q)SAR models:

- Biodegradation
  - The US EPA EPISuite™'s BIOWIN 1–3 & 5–6 models {U.S. Environmental Protection Agency, 2000, EPI Suite™}; the training sets for these models is reported in the EPISuite™ help file

- Acute toxicity to fish
    - Two linear log Kow–based models for non-polar and polar narcosis, recommended in the TGD {European Commission, 1996, Technical guidance document in support of Commission Directive 93/67/EEC on…}; the training sets used for these models were the training sets as originally selected by the authors of the (Q)SAR. ECB recalculated the (Q)SARs using updated log Kow values.
        - It should be noted here that the Kow-values used were obtained from multiple sources, and included both experimental and calculated Kow-values.
    - A linear log Kow–based model for narcosis in general, developed at the EU Joint Research Centre's European Chemicals Bureau (ref); the training set used for this model consisted of the combined training sets of the two models mentioned above.
        - It should be noted here that the Kow-values used were obtained from multiple sources, and included both experimental and calculated Kow-values.
    - The US EPA EPISuite™'s {U.S. Environmental Protection Agency, 2000, EPI Suite™} model for non-polar narcosis; training set described in the model's documentation.
    - The PropertEst (ref) model for non-polar narcosis; training set described in the model's documentation.
- Mutagenicity
    - DEREK for Windows' {LHASA Ltd, DEREK for Windows} model for activity in the Ames test;
    - The published structural alert system for carcinogenicity developed by Ashby and Tennant {Ashby and Tennant, 1991, Mutation Research, 257, 229-306};

In Work Package 1, the models were checked against the OECD principles for (Q)SAR models (ref), the general interoperability of the models was evaluated, and the (overall) results were compared to the results obtained in the Danish EPA's effort.

2.2.1 Applicability domain

The applicability domain of the acute fish toxicity models is defined as a mode of action, viz. non-polar or polar narcosis, together with a log Kow-range. A comprehensive overview of this mode of action and the associated chemical applicability domain is given a.o. by {Verhaar et al., 1992, Chemosphere, 25, 471-491}. Selection of SIDS substances within the applicability domain of the linear log Kow-based models was done by two automated substructure-recognition based procedures, viz. the US-EPA ASTER selection model, and a model developed by the EU-ECB, as well as by a panel of experts. Selection for the EPISuite™ models was done by the substructure-recognition model included in EPISuite™, which is quite similar to the ASTER model.

No explicit information on the applicability domain of the BIOWIN models is available. Therefore, no initial selection of SIDS substances within such a domain was done. For the mutagenicity models, either no pertinent information on the applicability domain is available, or the models used will automatically flag substances outside their domain. In either case, no initial selection of SIDS substances was done.

2.3 Statistics

Note that neither the Danish EPA effort nor the (Q)SAR experience project was meant to be a validation exercise. Statistics were therefore not used to determine the ultimate performance and quality of models, but only as a tool for understanding model results and thereby gaining experience and confidence in (Q)SAR modelling.

2.3.1 Concordances

For the qualitative end points (yes/no biodegradable; yes/no mutagenic), model performance was expressed in terms of true and false positives and negatives, with the following metrics:

- Sensitivity; TP/(TP + FN)
- Specificity; TN/(TN + FP)
- Concordance; (TN + TP)/(TP + FP + TN + FN)
- Positive Predictive Value; TP/(TP + FP)
- Negative Predictive Value; TN/(TN + FN)

No attempt was made to differentiate between different levels of prediction (such as differentiating e.g. between 0.1 and 0.4 on a scale of 0–1 with 0.5 as a cut-off point), other than recognizing some experimental results, as well as some predictions, as equivocal results.

2.3.2 Quantitative statistics

For the quantitative end points, the model performance was expressed as Goodness of Fit parameters (coefficient of determination, standard error of the estimate, Fisher and Friedman statistics, standard deviation error in calculation, Akaike information criterion and the Kubinyi statistic), internal cross-validation (leave-one-out, bootstrapping, Y-scrambling, as well as leverage and outlier detection. Outliers were defined as objects with a residual ≥ 2*$SE_{prediction}$.

Leverage is a measure of the ‘width’ and especially the uniformity of the descriptor space, with objects that are widely separated from the bulk of the objects in descriptor space having a high leverage (i.e. potential influence) on the model[5]. Leverage is defined as the diagonal of the regression model’s ‘hat matrix’ $X(X'X)^{-1}X'$. High leverage was defined as a leverage ≥ 3p/n, with p being the number of descriptors in the model including the constant, and n being the number of objects in the model.[6]

Leverage and residuals were plotted together, to yield a so-called William’s plot.

