SciMatics SciQSAR version of commercial MultiCASE MC4PC model A49 for teratogenic potential in humans

1. QSAR identifier

1.1 QSAR identifier (title)

SciMatics SciQSAR version of commercial MultiCASE MC4PC model A49 for teratogenic potential in humans, Danish QSAR Group at DTU Food.

1.2 Other related models

MultiCASE CASE Ultra version of commercial MultiCASE MC4PC model A49 for teratogenic potential in humans, Danish QSAR Group at DTU Food.

Leadscope Enterprise version of commercial MultiCASE MC4PC model A49 for teratogenic potential in humans, Danish QSAR Group at DTU Food.

1.3. Software coding the model

SciQSAR version 3.1.00.

2. General information

2.1 Date of QMRF

January 2015.

2.2 QMRF author(s) and contact details

QSAR Group at DTU Food;

Danish National Food Institute at the Technical University of Denmark;

http://qsar.food.dtu.dk/;

qsar@food.dtu.dk

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Marianne Dybdahl;

National Food Institute at the Technical University of Denmark;

Sine Abildgaard Rosenberg;

National Food Institute at the Technical University of Denmark;

2.3 Date of QMRF update(s)

2.4 QMRF update(s)

2.5 Model developer(s) and contact details

MultiCASE Inc.;

23811 Chagrin Blvd Ste 305, Beachwood, OH, 44122, USA;

www.multicase.com

MultiCASE Inc. has kindly given their permission that remodelling of their training set from the commercial MC4PC A49 model in SciQSAR was performed by:

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Jay Russel Niemelä;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Danish QSAR Group at DTU Food;

National Food Institute at the Technical University of Denmark;

http://qsar.food.dtu.dk/;

qsar@food.dtu.dk

2.6 Date of model development and/or publication

January 2014.

2.7 Reference(s) to main scientific papers and/or software package

Contrera, J.F., Matthews, E.J., Kruhlak, N.L., and Benz, R.D. (2004) Estimating the safe starting dose in phase I clinical trials and no observed effect level based on QSAR modelling of the human maximum recommended daily dose. *Regulatory Toxicology and Pharmacology*, 40, 185 – 206.

SciQSAR (2009) Reference guide: *Statistical Analysis and Molecular Descriptors*. Included within the SciMatics SciQSAR software.

2.8 Availability of information about the model

The training set is commercially available from MultiCASE Inc The training set was originally compiled by MultiCASE Inc. and used to train their commercial MultiCASE MC4PC A49 model. The Danish QSAR Group bought this model from MultiCASE Inc. in 1999. Permission to remodel the training set in SciQSAR was kindly granted by MultiCASE Inc. The model algorithm is proprietary from commercial software.

2.9 Availability of another QMRF for exactly the same model

3. Defining the endpoint

3.1 Species

Human (based on epidemiological, clinical and animal data).

3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.16. In vivo pre-natal-developmental toxicity

3.3 Comment on endpoint

It was previously believed that the mammalian embryo, developed in the impervious uterus of the mother, was protected from all extrinsic factors. However, after the thalidomide disaster of the 1960s, it became apparent and more accepted that the developing embryo and fetus could be highly vulnerable to certain environmental agents that have negligible or non-toxic effects to adult individuals.
Birth defects are known to occur in 3-5 % of newborns. It is estimated that approximately 10% of all birth defects are caused by prenatal exposure to a teratogenic agent. A teratogen is defined as an agent that causes malformations of the embryo or fetus *in utero* and include radiation, infections, maternal metabolic imbalances (e.g. diabetes, folic acid deficiency), drugs (e.g. anticancer drugs, tetracyclines, many hormones, thalidomide), and environmental chemicals (e.g. mercury, lead, dioxins PBDEs, HBCD, tobacco smoke).
Many biological mechanisms are involved in developmental toxicity and this complicates the identification of teratogenic agents by simpler models (e.g. *in vitro*). Six main teratogenic mechanisms associated with medication use has been identified: folate antagonism, neural crest cell disruption, endocrine disruption, oxidative stress, vascular disruption and specific receptor- or enzyme-mediated teratogenesis (van Gelder *et al.* 2010). The time at which the embryo or fetus is exposed to the teratogenic agent also plays a role in the potential outcome as the critical window of exposure differs between the different organ systems.
The training set for this model was compiled by MultiCASE Inc. and is composed of data from the Teratogen Information System (TERIS) and a compilation in which the United States Food and Drug Administration (FDA) definitions were used to quantify potential for developmental toxicity from drugs used during pregnancy (Ghanooni *et al.* 1997). The training set consists of clinical, epidemiologic and animal data. Results from the animal data are extrapolated to humans (Jensen *et al.* 2008).

