	QMRF identifier (JRC Inventory): **Q17-412-0026**
	QMRF Title: **Lazar models for carcinogenic potency (TD50) in the rat and mouse**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

Lazar models for carcinogenic potency (TD50) in the rat and mouse

1.2.Other related models:

1.3.Software coding the model:

Lazar

opentox-ruby v3.1.0

http://in-silico.ch

http://github.com/opentox

2.General information

2.1.Date of QMRF:

19/03/2014

2.2.QMRF author(s) and contact details:

Elena Lo Piparo Nestlé Research Center elena.lopiparo@rdls.nestle.com

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Andreas Maunz andreas@maunz.de

2.6.Date of model development and/or publication:

July 2014 (publication)

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

[1]Lo Piparo E, Maunz A, Helma C, Vorgrimmler D & Schilter B (2014). Automated and reproducible read-across like models for predicting carcinogenic potency. Regulatory Toxicology and Pharmacology 70: 370–378. http://www.ncbi.nlm.nih.gov/pubmed/25047023

[2]Maunz A, Gütlein M, Rautenberg M, Vorgrimmler D, Gebele D, Helma C (2013). Lazar: a modular predictive toxicology framework. Frontiers in Pharmacology 4, 38. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3669891/

2.8.Availability of information about the model:

The models are proprietary but our aim is to make them freely available in the near future thought a user friendly internet web site. The training and test set are available.

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

No

3.Defining the endpoint - OECD Principle 1

3.1.Species:

mouse and rat

3.2.Endpoint:

4.Human Health Effects 4.12.Carcinogenicity

3.3.Comment on endpoint:

carcinogenic potency (TD50) = daily dose that causes a tumor type in 50% of the exposed animals that otherwise would not develop the tumor in a standard lifetime

3.4.Endpoint units:

mg/kg/day

3.5.Dependent variable:

pTD50= -log(TD50/1000*MW)

3.6.Experimental protocol:

The datasets were composed from CPDB entries by Bercu et al. 2010 (Regul. Tox. and Pharmacol. 57, 300-306) available in supplementary material for download.

3.7.Endpoint data quality and variability:

The data consist of two datasets, one for rat and one for mouse, each being split into 90% training and 10% test data. The split was done by selecting every tenth compound from the full data, sorted on TD_{50} values, which allowed full coverage of training TD_{50} values in the test set. Dividing by molecular weight transforms the cancer potency value on a molar basis. This study made no changes to the data whatsoever, neither to compounds nor to activity values. Therefore the dataset employed by this model, such as the one from Bercu *et al.*, contains a total of 460 training set plus 51 test set compounds for the rat, and 362 training set plus 40 test set compound for the mouse.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Automated read-across

4.2.Explicit algorithm:

Locally weighted Support Vector Machine (SVM) regression

see equation

Lazar searches the training set with chemical structures and experimental measurements (pTD50) for neighbour compounds (similar to the current query structure) and calculates a prediction from the experimental measurements of the neighbours. Calculating the prediction is in three steps: 1. Identification of similar compounds in the training dataset (neighbours). 2. Creation of a local or read-across model for predictions based on structures and experimental activities of these neighbours. 3. Application of the local or read-across model to predict the activity of the query compound.

4.3.Descriptors in the model:

[1]Largest Chain

[2]Aromatic Bonds Count

[3]Longest Aliphatic Chain

[4]Rule Of Five

[5]Atom Count

[6]XLogP

[7]ALOGP
[8]Aromatic Atoms Count
[9]Mannhold LogP
[10]Bond Count
[11]Rotatable Bonds Count
[12]Largest Pi System
[13]APol
[14]BPol
[15]H-Bond Acceptor Count
[16]H-Bond Donor Count
[17]CPSA
[18]Chi Path
[19]Fragment Complexity
[20]Kier-Hall Smarts
[21]Kappa ShapeIndices
[22]Petitjean Number
[23]Autocorrelation Mass
[24]VAdjMa
[25]Chi Path Cluster
[26]Wiener Numbers
[27]Autocorrelation Polarizability
[28]Carbon Types, Eccentric Connectivity Index
[29]Chi Chain
[30]MDE
[31]Petitjean Shape Index
[32]TPSA
[33]Chi Cluster
[34]Zagreb Index
[35]Autocorrelation Charge

4.4.Descriptor selection:

Recursive Feature Elimination (RFE) was applied to cut down on the number of features. RFE first learns a model on all features and on the complete training data, thereby ranking features according to their influence on the model. Then, it learns several models on the top-k features, for several values of k, and validates each one on some held out data in order to determine a best feature selection.

4.5.Algorithm and descriptor generation:

All descriptors were calculated by OpenTox compliant descriptor calculation services using Lazar (opentox-ruby v3.1.0). The services in turn employ publicly available software libraries such as the Chemistry Development Kit (CDK), OpenBabel, and Joelib.