3 Results

3.1 DK

The results from the Danish EPA effort are reported in OECD report ENV/JM/TG(2004)26/REV1.

3.2 Ready Biodegradability

3.2.1 Work Package 1

All predictions were done with the suite of BIOWIN models (BIOWIN 1–3 & 5–6) available in the EPISuite package of programs {U.S. Environmental Protection Agency, 2000, EPI Suite™}. These models differ in their training set, algorithm used, as well as the (ready) biodegradability test that the biodegradability judgement is based upon. The experimental data that the predictions were checked against are validated (SIDS) data.

3.2.1.a OECD Principles

The available BIOWIN models partly comply with the OECD principles on (Q)SAR. It should

[5] It should be noted here that only high-leverage objects that are also outliers have a high influence on the model.

[6] Or differently, if X is a matrix with dimensions a,b (a rows, b columns), then n = a and p = b + 1.

be noted that while BIOWIN 5 and 6 have roughly the same end point as the SIDS experimental data (i.e. ready biodegradability), the other models are based on different ready biodegradability ‘tests’; moreover, there is no clear mechanistic underpinning of any of the models. Moreover, concerning the applicability domain, the models do not exclude substances that do not contain recognized substructure. In such cases, the model defaults to a molecular weight-based prediction. In fact the models have no definitive applicability domain at all. It was recently argued in an article by Tunkel et al. that in the case of the BIOWIN models this is not a major concern.

3.2.1.b Comparison between SIDS data and model training sets

A lot of the SIDS substances (ca 25% on average) are part of one or more BIOWIN model training sets; this means that including these substances in a so-called test set will give biased results. On the other hand, the SIDS data on ready biodegradability for these substances may come from different sources than the data used in the model training set. It was observed that for a number of these substances the SIDS data were not in agreement with the BIOWIN training set data, i.e. substances were classified as readily degradable in the SIDS data set but as not readily degradable in the training set data, or vice versa. This observation highlights the fact that experimental data are not (always) unequivocal or without variability or even uncertainty. Incidentally, the BIOWIN models do not flag input as being part of their training set; one has to manually check the documentation for that.

3.2.1.c Prediction results

Neither the SIDS data nor the BIOWIN models flag substances for which standard biodegradability data may be less than reliable; such as highly volatile substances, or rapidly hydrolyzing substances. Such substances are among the substances for which experimental results are equivocal or unreliable, or both. Exclusion of these substances resulted in a reduced test set of 170 SIDS substances.

Substances for which a BIOWIN model did not recognize any substructure, and therefore were predicted based on a correlation with molecular weight alone, were excluded from the test set.

Table 2: Statistical results for the 5 different BIOWIN models

	BIOWIN 1	BIOWIN 2	BIOWIN 3	BIOWIN 5	BIOWIN 6
Equivocals	19				21
Sensitivity	50 (48)	58	60-61	70	80 (79)
Specificity	93 (93)	87	86	82-81	81-80 (80)
Concordance	72 (70)	73	73	76	80 (79)
PPV	87 (87)	81	80	79-78	80-79 (80)
NPV	66 (64)	68	69-70	74	81 (79)

Table 2 presents the overall statistical results for the 5 individual BIOWIN models when comparing predicted with recorded ready biodegradability for the reduced SIDS data set, training sets included. The results are nearly identical to the results obtained in the Danish EPA effort. Values in brackets indicate the results for a test set with the equivocal predictions excluded. There is no guidance within the BIOWIN models, or external, on how to evaluate numerical degradability predictions in the range of 0.4–0.6, or close to the cut-off value of 0.5. Are these values to be regarded as accurate or equivocal, or discarded as unreliable? From the results presented in table 2 it can be provisionally concluded that equivocal prediction results do not significantly degrade the overall performance of the models, although (results not shown) the predictions for these particular substances are significantly less reliable than the overall model performance.

Overall, it can be concluded that BIOWIN 6, which incidentally is a model that was developed from experimental (MITI) ready biodegradability data, shows the best as well as the most consistent behaviour on the SIDS data, with about 80% correct results.

3.2.2 Work Package 2

For Work Package 2, ready biodegradability predictions for 10 SIDS substances were evaluated in more detail. The selected substances are presented in table 1. Participants involved in providing hands-on experience for biodegradability (Q)SARs were asked to give their personal judgement on the reliability of each prediction.

Interestingly, there was a consistent difference in reliability scores between the different participants, where 1 participant used the overall statistics result as a model reliability parameter, whereas 2 other participants judged reliability on a substance by substance basis. Between these two participants, one considered all predictions reliable, whereas the other participant considered that all predictions be augmented by expert knowledge.

