Leadscope Enterprise model for cytochrome P450 isoenzyme 2D6 (CYP2D6) substrates (human clinical data)

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise model for cytochrome P450 isoenzyme 2D6 (CYP2D6) substrates (human clinical data), Danish QSAR Group at DTU Food.

1.2 Other related models

MultiCASE CASE Ultra model for cytochrome P450 isoenzyme 2D6 (CYP2D6) substrates (human clinical data), Danish QSAR Group at DTU Food.

SciMatics SciQSAR model for cytochrome P450 isoenzyme 2D6 (CYP2D6) substrates (human clinical data), Danish QSAR Group at DTU Food.

1.3. Software coding the model

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

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

Sine Abildgaard Rosenberg;

National Food Institute at the Technical University of Denmark;

Trine Klein Reffstrup;

National Food Institute at the Technical University of Denmark;

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.

2.3 Date of QMRF update(s)

2.4 QMRF update(s)

2.5 Model developer(s) and contact details

Tine Ringsted;

National Food Institute at the Technical University of Denmark;

Jay Russel Niemelä;

National Food Institute at the Technical University of Denmark;

Svava Ósk Jónsdóttir;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Eva Bay Wedebye;

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

Roberts, G., Myatt, G. J., Johnson, W. P., Cross, K. P., and Blower, P. E. J. (2000) LeadScope: Software for Exploring Large Sets of Screening Data. *Chem. Inf. Comput. Sci.*, 40, 1302-1314.

Cross, K.P., Myatt, G., Yang, C., Fligner, M.A., Verducci, J.S., and Blower, P.E. Jr. (2003) Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.*, *46,* 4770-4775.

Valerio, L. G., Yang, C., Arvidson, K. B., and Kruhlak, N. L. (2010) A structural feature-based computational approach for toxicology predictions. *Expert Opin. Drug Metab. Toxicol.*, 6:4, 505-518.

2.8 Availability of information about the model

The training set is non-proprietary and was compiled from the published literature (see 6.5 for more details). 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 (primarily clinical data).

3.2 Endpoint

QMRF 5. Toxicokinetics

QMRF 5. 8. Toxicokinetics. Metabolism (including metabolic clearance)

3.3 Comment on endpoint

The cytochrome P450 (CYP) superfamily of heme-containing enzymes plays a significant role in phase I metabolism of a wide range of endogenous compounds and xenobiotics. It is therefore an important factor in drug development and drug therapy to determine if a drug is metabolized by CYP enzymes. Beside drugs, CYP enzymes detoxify environmental compounds and chemicals in consumer products. They also have the ability to form reactive intermediates that can damage DNA, lipids and proteins, and potentially lead to tumor initiation and cancer after long term exposure. The human genome encodes 57 different CYP genes, with five of these enzymes being responsible for the metabolism of 95% of drugs, namely CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4. The modulation of CYP activity by inhibition or induction of drugs or other chemicals can cause problems ranging from insufficient therapeutic effect to fatal toxic consequences.

The isoenzyme CYP2D6 is primarily expressed in the liver and the mucosa of the small intestine and is considered to be involved in the metabolism of 12–15% of prescribed drugs including antidepressants, antipsychotics, some antiarrhythmics, beta-blockers and analgesics. CYP2D6 is associated with substantial genetic polymorphism causing large inter-individual differences in enzyme activity, and the difference in CYP2D6 activity makes some people more vulnerable to CYP2D6 substrates because of inadequate excretion (poor metaboliser (PM)) or rapid excretion (ultrarapid metaboliser (UM)) of drugs. CYP2D6 also has the possibility of forming harmful intermediates.

The metabolism activity of CYP2D6 is, in many cases, facilitated by an ion pair interaction between an aspartic acid residue at the active site and a protonated nitrogen atom of the substrate. Substrates of CYP2D6 contain a basic nitrogen atom 5Å or 7Å from the point of oxidation with aromatic rings positioned coplanar in the substrate.

Data for this model is based on human clinical data for CYP2D6, primarily drugs, gathered from the literature.