4.6.Software name and version for descriptor generation:

Chemistry Development Kit (CDK), v1.4.17

The Chemistry Development Kit (CDK) is a Java library for structural chemo- and bioinformatics. It is now developed by more than 50 developers all over the world and used in more than 10 different academic as well as industrial projects world wide.

Egon Willighagen (egon.willighagen@maastrichtuniversity.nl)
http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page

Open Babel, v2.3.2
Open Babel is a chemical toolbox designed to speak the many languages of chemical data. It's an open, collaborative project allowing anyone to search, convert, analyze, or store data from molecular modeling, chemistry, solid-state materials, biochemistry, or related areas.
http://openbabel.org/wiki/Main_Page

JOElib2, v20090613
JOELib/JOELib2 is a cheminformatics library which supports SMARTS substructure search, descriptor calculation, processing/filtering pipes, conversion of file formats, 100% pure Java, and interfaces to external programs (e.g. Ghemical) are available.
Joerg Kurt Wegner (me@joergkurtwegner.eu)
http://sourceforge.net/projects/joelib/

4.7.Chemicals/Descriptors ratio:

Not applicable: Model training is done separately for each prediction (instance based learning).Training structures similar to the query structure (*neighbours*) are derived through suitable transformations on the features (involving standardization and normalization), and similarity calculation. Then a model is trained using the neighbours as training set and the query structure is predicted by the model. This process repeats for each query structure from scratch.

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

A prediction based on a large number of neighbors with high similarity and concordant experimental data will be more reliable than a prediction based on a low number of neighbors with low similarity and contradictory experimental results. Hence, the confidence of the lazar algorithm is even more comprehensive than classical applicability domain approach that only considers the feature value space, but not the coherence of the endpoint values. More formally, the confidence of a prediction is defined by the mean neighbour similarity.

If a query molecule is not well represented in the training dataset, it will be outside of the applicability domain of the model and it will have a poor regression statistic. In such cases, Lazar does not make a prediction. Instead it warns the user that the compound was outside the AD.

5.2.Method used to assess the applicability domain:

Lazar features a built-in assessment of applicability domain in the form of a confidence index. The confidence index is a raw, uncalibrated number, not a probability, between 0 and 1. The higher the confidence,

the more reliable the prediction. The confidence of a prediction is defined by the mean neighbour similarity.

5.3.Software name and version for applicability domain assessment:

Lazar
opentox-ruby v3.1.0
http://in-silico.ch
http://github.com/opentox

5.4.Limits of applicability:

Statistical view: As Lazar is an instance-based method, no hard cutoffs can be determined in terms of descriptor values. However, a confidence index is a number calculated with every prediction in Lazar. Based on test set validation, confidence values below 0.55 should be considered unreliable in both the mouse and rat models.

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN: No
Chemical Name: No
Smiles: Yes
Formula: No
INChI: No
MOL file: 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:

The training data was composed from CPDB entries by Bercu et al., who offer them as supplementary material to their article for download. They consist of two datasets, RAT and MOUSE, each being split into 90% training and 10% test data. The split was done by selecting every tenth compound from the full data, sorted on TD50 values, which allowed full coverage of training TD50 values in the test set. The test set sizes were 53 (MOUSE) and 40 (RAT).

6.6.Pre-processing of data before modelling:

Bercu et al. converted TD50 values to pTD50, in order to improve on normality of activity distribution, defined as pTD50=-log(TD50/(1000*Molecular Weight)). This study made no changes to the data whatsoever, neither to compounds nor to activity values.

6.7.Statistics for goodness-of-fit:

Classification Results (hard cutoff, in percent):
Coverage: 82 (RAT), 73 (MOUSE)
Specificity: 71 (RAT), 100 (MOUSE)
Sensitivity: 67 (RAT), 57 (MOUSE)

Concordance: 69 (RAT), 90 (MOUSE)
Positive predictivity: 70 (RAT), 100 (MOUSE)
Negative predictivity: 68 (RAT), 88 (MOUSE)
ROC-Score: 2.33 (RAT), +Infinity (MOUSE)

Classification Results (without indeterminate compounds, in percent):
Coverage: 43 (RAT), 40 (MOUSE)
Specificity: 80 (RAT), 100 (MOUSE)
Sensitivity: 93 (RAT), 80 (MOUSE)
Concordance: 88 (RAT), 94 (MOUSE)Positive predictivity: 88 (RAT), 100 (MOUSE)Negative predictivity: 89 (RAT), 92 (MOUSE)
ROC-Score: 4.65 (RAT), +Infinity (MOUSE)

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

Unfortunately there are no regression statistics in the classical sense of inferential statistics. Machine learning methods employ advanced models that are not described by an equation. It would only be possible to extract SVM model-specific parameters, such as kernel parameters. As individual models are generated for each prediction, the number of models created for a given test dataset is equal to the number of test dataset instances (instance-based learning).