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Potential teratogenic in humans, positive or negative.

3.6 Experimental protocol

Positive compounds consists of compounds whose FDA classification is D (evidence of human fetal potential) or X (a clear adverse fetal effect in animals, humans or both) or for which the TERIS score assigned is high or moderately high. Negative compounds possess an FDA classification of A (no potential of developmental toxicity on the basis of controlled studies in women) or a TERIS score of "none" or "unlikely." (Ghanooni *et al.* 1997).

3.7 Endpoint data quality and variability

Only TERIS data for which data quality was reported fair, good or excellent were considered (Ghanooni *et al.* 1997).

4. Defining the algorithm

4.1 Type of model

This is a categorical (Q)SAR model based on calculated molecular descriptors, and if available the modeller's own or third-party descriptors or measured endpoints can be imported and used as descriptors.

4.2 Explicit algorithm

This is a categorical (Q)SAR model made by use of the non-parametric discriminant analysis (DA) k-nearest-neighbor (kNN) (K=1, Mahalanobis distance used to determine proximity) method (see 4.5). The specific implementation is proprietary within the SciQSAR software.

4.3 Descriptors in the model

Molecular connectivity indices

Molecular shape indices

Topological indices

Electrotopological (Atom E and HE-States) indices

Electrotopological bond types indices

SciQSAR software provides over 400 built-in molecular descriptors. Additionally, SciQSAR makes it possible to import the modeller's own or third-party descriptors or use measured endpoints as custom descriptors.

4.4 Descriptor selection

The initial descriptor set is manually chosen by the model developer from the total set of built-in descriptors. Furthermore, the set of descriptors applied in the modelling by the program is on top of this selection determined by thresholds for descriptor variance and number of nonzero values likewise defined by the model developer.

48 descriptors were selected from the initial pool of descriptors by the system and used to build the model.

4.5 Algorithm and descriptor generation

For a binary classification problem SciQSAR uses discriminant analysis (DA) to make a (Q)SAR model. SciQSAR implements a broad range of discriminant analysis (DA) methods including parametric and non-parametric approaches. The classic parametric method of DA is applicable in the case of approximately normal within-class distributions. The method generates either a linear discriminant function (the within-class covariance matrices are assumed to be equal) or a quadratic discriminant function (the within-class covariance matrices are assumed to be unequal). When the distribution is assumed to not follow a particular law or is assumed to be other than the multivariate normal distribution, non-parametric DA methods can be used to derive classification criteria. The non-parametric DA methods available within SciQSAR include the kernel and *k*-nearest-neighbor (kNN) methods. The main types of kernels implemented in SciQSAR include uniform, normal, Epanechnikov, bi-weight, or tri-weight kernels, which are used to estimate the group specific density at each observation. Either Mahalanobis or Euclidean distances can be used to determine proximity between compound-vectors in multidimensional descriptor space. When the kNN method is used, the Mahalanobis distances are based on the pooled covariance matrix. When the kernel method is used, the Mahalanobis distances are based on either the individual within-group covariance matrices or the pooled covariance matrix. (Contrera *et al*. 2004)