Table 3: Ready biodegradability results for work package 2

	EXP	1	2	3	5	6	TGD
78-59-1	Ready	+	-	-	+	+	+
88-12-0	Ready	+	+	+	+	+	+
99-54-7	Inherent	-	-	+	+	+	-
107-41-5	Inherent	+	+	?	+	+	+
127-19-5	Ready?	+	+	?	+	+	+
288-32-4	Ready	+	+	+	+	+	+
505-32-8	Inherent?	-	-	+	+	+	+
839-90-7	Not	-	-	-	-	-	-
1477-55-0	Not	-	-	?	+	+	-
6864-37-5	Not	-	-	?	+	+	-

Additionally, several different approaches were taken to arrive at a final prediction for individual substances, where one participant used different combinations of BIOWIN models, whereas the two other participants used the TGD formula, which can be interpreted as a prediction of (non-)persistence rather than ready biodegradability.

Table 3 gives the individual results for the 10 substances chosen for Work Package 2 for the 5 BIOWIN models and the TGD recommendation; note that a + sign indicates a correct prediction.

It appears that, for these 10 substances, the BIOWIN 5 and 6 models offer the most correct predictions. What can also be seen is that, for this selection of compounds, the overall performance (how many models offer a correct prediction for a certain substance) depends on the ‘complexity’ of the substance, and the relevance of the substructures recognized by the BIOWIN models within the context of the model. To illustrate this, a substance like 2-methyl-2,5-pentanediol (107-41-5), which is inherently biodegradable, is basically predicted correctly by all models, whereas tris(2-hydroxyethyl)isocyanurate (839-90-7) is incorrectly predicted by all models. In the latter case, the models are unable to cope with the complex heterocyclic structure, and a.o. misinterpret the reactivity of the carboxyl groups due to context issues.

During the meeting there was some discussion concerning the usefulness of the ready biodegradability test(s) as such. Unlike in the inherent biodegradability test, the only end point in ready biodegradability is ultimate mineralization, while adaptation/acclimation is not considered at all. Another drawback is that in ready biodegradability tests, the concentration of the substance under investigation is orders of magnitude higher than in relevant environmental circumstances, and that it is the only source of organic carbon for the micro-organisms present. It is generally felt that there should be a better way to augment it when a more realistic biodegradability test is called for. However, the overall consensus was that the ready biodegradability test, with all its shortcomings and multiple incarnations, is a useful first tier screening test.

A suggestion to use all BIOWIN models as a ‘predictive battery’ met with mixed enthusiasm, mainly since the BIOWIN 5 and 6 models are so much more comprehensive and methodologically better than the other models. Moreover, these models are largely based on the same training set. Therefore, these models are not independent, and are not at the same level, making it methodologically unsound to combine them in a so-called test (or rather model) battery. From a practical point of view, it was not expected that such an approach would improve overall predictivity *or* reliability.

It was generally agreed that evaluating the reliability and suitability of a model for a certain substance is ‘compromised’ by actually knowing the experimental result for that compound. This makes is somewhat difficult to judge the suitability of a certain model to real ‘unknown’ chemicals. To address that issue, it was agreed that in a follow-up effort, participants would be asked to generate predictions for 10 substances without having prior knowledge and information on these substances (i.e. no SIDS substances), or so-called ‘blind testing’.

3.2.3 Availability and operability of models

The BIOWIN models are included in the US-EPA’s EPISuite™ package, and as such are freely available for download from the internet. Operation is straightforward, with CAS number or SMILES code input, and options for summarized or detailed output. The user manual is comprehensive and well written, but is the only reference for checking whether a substance is part of the models’ training set(s). No information concerning the applicability domain is provided by either the manual or the program, possibly because the models are intended to be applicable to all organic substances.

3.2.4 Conclusions

- The results were in excellent agreement with the Danish EPA effort.
- (Q)SARs don’t have to be more accurate than experimental results. Generally speaking, the available dataset indicates that experimental results are not more than 90% reliable.
- For the prediction of Ready Biodegradability (relevant for classification and labelling) it appears that the BIOWIN 5 and 6 models are the most reliable; however, BIOWIN models 1-3 appear to more closely correlate with inherent biodegradability, which is an important parameter in itself, for determining ultimate persistence, as well as for determining a compound’s PBT profile.
- Borderline predictions (those predictions in which the numerical values are close to the decision cut-off value) are either suspect, or should be considered equivocal. Not enough guidance is available on how to treat equivocal predictions, but it was generally agreed that a model like the BIOWIN models, should be trimodal (yes/maybe/no) rather than bimodal (yes/no).
- For every prediction it needs to be checked whether credible substructures were identified that support the prediction, and preferably should be augmented with

available mechanistic information on structural analogues; expert knowledge is required to fully appreciate (Q)SAR results.

