	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **CASE Ultra model for Female Mouse Carcinogenicity (CARC_MOUSE_F), v.1.9.0.8.1361.400**
	Printing Date: **Sep 25, 2023**

1.QSAR identifier

1.1.QSAR identifier (title):

CASE Ultra model for Female Mouse Carcinogenicity (CARC_MOUSE_F), v.1.9.0.8.1361.400

1.2.Other related models:

This model is part of the set of CASE Ultra Carcinogenicity models: Male Mouse Carcinogenicity (CARC_MOUSE_M), Female Rat Carcinogenicity (CARC_RAT_F), Male Rat Carcinogenicity (CARC_RAT_M)

1.3.Software coding the model:

CASE Ultra Version 1.9.0.8

QSAR based bioactivity and toxicity prediction software.

sales@multicase.com, MultiCASE Inc, 5885 Landerbrook Dr. #210 Mayfield Heights, OH 44124 USA www.multicase.com

http://www.multicase.com/case-ultra

2.General information

2.1.Date of QMRF:

July 10th, 2019

2.2.QMRF author(s) and contact details:

[1]Gianna Cioffi MultiCASE Inc +1-440-565-7221 gianna@multicase.com www.multicase.com

[2]Mounika Girireddy MultiCASE Inc +1-440-565-7221 girireddy@multicase.com www.multicase.com

2.3.Date of QMRF update(s):

September 22nd 2023

2.4.QMRF update(s):

- Added data from FDA drug labels
- Rebuild and revalidated models

2.5.Model developer(s) and contact details:

MultiCASE Inc. +1-440-565-7221 girireddy@multicase.com, saiakhov@multicase.com, chakravarti@multicase.com www.multicase.com

2.6.Date of model development and/or publication:

September 21st 2023

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

[1]FDA CDER Archives

[2]FDA Drug Labels https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files

[3]Matthews EJ, Contrera JF. A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software. Regulatory Toxicology and Pharmacology 1998, 28:242-64. www.ncbi.nlm.nih.gov/pubmed/10049796

[4]Contrera JF, Kruhlak NL, Matthews EJ, Benz RD. Comparison of MC4PC and MDL-QSAR rodent carcinogenicity predictions and the enhancement of predictive performance by combining QSAR models. Regulatory Toxicology and Pharmacology 2007, 49:172-82.

www.ncbi.nlm.nih.gov/pubmed/17703860
[5]Kruhlak NL, Guo D, Cross KP, Stavitskaya L. Enhanced (Q)SAR models for prediction rodent carcinogenicity. Abstracts of Papers, 54th Society of Toxicology Annual meetings, San Diego CA, March 22-26, 2015, poster presentation.

2.8.Availability of information about the model:

Model is commercial. Although the training set is not publicly available, Information about the non-proprietary training set chemicals, assay conditions and details, information about the alerts are available through CASE Ultra interface. For any other specific details contact: sales@multicase.com, MultiCASE Inc. 5885 Landerbrook Dr. #210 Mayfield Heights, OH 44124 USA. Phone: +1-440-565-7221.

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

none

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Mouse

3.2.Endpoint:

Carcinogenicity, Female Mouse

3.3.Comment on endpoint:

The objective of a long-term carcinogenicity study is to observe test animals for a major portion of their life span for the development of neoplastic lesions during or after exposure to various doses of a test substance by an appropriate route of administration. This test is intended primarily for use with rats and mice, and for oral administration. Both sexes should be used. Each dose group and concurrent control group should contain at least 50 animals of each sex. At least three dose levels and a concurrent control should be used. Animals are dosed with the test substance daily (oral, dermal or inhalation administration) and the mode of exposure should be adjusted according to the toxicokinetic profile of the test substance. The duration of the study will normally be 24 months for rodents. For specific strains of mice, duration of 18 months may be more appropriate. Termination of the study should be considered when the number of survivors in the lower dose groups or the control group falls below 25 per cent. The results of these studies include measurements (weighing, food consumption), and, at least, daily and detailed observations, as well as gross necropsy and histopathology.

3.4.Endpoint units:

Binary score

3.5.Dependent variable:

The binarized carcinogenic potential of the test chemical in the form of compounds exhibiting carcinogenicity (1) and not exhibiting carcinogenicity (0)

3.6.Experimental protocol:

Rodent Carcinogenicity models are based on two-year animal studies conducted according to ICH S1A, ICH S1B, and OECD TG-451 regulatory guidelines.

3.7.Endpoint data quality and variability:

High quality curated data. Structures of training chemicals and names were verified. Duplicates were removed. Mixtures components were manually reviewed and used as separate entries if applicable.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR model with binary classification ability. Consists of a logistic regression model with molecular fragment/substructures as the descriptors. The descriptors cover both potentiating and deactivating/mitigating molecular features for compounds exhibiting or not-exhibiting carcinogenic potential. The molecular features related to carcinogenicity were identified from training data using various machine learning techniques.

4.2.Explicit algorithm:

Logistic regression Binary QSAR

Training: Multi-parameter logistic regression modeling with occurrence of sub-structural features as independent and binary carcinogenic potential as dependent variables. Prediction: Application of the logistic regression model using the identification of structural features in the query compounds and computing carcinogenic potential using the fitted parameters of the model.

4.3.Descriptors in the model:

Molecular fragment based descriptors.

4.4.Descriptor selection:

An initial pool of approximately thousands of molecular fragment descriptors were subjected to a descriptor selection process which picks up a subset of the fragments with positive and negative contributions so as to give the best predictive ability to the whole model. The final model contains 182 alerts.