If the data outcome is continuous, regression analysis is used to build the predictive model. Within SciQSAR several regression methods are available: ordinary multiple regression (OMR), stepwise regression (SWR), all possible subsets regression (PSR), regression on principal components (PCR) and partial least squares regression (PLS). The choice of regression method depends on the number of independent variables and whether correlation or multicollinearity among the independent variables exists: OMR is acceptable with a small number of independent variables, which are not strongly correlated. SWR is used under the same circumstances as OMR but with greater number of variables. PSR is used for problems with a great number of independent variables. PCR and PLS are useful when a high correlation or multicollinearity exist among the independent variables. (SciQSAR 2009)

To test how stable the developed models are, SciQSAR have built-in cross-validation procedures (see 6.).

For this model, the kernel method was used.

4.6 Software name and version for descriptor generation

SciQSAR version 3.1.00.

4.7 Descriptors/chemicals ratio

In this model 48 descriptors were used. The training set consists of 323 compounds. The descriptor/chemical ratio is 1:6.7 (48:323).

5. Defining Applicability Domain

5.1 Description of the applicability domain of the model

The definition of the applicability domain consists of two components; the definition in SciQSAR and the in-house further refinement algorithm on the output from SciQSAR to reach the final applicability domain call.

1. SciQSAR

The first criterion for a prediction to be within the models applicability domain is that all of the descriptor values for the test compound can be calculated by SciQSAR. If SciQSAR cannot calculate each descriptor value for the test chemical no prediction value is given by SciQSAR and it is considered outside the model's applicability domain.

2. The Danish QSAR group

The Danish QSAR group has applied a stricter definition of applicability domain for its SciQSAR models. In addition to the applicability domain definition made by SciQSAR a second criterion has been applied for predictions generated from (Q)SAR models with a binary endpoint. For each prediction SciQSAR calculates the probability (p) for the test compound's membership in one of the two outcome classes (positive or negative). The probability of membership in a class is a measure of how well training set knowledge is able to discriminate a positive prediction from a negative prediction within the nearest space of the subject compound-vector. The probability of membership value is also a measure of the degree of confidence of a prediction. The Danish QSAR group uses this probability for a prediction to further define the model's applicability domain. Only positive predictions with a probability equal to or greater than 0.7 and negative predictions with a probability equal to or less than 0.3 are accepted. Positive predictions with a probability between 0.5 and 0.7 as well as negative predictions with a probability between 0.3 and 0.5 are considered outside the model's applicability domain. When these predictions are wed out the accuracy of the model in general increases at the expense of reduced model coverage. Furthermore, as SciQSAR does not define a structural domain, only predictions which were within either Leadscope structural domain (defined as at least one training set chemical within a Tanimoto distance of 0.7) or CASE Ultra structural domain (no unknown fragments for negatives and maximum 1 unknown fragment for positives) were defined as being inside the SciQSAR applicability domain.

5.2 Method used to assess the applicability domain

The system does not generate predictions if it cannot calculate each descriptor value for the test compound.

Only positive predictions with probability equal to or greater than 0.7 and negative predictions with probability equal to or less than 0.3 were accepted.

5.3 Software name and version for applicability domain assessment

SciQSAR version 3.1.00.

5.4 Limits of applicability

The Danish QSAR group applies an overall definition of structures acceptable for QSAR processing which is applicable for all the in-house QSAR software, i.e. not only SciQSAR. According to this definition accepted structures are organic substances with an unambiguous structure, i.e. so-called discrete organics defined as: organic compounds with a defined two dimensional (2D) structure containing at least two carbon atoms, only certain atoms (H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and I), and not mixtures with two or more ‘big components’ when analyzed for ionic bonds (for a number of small known organic ions assumed not to affect toxicity the ‘parent molecule’ is accepted). Structures with less than two carbon atoms or containing atoms not in the list above (e.g. heavy metals) are rendered out as not acceptable for further QSAR processing. Calculation 2D structures (SMILES and/or SDF) are generated by stripping off accepted organic and inorganic ions. Thus, all the training set and prediction set chemicals are used in their non-ionized form. See 5.1 for further applicability domain definition.

6. Internal validation

6.1 Availability of the training set

No

6.2 Available information for the training set

6.3 Data for each descriptor variable for the training set

6.4 Data for the dependent variable for the training set

6.5 Other information about the training set

323 compounds are in the training set: 130 positives and 193 negatives.

6.6 Pre-processing of data before modelling

Only structures acceptable for SciQSAR were used in the final training set. That is, only discrete organic chemicals as described in 5.4 were used. In case of replicate structures, one of the replicates was kept if all the compounds had the same activity and all were removed if they had different activity. No further structures accepted by the software were eliminated (i.e. outliers).

6.7 Statistics for goodness-of-fit

SciQSARs own internal performance test of the model gave the following Cooper's statistics for predictions within the applicability domain as defined by SciQSAR (i.e. the first criterion described in 5.1):

- Sensitivity (true positives / (true positives + false negatives)): 100%
- Specificity (true negatives / (true negatives + false positives)): 100%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 100%

6.8 Robustness – Statistics obtained by leave-one-out cross-validation

Not performed.

6.9 Robustness – Statistics obtained by leave-many-out cross-validation

SciQSAR's own internal 10-fold cross-validation (10*10% out) procedure was used for predictions within the applicability domain as defined by SciQSAR (i.e. the first criterion described in 5.1). As the probability domain was not applied (i.e. the second criterion described in 5.2) the accuracy of the predictions when applying this domain can be expected to be higher than reflected in these cross-validation results. This gave the following Cooper's statistics:

- Sensitivity (true positives / (true positives + false negatives)): 64.6%
- Specificity (true negatives / (true negatives + false positives)): 92.7%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 81.4%

6.10 Robustness - Statistics obtained by Y-scrambling

Not performed.

6.11 Robustness - Statistics obtained by bootstrap

Not performed.

6.12 Robustness - Statistics obtained by other methods

Not performed.

7. External validation

7.1 Availability of the external validation set

7.2 Available information for the external validation set

7.3 Data for each descriptor variable for the external validation set

7.4 Data for the dependent variable for the external validation set

7.5 Other information about the training set

7.6 Experimental design of test set

7.7 Predictivity – Statistics obtained by external validation

7.8 Predictivity – Assessment of the external validation set

7.9 Comments on the external validation of the model

External validation has not been performed for this model.

8. Mechanistic interpretation

8.1 Mechanistic basis of the model

The SciQSAR software provides over 400 calculated physico–chemical, electrotopological E-state, connectivity and other molecular descriptors. The descriptors selected for the model may indicate modes of action that are obvious for persons with expert knowledge about the endpoint.

8.2 A priori or posteriori mechanistic interpretation

A posteriori mechanistic interpretation. The descriptors selected for the model may provide a basis for mechanistic interpretation.

8.3 Other information about the mechanistic interpretation

9. Miscellaneous information

9.1 Comments

The model is useful for identifying potential human developmental toxicants, as well as serving as a starting point for further mechanistic investigations (Ghanooni *et al.* 1997).

9.2 Bibliography

Ghanooni, M., Mattison, D. R., Zhang, Y.P., Macina, O. T., Rosenkranz, H. S., and Klopman, G. (1997) Structural determinants associated with risk of human developmental toxicity. *Am. J. Obstet. Gynecol.*, 176:4, 799-805.

Jensen, G. E., Niemelä, J. R., Wedebye, E. B., and Nikolov, N. G. (2008) QSAR models for reproductive toxicity and endocrine disruption in regulatory use - a preliminary investigation. *SAR and QSAR in Environmental Research*, 19:7,631-641.

van Gelder, M.M.H.J., van Rooij, I.A.L.M., Miller, R.K., Zielhuis, G.A., de Jong-van den Berg, L.T.W, and Roeleveld, N. (2010) Teratogenic mechanisms of medical Drugs. *Human Reproduction Update*, 16:4, 378–394.

9.3 Supporting information