- The OECD principles are helpful in the discussion/exchange of experience; however, a lack of full compliance should not automatically discredit a model.
- Not enough experience was obtained with applying (Q)SAR to true unknowns. This will be addressed in a follow-up effort.
- There was a general consensus that the overall dataset statistics do not significantly add to the advancement of experience and confidence.

3.3 Mutagenicity

3.3.1 Work Package 1

All predictions were done with the computerized expert system DEREK for Windows[7] {LHASA Ltd, DEREK for Windows}, and with the manual structural alert scheme as published by Ashby and Tennant {Ashby and Tennant, 1991, Mutation Research, 257, 229-306}.

DEREKfW accepted all 177 substances for prediction, whereas the Ashby/Tennant model was applicable to 176 out of 177 substances. The experimental results (SIDS data) for the 177 SIDS substances indicated that most substances (142, or 80%) tested negative in the reported Ames test(s), whereas only 23 (13%) tested positive; 12 substances (7%) showed equivocal results.

3.3.1a OECD Principles

Again, it was concluded that neither model fully complies with the OECD principles on (Q)SAR. Whereas both models are based on a defined end point, the Ashby model's end point is carcinogenicity, not mutagenicity in the Ames test. For the DEREKfW model, it should be noted that not all Ames tests are created equal. The model documentation does not elaborate on what types of Ames test were used for or excluded from the training set.

For the DEREKfW model, no information on the algorithm is available, whereas the 'algorithm' for the Ashby model is very simple and rather non-exact. For neither model, the applicability domain was explicitly defined.

Both models have been validated by several research groups, but no formal validation information according to the OECD principles for (Q)SAR is available. Since neither model supply information on their training set, no formal comparison between SIDS data and training set data could be performed.

3.3.1b Prediction results

Table 4 contains the prediction statistics for the DEREKfW and Ashby model for mutagenicity, as compared with the SIDS experimental data on mutagenicity (Ames test). Immediately noticeable is the skewed results for Positive and Negative Predictive Values. This is a direct consequence of the skewed nature of the data set, with 80% negative substances and 20% positive substances. It can be shown that models with this distribution of positives and negatives and sensitivity and specificity oth 90%, has a PPV of 69% and an NPV of 97%.

[7] LHASA Ltd supports the use of DEREK for Windows results in the (Q)SAR experience project; however they do not encourage the use of positive and negative classification of results outside of their context. More specifically, LHASA Ltd objects to using these predictions as equivalent substitutes for test results per se.

Table 4: Statistical results for the DEREKfW and Ashby 'mutagenicity' models

	DEREKfW			Ashby		
	Overall	**Domain**	**Dom - TS**	**Overall**	**Domain**	**Dom - TS**
Sensitivity	70	71	50	65	67	50
Specificity	89	87	87	87	87	87
Concordance	86	83	81	84	85	85
PPV	50	63	40	44	35	23
NPV	95	91	91	94	96	96

It appears that the overall results for both models are about equal. However, the Ashby model appears to be applicable to a wider range of substances. From the results of Work Package 1 no conclusion can be drawn as to the suitability or preferability of a certain model to certain substances. The results of the Danish EPA MultiCASE model for the SIDS substances were however substantially better.

3.3.1.c Comparison with Danish results

The Danish EPA effort focused primarily on the commercial TOPKAT and MultiCASE models (including a Danish EPA in-house model developed with MultiCASE). As these models were not made available for the (Q)SAR experience project, no direct comparison in running and interpreting mutagenicity models could be made between the Danish EPA effort and the current project.

3.3.2 Work Package 2

Table 5 presents the 10 substances selected for hands-on/eyes-on experience with the mutagenicity (Q)SARs. This set consisted of five positive (i.e. mutagenic) substances and five negative (i.e. non-mutagenic) substances.

It can be seen that the correspondence between predictions and experimental results is comparable between the Ashby structural alert 'model' and the DEREKfW model. In fact, all negative structures are predicted correctly by either model, whereas two of the positives are incorrectly predicted as negative by both models. However, the reliability scores of the individual predictions, as awarded by the participants working with the models, tend to differ quite a lot.
The Ashby model prediction for formaldehyde e.g. was awarded the 'unknown' reliability score since, although the model contains the 'aliphatic aldehyde' structural alert, no positive aliphatic aldehyde was in fact present in the model's training set.

The results presented for the individual substances suggest that a weight-of-evidence approach (using both the models featured in the (Q)SAR experience project *and* the models that were used in the Danish EPA effort) seems to give the most reliable results, both for positive and negative predictions. This does result in a significant fraction of substances being classified as 'equivocal'. This could be a trigger for testing such substances.