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Human cytochrome P450 isoenzyme 2D6 (CYP2D6), substrates or non-substrates.

3.6 Experimental protocol

Data were either obtained from *in vivo* clinical experiments or various *in vitro* models like tissues slices, microsomes, cell cultures and purified and recombinant enzymes that have formed the basis for a clinical decision. Epidemiological observations and case studies may also have served as input for such decisions. Since data is obtained from various sources a common protocol have not been used, please see references in 9.2 for further information on the experimental protocols.

3.7 Endpoint data quality and variability

As data for the training set were compiled from multiple sources and no common experimental protocol has been described a certain degree of variability in data exists. In addition to this the main part of the training set consist of human clinical data and this type of data is associated with a high degree of variability because of the many factors affecting humans, such as genetics, lifestyle etc.

4. Defining the algorithm

4.1 Type of model

A categorical (Q)SAR model based on structural features and numeric molecular descriptors.

4.2 Explicit algorithm

This is a categorical (Q)SAR model made by use of partial logistic regression (PLR). The specific implementation is proprietary within the Leadscope software.

4.3 Descriptors in the model

structural features,

aLogP,

polar surface area,

number of hydrogen bond donors,

Lipinski score,

number of rotational bonds,

parent atom count,

parent molecular weight,

number of hydrogen bond acceptors

4.4 Descriptor selection

Leadscope Predictive Data Miner is a software program for systematic sub-structural analysis of a chemical using predefined structural features stored in a template library, training set-dependent generated structural features (scaffolds) and calculated molecular descriptors. The feature library contains approximately 27,000 pre-defined structural features and the structural features chosen for the library are motivated by those typically found in small molecules: aromatics, heterocycles, spacer groups, simple substituents. Leadscope allows for the generation of training set-dependent structural features (scaffold generation), and these features can be added to the pre-defined structural features from the library and be included in the descriptor selection process. It is possible in Leadscope to remove redundant structural features before the descriptor selection process and only use the remaining features in the descriptor selection process. Besides the structural features Leadscope also calculates eight molecular descriptors for each training set structure: the octanol/water partition coefficient (alogP), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), Lipinski score, atom count, parent compound molecular weight, polar surface area (PSA) and rotatable bonds. These eight molecular descriptors are also included in the descriptor selection process.

Leadscope has a default automatic descriptor selection procedure. This procedure selects the top 30% of the descriptors (structural features and molecular descriptors) according to X^2-test for a binary variable, or

the top and bottom 15% descriptors according to *t*-test for a continuous variable. Leadscope treats numeric property data as ordinal categorical data. If the input data is continuous such as IC_{50} or cLogP data, the user can determine how values are assigned to categories: the number of categories and the cut-off values between categories. (Roberts *et al.*2000).

When developing this model, intermediate models with application of different modelling approaches in Leadscope were tried:

1. 'Single model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors for descriptor selection.
2. 'Single model' using both the Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection.
3. 'Single model' using both Leadscope pre-defined structural features and the training set dependent features (scaffolds generation), with subsequent removal of redundant structural features, and calculated molecular descriptors for descriptor selection.
4. 'Composite model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors in the descriptor selection.
5. 'Composite model' using both Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection.

Based on model performance as measured by a preliminary cross-validation the model developed using approach number 3. was chosen.

For this model scaffolds were generated by Leadscope for the training set structures and added to the Leadscope library of structural features. The number of structural features was then reduced further using the built-in filter to remove similar (reduntant) features (the "less similar" features removed). Descriptors were then automatically selected among the remaining structural features and the eight molecular descriptors.

4.5 Algorithm and descriptor generation

For descriptor generation see 4.4.

After selection of descriptors the Leadscope Predictive Data Miner program performs partial least squares (PLS) regression for a continuous response variable, or partial logistic regression (PLR) for a binary response variable, to build a predictive model. By default the Predictive Data Miner performs leave-one-out or leave-groups-out (in the latter case, the user can specify any number of repetitions and percentage of structures left out) cross-validation on the training set depending on the size of the training set. In the cross-validation made by Leadscope the descriptors selected for the 'mother model' are used when building the validation submodels and they therefore have a tendency to be overfittet and give overoptimistic validation results.

