

	QMRF identifier (JRC Inventory): To be entered by JRC
	QMRF Title: Apical Cardiotoxicity Model V1.0
	Printing Date: January 23 2025

1. QSAR identifier

1.1. QSAR identifier (title):

Apical Cardiotoxicity Model V1.0

1.2. Other related models:

- Inhibition Mitochondrial Complexes Model V2.0
- Mitochondrial Dysfunction Model V1.0
- Oxidative stress Model V1.0
- Aryl Hydrocarbon Receptor Model V1.0
- hERG channel blockade Model V1.0

1.3. Software coding the model:

The alternative cloud Platform <https://platform.alternative-project.eu/>

VEGA platform <https://www.vegahub.eu/portfolio-item/vega-qsar/>

2. General information

2.0. Abstract:

Drug-induced cardiotoxicity (DICT) is one of the leading causes of drug attrition in clinical trials or withdrawal from the market. Many studies have been conducted to detect DICT in the early stage of drug development and clinical diagnosis, but the success is limited, as evident by the high attrition rate at all clinical phases due to DICT.

2.1. Date of QMRF:

January 2025

2.2. QMRF author(s) and contact details:

Edoardo Luca Viganò – Istituto di ricerche farmacologiche Mario Negri – edoardo.vigano@marionegri.it

2.3. Date of QMRF update(s):

NA

2.4. QMRF update(s):

NA

2.5. Model developer(s) and contact details:

Edoardo Luca Viganò – Istituto di ricerche farmacologiche Mario Negri – edoardo.vigano@marionegri.it

2.6. Date of model development and/or publication:

2024

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

NA

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:

Human: Drug-induced cardiotoxicity (DICT).

3.2. Endpoint:

Apical Cardiotoxicity drugs side effects report from FDA (U.S. Food and Drug Administration): Drug-induced cardiotoxicity (DICT).

3.4. Endpoint units:

Adimensional

3.5. Dependent variable:

The dependent variable is DICT, as binary classification: 0 (non-DICT), 1 (DICT)

3.6. Experimental protocol:

Drugs side effects reported from FDA (U.S. Food and Drug Administration).

3.7. Endpoint data quality and variability:

The last data source we use is a dataset of drugs annotated with ranked drug-induced cardiotoxicity risk in humans. These data were collected by utilizing labeling documents for FDA (U.S. Food and Drug Administration)-approved drugs. To the best of our knowledge, this is the largest dataset of drugs annotated with ranked DICT risk in humans (DICTrank). The dataset initially consisted of 1,318 drugs, classified as follows: Most-DICT-Concern (341), Less-DICT-Concern (528), No-DICT-Concern (343), and Ambiguous-DICT-Concern (106; lacking sufficient information in the labeling document to determine cardiotoxicity potential). Drugs in the Ambiguous category were removed, and the Less-DICT-Concern and Most-DICT-Concern categories were combined and defined as "active" to create a binary classification endpoint.

<https://doi.org/10.1101/2023.07.06.548029>.

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

K Nearest Neighbors Classifier metric=manhattan, n_neighbors=3, weights=distance

4.2. Explicit algorithm:

The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

4.3. Descriptors in the model:

Molecular descriptors are not used. Instead, the models employ the semantic and grammatical concepts behind SMILES notations to encode chemical information suitably for modeling using CDDD descriptors [1].

4.4. Descriptor selection:

Variable selection was made using the Mutual information (MI) concept. MI between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

The function relies on nonparametric methods based on entropy estimation from k-nearest neighbors distances.

4.5. Algorithm and descriptor generation:

CDDD descriptors [1].

4.6. Software name and version for descriptor generation:

Python 3.11.7 Scikit-Learn 1.4.0 Rdkit 2023.09.4

4.7. Chemicals/Descriptors ratio:

10

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.

ADI is defined in this way for this QSAR model's predictions:

If $1 \geq \text{AD index} > 0.80$, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to "good reliability" of prediction.

If $0.80 \geq \text{AD index} > 0.6$, the predicted substance could be out of the Applicability Domain of the model. It corresponds to "moderate reliability" of prediction.

If $\text{AD index} \leq 0.6$, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to "low reliability" of prediction.

5.2. Method used to assess the applicability domain:

The Applicability domain 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 [1]. 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 $1 \geq \text{index} > 0.80$, strongly similar compounds with known experimental value in the training set have been found.

If $0.80 \geq \text{index} > 0.6$, only moderately similar compounds with known experimental value in the training set have been found.

If $\text{index} \leq 0.6$, 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 $\text{index} < 0.6$, accuracy of prediction for similar molecules found in the training set is good

If $0.8 > \text{index} \geq 0.6$, accuracy of prediction for similar molecules found in the training set is not optimal

If $\text{index} \geq 0.8$, accuracy of prediction for similar molecules found in the training set is not adequate

Concordance for 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 $\text{index} < 0.6$, molecules found in the training set have experimental values that agree with the target compound predicted value

If $0.8 > \text{index} \geq 0.6$, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If $\text{index} \geq 0.8$, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

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 $\text{RARE} * \text{NOTFOUND}$. Defined intervals are:

If $\text{index} = 1$, all atom centered fragment of the compound have been found in the compounds of the training set

If $1 > \text{index} \geq 0.7$, some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments

If $\text{index} < 0.7$, a prominent number of atom 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:

Scikit-Learn 1.4.0, RDKit 2023.09.4 VEGA (www.vegahub.eu)

5.4. Limits of applicability:

The model is not applicable to inorganic chemicals and salts.

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

Smiles: Yes

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:

n 846

6.6. Pre-processing of data before modelling:

The SMILES retrieved are then curated using the preprocess methods developed by winter at all [1].

6.7. Statistics for goodness-of-fit:

F1-Score: 0.72 ± 0.04 (10 fold cross validation median value and standard deviation)

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:

F1-Score: 0.72 ± 0.04 (10 fold cross validation median value and standard deviation)

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

Smiles: Yes

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:

n: 169

7.6. Experimental design of test set:

NA

7.7. Predictivity - Statistics obtained by external validation:

Balanced Accuracy 0.66

Precision 0.85

Sensitivity 0.63

Specificity 0.70

MCC 0.30

F1-Score: 0.72

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:

NA

8.2. A priori or a posteriori mechanistic interpretation:

NA

8.3. Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1. Comments:

NA

9.2.Bibliography:

[1] Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. Chem. Sci. 2019, 10, 1692–1701.

9.3. Supporting information:

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

All available datasets are present in alternative cloud platform and will be in the model inside the VEGA software.