It was remarked that proprietary models should provide enough information to judge whether negative predictions are true negatives or just 'don't know' responses, in order to be able to a priori judge the reliability of such a prediction. Additionally, the example of the (negative) substance isophorone was discussed, where all models yield 'negative' predictions, but the Ashby-Tennant scheme makes a reservation based on the presence of a conjugated double

Table 5: Mutagenicity results for Work Package 2

Mutagenicity			Ashby		DEREKfW	
CAS	name	EXP	cf model	rel[8]	cf model	rel[4]
50-00-0	Formaldehyde	Pos.	+	4	+	1
75-38-7	1,1-difluoroethylene	Pos.	-	2	-	4
98-59-9	4-methylbenzenesulfonyl chloride	Pos.	-	4	-	4
106-88-7	1,2-epoxybutane	Pos.	+	1	+	1
25321-14-6	dinitrotoluene	Pos.	+	1	+	1
78-59-1	Isophorone	Neg.	+	1	+	1
140-88-5	Ethyl acrylate	Neg.	+/-	4	+	1
1163-19-5	Bis(pentabromodiphenyl)ether	Neg.	+	1	+	1
123-54-6	2,4-pentadione	Neg.	+	1	+	1
98-92-0	Nicotinamide	Neg.	+	2	+	1

bond. When asked whether this would constitute an acceptable result for waiving further testing, it became clear that some participants would regard even a minor warning as described as a clear trigger for performing a test. This of course would have severe repercussions on the usefulness of (Q)SAR. It was concluded that at a minimum, all available additional information should be presented to corroborate the negative findings, such as background information provided by the model program, information available in the open literature, information on structural analogues, results from additional models, etc., despite the fact that the prediction were in the applicability domain of the models.

3.3.3 Availability and operability of models

The Ashby ‘model’ entails manual application of the instructions contained in the relevant Ashby and Tennant publications. As such the model is freely available, but requires a fair amount of user expertise and experience to use it with confidence. Information on the applicability domain is not formally defined and therefore up to expert judgment. The DEREK for Windows program is a semi-commercial package. Operation and interpretation require training and experience, as well as a firm understanding of both (Q)SAR and mutagenicity in the Ames test.

3.3.4 Conclusions

- It became clear (again) that for the evaluation of (Q)SAR results expert knowledge is required; just as for the evaluation of experimental results.
- There are not enough models available for human toxicology end points that are reliable, relevant, and universal; those that are, are difficult or expensive to obtain.
- The model applicability domain is generally accepted as an important principle. However, confidence in the predictions was limited even when substances were within the (positive or negative) applicability domain of a model. The positive applicability domain of DEREKfW is better defined than the negative applicability domain. As a result, Ames test mutagenicity predictions were not considered reliable.
- (Q)SAR cannot just be used as alternative for testing, but also as corroborating or

[8] 1 = reliable;
2 = can be used with expert knowledge
3 = unreliable
4 = unknown

additional information, e.g. in the case of equivocal experimental results. A mutagenicity model, while not necessarily useful within REACH for substituting animal testing, can therefore be very worthwhile. This also includes the approach of using (Q)SAR not as a stand-alone tool but as a part of a Weight-of-Evidence approach.

- The concept of 'end point' needs to be more clearly defined, where it is clear that a single end point may not be a well-defined end point at all, such as is the case in biodegradation, where the result depends a.o. on the composition and adaptation of the microbial community, or in mutagenicity, where the result depends a.o. on the Salmonella strain used or the enzyme systems added to the test.
- Within the evaluation of chemicals, the Ames test is a screening test for genotoxicity and as such is part of a testing strategy. It should be clearly defined how (Q)SAR predictions fit into this testing strategy. For instance, it was the feeling of most participants that even with all models predicting a negative outcome in the Ames test an experimental Ames test might still be required. Therefore guidance is needed on how to apply (Q)SARs to real unknowns in different evaluation settings and depending on the consequences.

3.4 Fish Toxicity

ECB selected 3 regression equation (Q)SAR models for this effort, based on Kow as the independent descriptor. The performance of these models was evaluated in light of the OECD principles; additionally the results were statistically evaluated. Some background of the ECB model development process is given. The end point chosen was the acute mortality (LC50) to *Pimephales promelas* (fathead minnow).

Table 6: Summary of training set statistics for fish (Q)SARs

	Non-polar narcosis	Polar narcosis	Combined model
n	58	86	144
r^2	92.2	90.1	87.6
s	0.41	0.33	0.46
F	661	763	998
LOF	0.18	0.11	0.21
SDEC	0.40	0.33	0.45
AIC	0.18	0.12	0.21
FIT	11.1	8.67	6.83
h*	0.103	0.070	0.042

The model performance for these three models was documented with statistics on the training set. A summary of these statistics is given in table 6. As can be seen from these results, both are adequate models within their applicability domain, with the individual models for non-polar and polar narcosis performing slightly, but not significantly better than the combined model.