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

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

The quality of the models was determined through statistical parameters such as coverage, specificity, sensitivity, concordance, positive and negative predictivity. For classification, the ROC (Receiver Operating Characteristic) score was calculated to provide an additional measure of the predictive performance of the models.

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: No
Chemical Name: Yes
Smiles: Yes
Formula: No
INChI: No
MOL file: 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:

Test set sizes: 53 compounds for mouse and 40 compounds for rat.

7.6.Experimental design of test set:

The datasets were composed from CPDB entries by Bercu et al., available in supplementary material for download. They consist of two datasets, one for rat and one for mouse, each being split into 90% training and 10% test data. The split was done by selecting every tenth compound from the full data, sorted on TD_{50} values, which allowed full coverage of training TD_{50} values in the test

7.7.Predictivity - Statistics obtained by external validation:

Numerical Predictions:

Percentage of compounds with ratios between predicted and experimental below certain thresholds

<=1-fold: 43 (RAT), 48 (MOUSE)

<=5-fold: 71 (RAT), 86 (MOUSE)

<=10-fold: 76 (RAT), 93 (MOUSE)

Classification Results (hard cutoff, in percent):

Coverage: 82 (RAT), 73 (MOUSE)

Specificity: 71 (RAT), 100 (MOUSE)

Sensitivity: 67 (RAT), 57 (MOUSE)

Concordance: 69 (RAT), 90 (MOUSE)

Positive predictivity: 70 (RAT), 100 (MOUSE)

Negative predictivity: 68 (RAT), 88 (MOUSE)

ROC-Score: 2.33 (RAT), +Infinity (MOUSE)Classification Results (without indeterminate compounds, in percent):

Coverage: 43 (RAT), 40 (MOUSE)Specificity: 80 (RAT), 100 (MOUSE)

Sensitivity: 93 (RAT), 80 (MOUSE)

Concordance: 88 (RAT), 94 (MOUSE)Positive predictivity: 88 (RAT), 100 (MOUSE)

Negative predictivity: 89 (RAT), 92 (MOUSE)ROC-Score: 4.65 (RAT), +Infinity (MOUSE)

7.8.Predictivity - Assessment of the external validation set:

The test set was used as defined by Bercu et al.

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:

It's not possible to provide a mechanistic interpretation because every time a new model is built based on the similar compounds found.

8.2.A priori or a posteriori mechanistic interpretation:

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

The Lazar model is a combination of mathematical and statistical approaches. Therefore, no single equation describing the relation between descriptors and endpoints can be given, as in traditional QSAR. The only way to transparently present the approach is a textual

description and/or reproducing the results via the freely available software and data sets. Training and test sets can be found as supporting information to Bercu et al [sect.9.2; ref.1]

9.2.Bibliography:

[1]Bercu JP, Morton SM, Deahl JT, Gombar VK, Callis CM & van Lier RB (2010). In silico approaches to predicting cancer potency for risk assessment of genotoxic impurities in drug substances. Regulatory Toxicology and Pharmacology 57 (2-3) 300-306. http://ac.elscdn. com/S0273230010000589/1-s2.0-S0273230010000589-main.pdf?_tid=5f6b3dd2-73c6-11e4- 8095-00000aab0f26&acdnat=1416825837_de5f671919e3d82631964c90df39545c

[2]Contrera JF (2011). Improved in silico prediction of carcinogenic potency (TD50) and the risk specific dose (RSD) adjusted Threshold of Toxicological Concern (TTC) for genotoxic chemicals and pharmaceutical impurities. Regulatory Toxicology and Pharmacology 59 (1) 133-141. http://ac.elscdn. com/S0273230010001789/1-s2.0-S0273230010001789-main.pdf?_tid=b9644720-73c6-11e4- b969-00000aab0f26&acdnat=1416825988_b7813e33039b315fba9f4818c640c7b2

[3]Maunz A, Gütlein M, Rautenberg M, Vorgrimmler D, Gebele D, Helma C(2013). lazar: a modular predictive toxicology framework. Frontiers in Pharmacology 4, 38. http://journal.frontiersin.org/Journal/10.3389/fphar.2013.00038/full

[4]Kuhn M (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28 (5) 1-26. http://www.jstatsoft.org/v28/i05

9.3.Supporting information:

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

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q17-412-0026

10.2.Publication date:

2017-09-21

10.3.Keywords:

Lazar;TD50;automated read/across;rat;mouse;

10.4.Comments:

old # Q29-44-39-423