	*QMRF identifier (JRC Inventory):* **To be entered by JRC**
	QMRF Title: **Carcinogenicity oral classification model (IRFMN) (version 1.0.1)**
	Printing Date: **13-lug-2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Carcinogenicity oral classification model (IRFMN) (version 1.0.1)

1.2.Other related models:

Carcinogenicity oral Slope Factor model (IRFMN) (version 1.0.1)

1.3.Software coding the model:

VEGA (https://www.vegahub.eu/)

The VEGA software provides QSAR models to predict tox, ecotox, environ, phys-chem and toxicokinetic properties of chemical substances.

emilio.benfenati@marionegri.it

CDK

The Chemistry Development Kit

The CDK developers https://github.com/cdk

2.General information

2.1.Date of QMRF:

13-07-2022

2.2.QMRF author(s) and contact details:

[1]Cosimo Toma Istituto di Ricerche Farmacologiche Mario Negri IRCCScosimo.toma@marionegri.it https://www.researchgate.net/profile/Cosimo_Toma

[2]Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri IRCCSemilio.benfenati@marionegri.it

[3]Alberto Manganaro Kode srl info@kode-solutions.net /

2.3.Date of QMRF update(s):

No update

2.4.QMRF update(s):

No update

2.5.Model developer(s) and contact details:

Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

2.6.Date of model development and/or publication:

2017

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

[1] Toma, C., Manganaro, A., Raitano, G., Marzo, M., Gadaleta, D., Baderna, D., ... & Benfenati, E. (2021). QSAR Models for Human Carcinogenicity: An Assessment Based on Oral and Inhalation Slope Factors. Molecules, 26(1), 127.

[2] Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, and Screening Tools: The VEGAHUB Platform as an Example. Advances in Computational Toxicology; Springer; 2019. p. 365-81.

2.8.Availability of information about the model:

The model is non-proprietary and the training set is available.

2.9.Availability of another QMRF for exactly the same model:

Another QMRF is not available.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Different species (Human, rat, mouse)

3.2.Endpoint:

TOX 7.7. Carcinogenicity

3.3.Comment on endpoint:

Linear extrapolation should be used when there are data to indicate that the dose-response curve has a linear component below the POD (point of departure). In this case there is a proportional (linear) relationship between risk and dose at low doses. The slope of this line, known as the slope factor, is an upper-bound estimate of risk per increment of dose for carcinogens that can be used to assess the increase over a lifetime in incidence of cancers in humans from oral or inhalation exposure to a dose of a carcinogenic chemical. For more information see https://rais.ornl.gov/tutorials/toxvals.html

3.4.Endpoint units:

Noncarcinogen/ Carcinogen

3.5.Dependent variable:

The dependent variable is cancerogenic effect on human, as binary classification: 0 (non-carcinogen), 1 (carcinogen)

3.6.Experimental protocol:

NA

3.7.Endpoint data quality and variability:

The model has been developed using data from the Risk Assessment Information System (RAIS)Toxicity values database (https://rais.ornl.gov). Data cover different pollutants categories including organic and inorganic compounds such as dioxins, polycyclic aromatic hydrocarbons (PAHs, pesticides and metals frequently found in contaminated sites.1110 values for mono-constituent organic substances were retrieved for oral slope factor (OSF, mg/kg-day)-1) The RAIS database include the oral slope factor (OSF) values only for chemicals with carcinogenic effects, so chemicals with a defined value (in our case OSF) were considered carcinogenic, and compounds with no value were considered non-carcinogenic.

Other relevant references regarding this endpoint in 9.2 [4].

Canonical SMILES were retrieved for each chemical from two sources (JChem for Office and ChemID plus) then chemicals showing incongruences between the various sources were rejected. Most of the QSAR models cannot handle inorganic compounds, metals and metal complexes, organic salts and data related to mixtures and these compounds have been also rejected. Ionized structures were neutralized and counterions eliminated. The datasets were further checked for the presence of duplicates. The final dataset for the classification model included 593 compounds.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Classification and regression trees (CART)

4.2.Explicit algorithm:

CART implementation in R (rpart package) Recursive partitioning for classification, regression and survival trees. An implementation of most of the functionality of the 1984 book by Breiman, Friedman, Olshen and Stone [2]

4.3.Descriptors in the model:

[1]nS number of Sulfur atoms

[2]nCIC number of rings (cyclomatic number)

[3]ATSC6s Centred Broto-Moreau autocorrelation of lag 6 weighted by I-state

[4]P_VSA_logp_6 P_VSA-like on LogP, bin 6

[5]SpMax_EA(dm) leading eigenvalue from edge adjacency mat. weighted by dipole moment

[6]B02[C-N] Presence/absence of C - N at topological distance 2

[7]B09[C-F] Presence/absence of C - F at topological distance 9

4.4.Descriptor selection:

An in-house tool developed in the R statistical platform has been used to select the best descriptors set and size to be employed for the final model. The approach was based on a forward selection technique: starting from the descriptor most correlated with the experimental data, at each iteration the descriptor leading to the best model (among all the available descriptors) was added, until the size of 25 descriptors. Models have been built with a Linear Discriminant Analysis (LDA) modelling and applied with a bootstrap cross-validation approach (n = 100 iterations). The fitness function has been calculated for each model as a linear combination of the mean values of the accuracy, sensitivity and specificity obtained from the models built in each bootstrap iteration. This function has been used to select the best descriptor to be added to proceed to the next iteration.