All three training sets adequately span their respective model's applicability domain, both from a mode of action (i.e., structural) point of view, as well as concerning the descriptor space (log Kow). Models were checked against the OECD principles on (Q)SAR. All three models have a defined end point, a well-defined algorithm, and are based on a mechanistic understanding of the end point. For all models, the applicability domain is well-defined. The applicability domain for the individual models was visualized in a so-called William's plot, using a 'leverage' based metric, for both the descriptor space and the effect parameter.

3.4.1 Work Package 1

For only 32 substances acute toxicity information on fathead minnow was available in the SIDS data; augmentation with data from the AQUIRE database {US-EPA ERL-Duluth, 1989, AQUatic toxicity Information REtrieval database (AQUIRE)} resulted in a data set of 57 substances. Substances were classified as to belonging to the applicability domain of the models, based on Mode of Action, leverage, and outlier status. Model results for the 57 substances range from 0.1–10 times the actual acute toxicity when used only on substances in the model's applicability domain. When taking into account all substances, regardless of applicability domain, the range became much larger, with the observation that for the non-polar narcosis model, this range was primarily extended to the higher end (i.e. substances outside the applicability domain are generally more toxic than predicted, in fact the expected behaviour for a baseline toxicity model). As such there was a large correlation between outlier status and mode of action, less so for leverage and mode of action.

Discussions on the results for work package 1 emphasized that statistics as presented give a very good overview of the performance of the models; but may not be too helpful in applying the model to unknowns. It was furthermore argued whether defining the applicability domain both on a mode of action criterion and on outlier status is helpful. The rationale would be that if the mode of action criterion is well-defined, there really should be no major outliers, and conversely, if the domain is defined based on statistical outliers, there is no way to a priori determine whether an unknown is in the applicability domain of the model.

It was furthermore suggested that experimental mode of action information, e.g. from *in vitro* testing, might be included in defining whether a substance is within the applicability domain of a model. It was again stressed that experimental results are not infinitely accurate and precise; this should be reflected in the way people look at (Q)SAR results. Several participants make the point that it is not expected that any assessment of hazardous properties of chemical substances, whether based on experimental results or on model predictions, will ever be able without some level of expert knowledge. The leverage-based approach of visualizing the descriptor applicability domain is seen as a very useful tool, especially for models with many, or highly technical (PCs, latent variables) descriptors.

3.4.2 Work Package 2

In work package 2 ten individual substances were predicted with the 3 selected models.

Table 7: Work Package 2 results for the non-polar narcosis (Q)SAR

CAS	name	EXP	Non-polar narcosis (Q)SAR	Domain	ratio (exp/pred)
78-87-5	1,2-dichloropropane	1.24	0.54	Y	2.31
79-20-9	acetic acid methyl ester	4.32	22.44	Y	0.19
95-50-1	1,2-dichlorobenzene	0.39	0.07	Y	5.57
96-18-4	1,2,3-trichloro propane	0.45	0.33	Y	1.38
100-41-4	ethylbenzene	0.11	0.11	Y	1.00
107-98-2	1-methoxy-2-propanol	230.67	123.59	Y	1.87
108-88-3	methyl benzene	0.28	0.30	Y	0.93
123-86-4	acetic acid butyl ester	0.15	1.19	N - MoA & outlier	0.13
770-35-4	1-phenoxy-2-propanol	1.84	2.29	Y	0.80
1490-04-6	5-methyl-2-(1-methylethyl)cyclohexanol	0.12	0.06	Y	1.97

Table 8: Work Package 2 results for the polar narcosis (Q)SAR

CAS	name	EXP	Polar narcosis (Q)SAR	Domain	ratio (exp/pred)
78-87-5	1,2-dichloropropane	1.24	0.16	N - MoA & outlier	7.64
79-20-9	acetic acid methyl ester	4.32	3.74	N - MoA	1.15
95-50-1	1,2-dichlorobenzene	0.39	0.03	N - MoA & outlier	13.37
96-18-4	1,2,3-trichloro propane	0.45	0.11	N - MoA	4.23
100-41-4	ethylbenzene	0.11	0.04	N - MoA	2.59
107-98-2	1-methoxy-2-propanol	230.67	15.70	N - MoA & outlier	14.69
108-88-3	methyl benzene	0.28	0.10	N - MoA	2.83
123-86-4	acetic acid butyl ester	0.15	0.32	N - MoA	0.49
770-35-4	1-phenoxy-2-propanol	1.84	0.55	N - MoA	3.36
1490-04-6	5-methyl-2-(1-methylethyl)cyclohexanol	0.12	0.02	N - MoA & outlier	4.80