4.5.Algorithm and descriptor generation:

Descriptors for this model are atom-centered molecular fragments which are generated from the training set compounds systematically creating a dictionary of unique fragments. After selecting a few most relevant fragments, a statistical logistic regression data fitting was applied between the X and Y variables to give the final model.

4.6.Software name and version for descriptor generation:

CASE Ultra V 1.9.0.8

QSAR based bioactivity and toxicity prediction software.

sales@multicase.com, MultiCASE Inc, 5885 Landerbrook Dr. #210 Mayfield Heights, OH 44124 USA www.multicase.com

http://www.multicase.com/case-ultra

4.7.Chemicals/Descriptors ratio:

No of chemicals = 1361 (546 Positives/815 Negatives)

No. of descriptors = 182

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model is defined by the chemical space based on small fragments from the training set chemicals and range in the computed prediction probabilities where the model has reasonable ability to differentiate between compounds exhibiting carcinogenicity and not exhibiting carcinogenicity.

5.2.Method used to assess the applicability domain:

The CASE Ultra program evaluates automatically whether a tested molecule is within the domain of applicability of the model it is tested with. A combination of two criteria were used:

1. Checking for 3-atom fragments that are not present in the training chemicals, and
2. Calculated prediction probabilities that fall between 0.3-0.5 where the model has weakest differentiability.

5.3.Software name and version for applicability domain assessment:

CASE Ultra Version 1.9.0.8

QSAR software for modeling and predicting bio-activity of chemicals

sales@multicase.com, MultiCASE Inc, 5885 Landerbrook Dr. #210 Mayfield Heights, OH 44124 USA www.multicase.com, +1-440-565-7221

http://www.multicase.com/case-ultra

5.4.Limits of applicability:

Inorganic compounds, mixtures and large biomolecules are not covered.

In addition,

1. Test chemicals with 3-atom fragments that are not present in the training chemicals potentially are out-of-domain, and
2. Test chemicals with computed prediction probabilities between 0.3 - 0.5 are in grey zone

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: Yes

NanoMaterial: null

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:

Within CASE Ultra interface, all the fragment descriptors are supported by the training chemicals. Every descriptor is supported by relevant statistics, e.g. number of positive and negative training chemicals that contain the fragment. Training set chemicals are explained with assay type, assay conditions, scientific publications etc.

6.6.Pre-processing of data before modelling:

1. Extracted new data from FDA drug safety label pdfs, also included data from existing carcinogenicity models (FDA CDER archives).

2. Verification of chemical structures, registry numbers, CID and names.

2. Duplicates were removed.

3. Mixture components were treated on case by case basis, components if necessary were separated and assigned activity if possible.

6.7.Statistics for goodness-of-fit:

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

not performed

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

Leave 10% out 10 times: Sensitivity 54.9%, Specificity 80.6%, Positive predictivity 66%, Negative predictivity 72.4%, Coverage 77.8%, Accuracy 0.695, Classification cut-off 0.40

6.10.Robustness - Statistics obtained by Y-scrambling:

Y-Scrambling 10% out 10 times: Sensitivity 29%, Specificity 73.7%, Positive predictivity 44.6%, Negative predictivity 59%, Coverage 73.7%, Accuracy 0.513, Classification cut-off 0.40

6.11.Robustness - Statistics obtained by bootstrap:

Bootstrapping 10% out 100 times: Sensitivity 54.9%, Specificity 79.6%, Positive predictivity 66.2%, Negative predictivity 71%, Coverage 76.8%, Accuracy 0.689, Classification cut-off 0.40

Bootstrapping 10% out 10 times: Sensitivity 56.8%, Specificity 81.2%, Positive predictivity 68.2%, Negative predictivity 72.4%, Coverage 77.4%, Accuracy 0.703, Classification cut-off 0.40

6.12.Robustness - Statistics obtained by other methods:

n/a

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No

7.2.Available information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

NanoMaterial: null

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

No

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

No

7.5.Other information about the external validation 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:

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The most important mechanistic basis of the model are the identified molecular substructural features that are part of the model. It should be noted that these features were mined automatically from the training data during the model building process. CASE Ultra models do not have any predefined knowledge of molecular mechanism that explains the activity of a molecule.

8.2.A priori or a posteriori mechanistic interpretation:

The mechanistic basis of the model was neither determined a priori nor a posteriori. The selected features were mined completely automatically from the training data during the model building process and they agree very well with the known chemical mechanisms of carcinogenicity. The training structures were also not selected with any specific mechanism in mind.

8.3.Other information about the mechanistic interpretation:

None

9.Miscellaneous information

9.1.Comments:

This model should be useful in the risk assessment of identifying compounds causing carcinogenicity. It will also be helpful in understanding various known and unknown mechanisms of compounds causing carcinogenicity. It will be particularly helpful in regulatory submission. When a prediction is made using this model in CASE Ultra program, the identified alerts (if any) are highlighted in the query chemical which is helpful in interpreting the results.

9.2.Bibliography:

[1]Optimizing predictive performance of CASE Ultra expert system models using the applicability domains of individual toxicity alerts; Chakravarti, S.K., Saiakhov, R.D. and Klopman, G., Journal of Chemical Information and Modeling, 2012, 52, 2609-2618. DOI: 10.1021/ci300111r

[2]Effectiveness of CASE Ultra Expert System in Evaluating Adverse Effects of Drugs; Saiakhov, R.D., Chakravarti, S.K. and Klopman, G.; Molecular Informatics, 2012, 32, 87-97. DOI: 10.1002/minf.201200081

[3]Computing similarity between structural environments of mutagenicity alerts, Chakravarti, S.K.,Saiakhov, R. D.; Mutagenesis, October 20, 2018, DOI: https://doi.org/10.1093/mutage/gey032

9.3.Supporting information:

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

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