In this model, because of the categorical outcome in the response variable, PLR was used to build the predictive model. For this model 204 descriptors were selected to build the model. These include 8 Leadscope calculated molecular descriptors, 114 hierarchy features, and 82 dynamic features. The 204 descriptors were distributed on 3 PLS factors.

4.6 Software name and version for descriptor generation

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

4.7 Descriptors/chemicals ratio

In this model 204 descriptors were used and distributed on 3 PLS factors. The training set consists of 745 compounds. The descriptor/chemical ratio is 1:3.7 (204:745).

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 of a structural domain in Leadscope and the in-house further probability refinement algorithm on the output from Leadscope to reach the final applicability domain call.

1. Leadscope

For assessing if a test compound is within the structural applicability domain of a given model Leadscope examines whether the test compound bears enough structural resemblance to the training set compounds used for building the model (i.e. a structural domain analysis). This is done by calculating the distance between the test compound and all compounds in the training set (distance = 1 - similarity). The similarity score is based on the Tanimoto method. The number of neighbours is defined as the number of compounds in the training set that have a distance equal to or smaller than 0.7 with respect to the test compound. The higher the number of neighbours, the more reliable the prediction for the test compound. Statistics of the distances are also calculated. Effectively no predictions are made for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated in Leadscope.

2. The Danish QSAR group

In addition to the general Leadscope structural applicability domain definition the Danish QSAR group has applied a further requirement to the applicability domain of the model. That is only positive predictions with a probability equal to or greater than 0.7 and negative predictions with probability equal to or less than 0.3 are accepted. Predictions within the structural applicability domain but with probability between 0.5 to 0.7 or 0.3 to 0.5 are defined as positives out of applicability domain and negatives out of applicability domain, respectively. When these predictions are weeded out the performance of the model in general increases at the expense of reduced model coverage.

5.2 Method used to assess the applicability domain

Leadscope does not generate predictions for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated.

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

5.3 Software name and version for applicability domain assessment

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

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 Leadscope. 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). Calculation 2D structures (SMILES and/or

SDF) are generated by stripping off ions (of the accepted list given above). 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

Yes

6.2 Available information for the training set

CAS

SMILES

6.3 Data for each descriptor variable for the training set

No

6.4 Data for the dependent variable for the training set

All

6.5 Other information about the training set

745 compounds are in the training set: 247 substrates and 498 non-substrates.

Data for the training set was compiled from the following sources: Hodgson (2001), Rendric (2002), Skinner *et al.* (2003), Dong (2005), Koumis and Samuel (2005), Yap and Chen (2005), Skerjanic (2006), Peters *et al.* (2007), Urichuk *et al.* (2008), and Draper (2009).

6.6 Pre-processing of data before modelling

Only structures acceptable for the commercial software were used in the 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.

6.7 Statistics for goodness-of-fit

Not performed.

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

Not performed. (It is not a preferred measurement for evaluating large models).

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

A five times two-fold 50 % cross-validation was performed. This was done by randomly removing 50% of the full training set used to make the “mother model”, where the 50% contains the same ratio of positive and negatives as the full training set. A new model (validation submodel) was created on the remaining 50% using the same settings in Leadscope but with no information from the “mother model” regarding descriptor selection etc. The validation submodel was applied to predict the removed 50% (within the defined applicability domain for the submodel). Likewise, a validation submodel was made on the removed

50% of the training set and this model was used to predict the other 50% (within the defined applicability domain for this submodel). This procedure was repeated five times.

Predictions within the defined applicability domain for the ten validation submodels were pooled and Cooper's statistics calculated. This gave the following results for the 61.4% (2290/(5*746)) of the predictions which were within the applicability domain:

- Sensitivity (true positives / (true positives + false negatives)): 60.0%
- Specificity (true negatives / (true negatives + false positives)): 89.4
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 80.1%

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 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

External validation not performed for this model.