Table 9: Work Package 2 results for the global narcosis (Q)SAR

CAS	name	EXP (mM)	Global (Q)SAR	Domain	ratio (exp/pred)
78-87-5	1,2-dichloropropane	1.24	0.27	Y	4.57
79-20-9	acetic acid methyl ester	4.32	9.04	Y	0.48
95-50-1	1,2-dichlorobenzene	0.39	0.04	Y	9.77
96-18-4	1,2,3-trichloro propane	0.45	0.17	Y	2.65
100-41-4	ethylbenzene	0.11	0.06	Y	1.80
107-98-2	1-methoxy-2-propanol	230.67	44.98	Y	5.13
108-88-3	methyl benzene	0.28	0.16	Y	1.79
123-86-4	acetic acid butyl ester	0.15	0.57	N - MoA	0.27
770-35-4	1-phenoxy-2-propanol	1.84	1.06	Y	1.74
1490-04-6	5-methyl-2-(1-methylethyl)cyclohexanol	0.12	0.03	Y	3.57

Substances were classified as to being within the applicability domain of the models, based on mode of action (structure), leverage, and outlier status. Again, use of the 3n/p cut-off criterion seemed overly conservative, while using an outlier-status based criterion depends on knowing the actual experimental toxicity of the substance, and would not be applicable to real-world unknowns. A general conclusion is that for acceptance of model application, a better, more operational definition of the applicability domain is needed.

Unfortunately the available models do not cover the complete chemical/toxicological domain of the 177 SIDS substances; moreover the ten selected substances represented 9(8) non-polar narcosis substances and 1(2) ester toxicity substances. Overall, the Non-polar narcosis model yields the better predictions, but the results did not provide much insight in the applicability of (Q)SAR to unknown substances.

Ensuing discussion suggested that there is a need for more (Q)SAR models for MoAs beyond baseline toxicity; alternatively, a competing approach, where predicted toxicity is a product of baseline toxicity and an excess factor, may be investigated and promoted. If models for additional MoAs are developed, this will mean that expert knowledge in evaluating Mode of Action will become relevant in evaluating (Q)SAR predictions. ECB data suggest that ca 50% of substances are amenable to baseline toxicity predictions.

Ideally, such models should be publicly available, to both industry and regulators, as well as to the general public. This suggests that a central (Q)SAR repository, accessible to all those involved in the hazard assessment of chemical substances is something that should be given serious consideration.

3.4.3 Availability and operability of models

All models presented here are published linear regression models, and as such are freely available, and can be used by anyone with a slide rule, electronic calculator, or spreadsheet program. The only requirement is having a value for the log Kow of the substance under concern. Determining whether a substance falls within the applicability domain of these models is slightly more involved; this can be done by a number of computer programs that perform substructure recognition, such as EPISuite™, or by using a manual decision tree, such as that published by {Verhaar et al., 1992, Chemosphere, 25, 471-491}.

3.4.4 Conclusions

- It was acknowledged that some basic understanding of statistics is required in order to fully appreciate the performance of (Q)SAR models.
- OECD principles on (Q)SARs help to ask the appropriate questions.
- (Q)SAR works very well for narcosis substances (both PN and NPN combined and apart); definition of the domain for this MoA can be based either on chemical structure or descriptor space or possibly a combination.
- For other substances, different approaches, such as MoA specific (Q)SARs, excess toxicity approaches, are required.
- Classification of chemicals into ‘mode of action’ classes, usually based on structural alerts, requires expert knowledge.
- The use of (Q)SAR models in a Weight-of-Evidence approach may reduce the concern for over- or underpredictions.
- A leverage-based approach to domain definition based on independent descriptors is useful

4 Discussion and Follow up to the 1st phase

In general the predictions generated in the Work packages 1 (the whole SIAM dataset) and the statistical calculations (correct in x% of the cases) were seen as very useful to get a feeling for the predictive capabilities of a model, but translating this information to the reliability of one specific prediction proved to be difficult. It was remarked by several participants that detailed knowledge of the model (training set, algorithm, applicability domain, predictivity) was needed as well as specific knowledge of the endpoint being predicted, as a lot of the uncertainty in the prediction is related to issues that are very specific to the endpoint of interest, and the (experimental) peculiarities playing a role in the experimental determination of the endpoint. A lot of properties (of the substance) that make a QSAR prediction unreliable, and also determine the applicability domain of the model, are identical to what would make an experimental result unreliable as well.

