EUROPEAN COMMISSION
DIRECTORATE GENERAL
JOINT RESEARCH CENTRE
Institute for Health and Consumer Protection
Toxicology and Chemical Substances Unit

Ispra, 26/05/2008

QSAR Model Reporting Format (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 1 for Biodegradation"; "TOPKAT Skin Irritation Acyclics (No Acids, Amines, Esters) MOD v SEV Model".*

1.2 Other related models: *If appropriate, identify any model that is related to the model described in the present QMRF. Example: "TOPKAT Skin Irritation Acyclics (Acids, Amines, Esters) NEG/MLD v MOD/SEV Model" is related to the model mentioned in 1.1: "TOPKAT Skin Irritation Acyclics (Acids, Amines, Esters) MOD v SEV 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 2.7).*

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.*

Supporting Information – QMRF v1.2

Templates will be provided for submitting information about the training and test sets. Storage of searchable information about the training and the test sets in the database of the ECB Inventory will be possible if the submitter uses specific file formats (excel file or sdf files) with predefined fields.

Fields to be included in the training set template

•Chemical Name (IUPAC)

- Chemical Name (Not IUPAC)
- CAS Number
- SMILES
- InChI
- MOL *(file name is reported for excel files; if it is an sdf file, coordinate can be simply included in it).*
- Structural Formula
- Dependent Variable Value
- Descriptor1 Value *(the name of the descriptors should be specified by the user).*
- Descriptor 2 Value
- Descriptor 3 Value
- Descriptor 4 Value
- Descriptor X Value

The template for the test set is identical to the one for the training set.

Endpoint classification

The predefined endpoint classification included in the QMRF is:

1. Physicochemical effects

1.1 Melting point
1.2 Boiling point
1.3 Water solubility
1.4 Vapour pressure
1.5 Surface tension
1.6 Octanol-water partition coefficient (Kow)
1.7 Octanol-water distribution coefficient (D)
1.8 Octanol-air partition coefficient (Koa)
1.9 Air- water partition coefficient (Henry's law constant, H)
1.10 Dissociation constant (pKa)

2. Environmental fate parameters

2.1 Persistence: Abiotic degradation in water
  a. Hydrolysis
  b. Oxidation
  c. Other
2.2 Persistence: Abiotic degradation in air (Phototransformation)
  a. Direct photolysis
  b. Indirect photolysis (OH-radical reaction, ozone-radical reaction, other)
2.3 Persistence: Biodegradation
  a. Ready/not ready biodegradability
  b. Biodegradation time frame (primary, ultimate degradation)
2.4 Bioconcentration
  a. BCF fish

b. BCF other organisms

2.5 Bioaccumulation

a. BAF fish

b. BAF other organisms

2.6 Organic carbon-sorption partition coefficient (organic carbon; Koc)

2.7 Adsorption/Desorption in soil

2.8 Adsorption/Desorption in sediment

2.9 Vegetation-water partition coefficient

2.10 Vegetation-air partition coefficient

2.11 Vegetation-soil partition coefficient

3. Ecotoxic effects

3.1 Short-term toxicity to Daphnia (immobilisation)

3.2 Short-term toxicity to algae (inhibition of the exponential growth rate)

3.3 Acute toxicity to fish (lethality)

3.4 Long-term toxicity to Daphnia (lethality, inhibition of reproduction)

3.5 Long-term toxicity to fish (egg/sac fry, growth inhibition of juveniles, early life stage, full life cycle)

3.6 Microbial inhibition (activated sludge respiration inhibition, inhibition of nitrification, other)

3.7 Toxicity to soil microorganisms (inhibition of C-mineralisation, inhibition of N-mineralisation, other)

3.8 Toxicity to earthworms (survival, growth, reproduction)

3.9 Toxicity to plants (leaves, seed germination, root elongation)

3.10 Toxicity to soil invertebrates (survival, growth, reproduction)

3.11 Toxicity to sediment organisms (survival, growth, reproduction)

3.12 Toxicity to birds

a. Short term toxicity (feeding, gavage, other)

b. Long-term toxicity (survival, growth, reproduction)

4. Human health effects

4.1 Acute inhalation toxicity

4.2 Acute oral toxicity

4.3 Acute dermal toxicity

4.4 Skin irritation/corrosion

4.5 Acute photoirritation

4.6 Skin sensitisation

4.7 Respiratory sensitisation

4.8 Photosensitisation

4.9 Eye irritation/corrosion

4.10 Mutagenicity

4.11 Photomutagenicity

4.12 Carcinogenicity

4.13 Photocarcinogenicity

4.14 Repeated dose toxicity

4.15 In vitro reproductive toxicity (e.g. embryotoxic effects in cell culture such as embryo stem cells)

4.16 In vivo pre-natal-developmental toxicity

4.17 In vivo pre-, peri-, post natal development and/or fertility (1 or 2 gen. study or enhanced 1 gen. study)

4.18 Endocrine Activity:

a. Receptor-binding (specify receptor)
b. Receptor binding and gene expression (specify receptor)
c. Other (e.g. inhibition of specific enzymes involved in hormone synthesis or regulation, specify enzyme(s) and hormone)

5. Toxicokinetics

5.1 Skin penetration

5.2 Ocular membrane penetration

5.3 Gastrointestinal absorption

5.4 Blood-brain barrier penetration

5.5 Placental barrier penetration

5.6 Blood-testis barrier penetration

5.7 Blood-lung barrier penetration

5.8 Metabolism (including metabolic clearance)

5.9 Protein-binding

5.10 DNA-binding

6. Other