8. Mechanistic interpretation

8.1 Mechanistic basis of the model

The global model identifies structural features and molecular descriptors which in the model development was found to be statistically significant associated with effect. Many predictions may indicate modes of action that are obvious for persons with expert knowledge for the endpoint.

8.2 A priori or posteriori mechanistic interpretation

A posteriori mechanistic interpretation. The identified structural features and molecular descriptors may provide basis for mechanistic interpretation.

8.3 Other information about the mechanistic interpretation

9. Miscellaneous information

9.1 Comments

The model can be used to predict if a chemical is a substrate of the cytochrome P450 2D6 isoenzyme in humans.

9.2 Bibliography

Dong, B.J. (2005) Cinacalcet: an oral calcimimetic agent for the management of hyperparathyroidism. *Clin. Ther.*, 27:4, 1725–1751.

Draper, E. (2009) New drug bulletin: Budesonide and Formoterol (Symbicort-AstraZeneca), University of Utah Hospitals and Clinics, E. Fox, ed., Accessed 12 February 2009, Available at http:// healthcare.utah.edu/pharmacy/bulletins/NDB_136.pdf.

Hodgson, E. (2001) In vitro human phase I metabolism of xenobiotics I: Pesticides and related chemicals used in agriculture and public health, September 2001. *J. Biochem. Mol. Toxicol.,* 15:6, 296–299.

Jónsdóttir, S.Ó., Ringsted, T., Nikolov, N.G., Dybdahl, M, Wedebye, E.B., Niemelä, J.R. (2012). Identification of cytochrome P450 2D6 and 2C9 substrates and inhibitors by QSAR analysis. *Bioorganic & Medicinal Chemistry* 20, 2042–2053.

Koumis, T., and Samuel, S. (2005) Tiotropium bromide: A new long-acting bronchodilator for the treatment of chronic obstructive pulmonary disease. *Clin. Ther*., 27:4, 377–392.

Peters F.T., Dragan, C.A., Wilde, D.R., Meyer, M.R., Zapp, J., Bureik, M., and Maurer, H.H. (2007) Biotechnological synthesis of drug metabolites using human cytochrome P450 2D6 heterologously expressed in fission yeast exemplified for the designer drug metabolite 40-hydroxymethyl-alphapyrrolidinobutyrophenone. *Biochem. Pharmacol.*, 74, 511–520.

Rendic, S. (2002) Summary of information on human CYP enzymes: Human P450 metabolism data. *Drug Metabolism Reviews,* 34(1&2), 83–448.

Ringsted, T., Nikolov, N., Jensen, G.E., Wedebye, E.B., and Niemelä, J. (2009) QSAR Models for P-450 (2D6) Substrate Activity. *SAR and QSAR in Environmental Research*, 20:3, 309-325.

Skerjanec, A. (2006) The clinical pharmacokinetics of darifenacin. *Clin. Pharmacokinet*., 45:4 ,325–350.

Skinner, M.H., Kuan, H.Y., Pan, A., Sathirakul, K., Knadler, M.P., Gonzales, C.R., Yeo, K.P. Reddy, S., Lim, M., Ayan-Oshodi, M., and Wise, S.D. (2003) Duloxetine is both an inhibitor and a substrate of cytochrome P4502D6 in healthy volunteers. *Clin Pharmacol Ther.,* 73:3, 170–177.

Urichuk, L., Prior, T.I., Dursun, S., and Baker, G. (2008) Metabolism of atypical antipsychotics: Involvement of cytochrome P450 enzymes and relevance for drug-drug interactions. *Curr. Drug Metab.,* 9, 410–418.

Yap, C. W., and Chen, Y. Z. (2005) Prediction of Cytochrome P450 3A4, 2D6, and 2C9 Inhibitors and Substrates by Using Support Vector Machines. *J. Chem. Inf. Model,* 45, 982-992.

9.3 Supporting information