Hands-on experience builds confidence and eyes on experience was considered to give insufficient confidence (as an example the TOPKAT predictions that were within the applicability domain were not considered to be valid as stand alone predictions, more “evidence” was considered necessary to draw a (regulatory) conclusion). Limited availability of (and therefore experience with) models was therefore perceived as a large drawback.

A central (Q)SAR repository may solve that problem, however knowing how the model works does not make it necessarily a better model. The uncertainties of the model will become clearer.

Endpoint

Expert knowledge on (Q)SAR is important as well as expertise on the endpoint, examples are BIOWIN and Mutagenicity. The variability in experimental animal testing outcome needs to be considered depending on the endpoint and test system.

Applicability domain

The practical use of the OECD principles should be discussed as some principles are important e.g. applicability domain. However, a (computer generated) indication that the substance is within the applicability domain of the model does not necessarily lead to more or higher confidence in the prediction (this was for example the the case with the eyes-on TOPKAT predictions). Again the limited availability and experience with certain models was the reason for less confidence in the prediction and/or more scepticism whether such a prediction could be trusted at all.

Implication of the prediction

When using a prediction it was considered necessary to know what the implication of the prediction is. Only then (Q)SARs can contribute to intelligent testing strategies, Weight-of-Evidence approaches, or sometimes even be used as stand alone, because the regulatory consequences (or lack thereof) are known! (e.g. (no) classification: or further testing at the next tonnage threshold)

The 'blind testing' experience effort, as the next major part of the (Q)SAR experience project, may need explicit statement of the regulatory context, and further testing proposals should be considered indicating the tonnage level for which that conclusion would be valid.

Differences in reliability scoring

It was observed that different participants rated the outcome of identical predictions from the (Q)SAR models used very differently. The Klimisch code was used with different interpretations, i.e. a BIOWIN 1 prediction was rated by one participant as quality 1, can be used on its own with the rationale that the prediction was very reasonable, not in the equivocal zone, a large part of the structure was covered by the identified substructures.

However, the same prediction was given a Klimisch reliability code 3 (unreliable, can not be used) because the endpoint used for establishing the BIOWIN 1 model (an evaluated biodegradability score taking into account all experimental data, including inherent studies, field studies etc.) did not comply with the regulatory endpoint of ready biodegradability (i.e. result of one of the OECD 301 tests). Others argued that a QSAR prediction could never be rated as 1, but only 2 at best (reliable outcome, but can not be used on its own).

One participant gave a ranking of 3 to a specific (Q)SAR prediction, because it contradicted the (known) experimental result.

The discussion on the Klimisch code showed that reliability of a (QSAR) prediction can be judged differently when viewing the prediction on its own vs. compared to other data, or when judging the prediction of the (toxicological) endpoint for which the model was optimized vs. the regulatory endpoint that is evaluated.

Follow-up to the first phase

Hands-on experience with actual (Q)SARs does increase the confidence people have in (Q)SARs as well as in improved understanding of why and when (Q)SARs will and won't work. As with any experimental result, expert knowledge will always be important to interpret (Q)SAR results and findings, no matter how advanced the (Q)SAR computer models become. Anyone evaluating chemical hazard information should be fully aware that experimental (test) results are not infinitely accurate and reliable; in fact (Q)SAR could sometimes be more reliable than test results. At a minimum, this should imply that (Q)SARs don't have to be more reliable, in a statistical sense, than experimental information.

In addition to providing substitute hazard estimates, (Q)SARs will be very valuable in defining intelligent testing strategies and formulating Weight-of-Evidence approaches in determining a substance's hazard profile. However, the *a priori* evaluation of the suitability of a specific (Q)SAR model to an 'unknown' substance remains an important issue. Therefore it is proposed to continue the experience project with a series of 'blind testing' examples[9]. These examples, being the next major part of the (Q)SAR experience project, should also result in guidance on how to properly document and evaluate the suitability of QSAR models and prediction for a specific regulatory purpose.

The *practical* use of the OECD principles should be discussed more thoroughly, and possibly implemented in the guidance on the use of QSARs as well.

The limited availability of human toxicology models in particular is perceived as a major drawback in the acceptance and use of (Q)SAR. This issue may be solved by developing a central (Q)SAR repository.

Acknowledgements

The support of LHASA UK Ltd in supplying the (Q)SAR experience project with their permission of using and publishing the results from DEREK for Windows within the context of the (Q)SAR experience project is kindly acknowledged.

[9] These blind testing examples were the reason for development of the reporting formats as described in main text of this report, and also formed the exercises/case studies needed for discussing the formats within the EU QSAR Experience Project.