4.5.Algorithm and descriptor generation:

The descriptors have been generated with VEGA software using CDK libraries

4.6.Software name and version for descriptor generation:

VEGA Alberto Manganaro (info@kode-solutions.net)www.vega-hub.com

4.7.Chemicals/Descriptors ratio:

593 (training)/7 (descriptors) = 84

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The Applicability Domain (AD) is assessed using the original algorithm implemented within VEGA. An overall AD index is calculated, based on a number of parameters, which relate to the results obtained on similar chemicals within the training and test sets and is defined in this way for this QSAR model´s predictions:

If 1 ≥ AD index > 0.8, the predicted substance is into the Applicability Domain of the model. It corresponds to good reliability of prediction.

If 0.8 ≥ AD index > 0.6, the predicted substance could be out of the Applicability Domain of the model. It corresponds to moderate reliability of prediction.

If AD index ≤ 0.6, the predicted substance is out of the Applicability Domain of the model and corresponds to low reliability of prediction.

Indices are calculated on the first k = 2 most similar molecules, each having S_k similarity value with the target molecule.

Similarity index (*IdxSimilarity*) is calculated as:

$$\frac{\sum_k S_k}{k} \times (1 - Diam^2)$$

where *Diam* is the difference in similarity values between the most similar molecule and the *k*-th molecule.

Accuracy index (*IdxAccuracy*) is calculated as:

$$\frac{\sum_c log\ (1 + S_c)}{\sum_k log\ (1 + S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the prediction of the model matches with the experimental value of the molecule.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c log\ (1 + S_c)}{\sum_k log\ (1 + S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the experimental value of the molecule matches with the prediction made for the target molecule.

ACF contribution (*IdxACF*) index is calculated as

$$ACF = rare \times missing$$

where: *rare* is calculated on the number of fragments found in the molecule and found in the training set in less than 3 occurences as following: if the number is 0, *rare* is set to 1.0; if the number is 1, *rare* is set to 0.6; otherwise *rare* is set to 0.4

missing is calculated on the number of fragments found in the molecule and never found in the training set as following: if the number is 0, *missing* is set to 1.0; if the number is 1, *missing* is set to 0.6; otherwise *missing* is set to 0.4

Descriptors Range (*IdxDescRange*) index is calculated as 1.0 if all molecular descriptors used in the prediction fall within the range of descriptors used in the whole training set, 0.0 otherwise.

AD final index is calculated as following:

$$ADI = (IdxSimilarity^{0.5} \times IdxAccuracy^{0.25} \times IdxConcordance^{0.25}) \times IdxACF \times IdxDescRange$$

5.2. Method used to assess the applicability domain:

The applicability domain and chemical similarity is measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [3]. The VEGA AD also evaluates the correctness of the prediction on similar compounds (accuracy), the consistency between the predicted value for the target compound and the experimental values of the similar compounds, the range of the descriptors, and the presence of unusual fragments, using atom centred fragments.

These indices are defined in this way for this QSAR model:

Similar molecules with known experimental value:

This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are:

If 1 ≥ index > 0.80, strongly similar compounds with known experimental value in the training set have been found

If 0.80 ≥ index > 0.6, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.6, no similar compounds with known experimental value in the training set have been found

Accuracy (average error) of prediction for similar molecules:

This index takes into account the classification accuracy in prediction for the two most similar compounds found. Values near 1 mean that the predicted compounds fall in an area of the model's space where the model gives reliable predictions (no misclassifications), otherwise the lower is the value, the worse the model behaves. Defined intervals are:

If index < 0.6, accuracy of prediction for similar molecules found in the training set is good

If 0.8 > index ≥ 0.6, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 0.8, accuracy of prediction for similar molecules found in the training set is not adequate

Concordance for similar molecules (average difference between target compound prediction and experimental values of similar molecules).:

This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:

If index < 0.6, molecules found in the training set have experimental values that agree with the target compound predicted value

If 0.8 > index ≥ 0.6, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 0.8, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

Model descriptors range check:

This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range. Defined intervals are:

If index = True, descriptors for this compound have values inside the descriptor range of the compounds of the training set

If index= False, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

Atom Centered Fragments similarity check:

This index takes into account the presence of one or more fragments that aren't found in the training set, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur less than three times in the training set), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND. Defined intervals are:

If index = 1, all atom centered fragment of the compound have been found in the compounds of the training set

If 1 > index ≥ 0.7, some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments

If index < 0.7, a prominent number of atoms centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

5.3.Software name and version for applicability domain assessment:

VEGA (www.vegahub.eu)

5.4.Limits of applicability:

The model is not applicable to inorganic chemicals and substances containing unusual elements (i.e., different from C, O, N, S, P, Cl, Br, F, I). Salts can be predicted only if converted to the neutralized form.

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

6.3.Data for each descriptor variable for the training set:

All

6.4.Data for the dependent variable for the training set:

All

6.5.Other information about the training set:

Chemicals with a defined value of OSF (Oral Slope Factor) were considered carcinogenic, and compounds with no value were considered non-carcinogenic.

6.6.Pre-processing of data before modelling:

Canonical SMILES were retrieved for each chemical from two sources (JChem for Office and ChemID plus) then chemicals showing incongruences between the various sources were rejected. Most of the QSAR models cannot handle inorganic compounds, metals and metal complexes, organic salts and data related to mixtures and these compounds have been also rejected. Ionized structures were neutralized and counterions eliminated. The datasets were further checked for the presence of duplicates. The final dataset for the classification model included 593 mono-organic constituent compounds.

6.7.Statistics for goodness-of-fit:

Training set (593 chemicals, 257 positive, 336 negative)

Accuracy = 0.81 Sensitivity = 0.82 Specificity = 0.79

TP 211, TN 267, FP 69, FN 46

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

NA

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

NA

6.10.Robustness - Statistics obtained by Y-scrambling:

NA

6.11.Robustness - Statistics obtained by bootstrap:

NA

6.12.Robustness - Statistics obtained by other methods:

NA

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

7.3.Data for each descriptor variable for the external validation set:

All

7.4.Data for the dependent variable for the external validation set:

All

7.5.Other information about the external validation set:

NA

7.6.Experimental design of test set:

For the training/test splitting, constitutional and ring descriptors blocks have been used, together with the experimental class value, as input for a principal components analysis (PCA). The first principal component (PC) has been used to rank the compounds, then a venetian blind approach has been used to split training and test set compounds with an 80-20% ratio

7.7.Predictivity - Statistics obtained by external validation:

Test set (149 chemicals, 58 positive, 91 negative)

Accuracy = 0.76 Sensitivity = 0.76 Specificity = 0.76

TP 44, TN 69, FP 22, FN 14

Considering ADI (see 5.1) thresholds:

Test set in AD: n=64; Sensitivity 0.90; Specificity 1.0; balanced accuracy 0.95;

TP 27; TN 34; FP 0; FN 3

Test set "could be out AD": n=47; Sensitivity 0.61; Specificity 0.57; balanced accuracy 0.59;

TP 11; TN 16; FP 10; FN 4

Test set out AD: n=39; Sensitivity 0.60; Specificity 0.66; balanced accuracy 0.63;

TP 6; TN 19; FP 10; FN 4

7.8.Predictivity - Assessment of the external validation set:

NA

7.9.Comments on the external validation of the model:

NA

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The mechanistic approach of the present model is supported by the use of the specific atoms, bonds, and molecular fragments involved in the model descriptors

8.2.A priori or a posteriori mechanistic interpretation:

The cyclomatic number (nCIC) identifies the number of rings contained in the molecule. The descriptor is related to the high carcinogenicity potency associated to ring complexes high number of rings in the same molecule, as typically occur in such as PAHs, that are carcinogenic through the formation of epoxides.

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1] Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, andScreening Tools: The VEGAHUB Platform as an Example. Advances in Computational Toxicology:Springer; 2019. p. 365-81

[2] Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). Classification And Regression Trees (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

[3] Floris, M., Manganaro, A., Nicolotti, O. et al. A generalizable definition of chemical similarity for read-across. J Cheminform 6, 39 (2014). https://doi.org/10.1186/s13321-014-0039-1

[4] Review of U.S. EPA's ORD Staff Handbook for Developing IRIS Assessments: 2020 Version (2022); EPA/600/R-20/137 www.epa.gov/ord ORD Staff Handbook for Developing IRIS Assessments Version 1.0 November 2020, EPA/630/P-02/002F December 2002 Final Report A REVIEW OF THE REFERENCE DOSE AND REFERENCE CONCENTRATION PROCESSES, EPA/630/P-02/002F December 2002 Final Report A REVIEW OF THE REFERENCE DOSE AND REFERENCE CONCENTRATION PROCESSES, EPA/630/P-03/001F March 2005 Guidelines for Carcinogen Risk Assessment Risk Assessment Forum U.S. Environmental Protection Agency Washington, DC

9.3.Supporting information:

Training set(s)Test set(s)Supporting information:

All available dataset are present in the model inside the VEGA software.

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC