Regulatory Toxicology and Pharmacology 70 (2014) 370–378

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Automated and reproducible read-across like models for predicting carcinogenic potency

Elena Lo Piparo [a,*,1], Andreas Maunz [b,1], Christoph Helma [b], David Vorgrimmler [b], Benoît Schilter [a]

[a] Chemical Food Safety Group, Nestlé Research Center, Lausanne, Switzerland
[b] In Silico Toxicology GmbH, Basel, Switzerland

ARTICLE INFO

Article history:
Received 8 April 2014
Available online 15 July 2014

Keywords:
Alternative method
Risk assessment
Quantitative structure activity relationship (QSAR)
Toxicity
Cancer potency (TD_{50})
Genotoxicity
Read-across

ABSTRACT

Several qualitative (hazard-based) models for chronic toxicity prediction are available through commercial and freely available software, but in the context of risk assessment a quantitative value is mandatory in order to be able to apply a Margin of Exposure (predicted toxicity/exposure estimate) approach to interpret the data. Recently quantitative models for the prediction of the carcinogenic potency have been developed, opening some hopes in this area, but this promising approach is currently limited by the fact that the proposed programs are neither publically nor commercially available. In this article we describe how two models (one for mouse and one for rat) for the carcinogenic potency (TD_{50}) prediction have been developed, using lazar (Lazy Structure Activity Relationships), a procedure similar to read-across, but automated and reproducible. The models obtained have been compared with the recently published ones, resulting in a similar performance. Our aim is also to make the models freely available in the near future thought a user friendly internet web site.

1. Introduction

Given increasing pressure to reduce animal testing, alternative methods relating chemical structure to toxicity have been increasingly valued in many regulatory organisations (ECHA, 2011; ICCR, 2012; U.S. EPA, 2008; EFSA, 2010; Arvidson et al., 2010). The contribution of computational toxicology to the future of regulatory decisions in public health has been addressed (NAS, 2007; Rusyn and Daston, 2010; U.S. EPA, 2012) and nowadays computational tools are widely promoted to support regulatory assessments and decision making in the field of food safety (Lo Piparo et al., 2011). In this context read-across has been mentioned as the most actionable short term strategy for reducing animal use.

From a food sector perspective, the application of such approaches may bring significant benefits not only in terms of saving time, cost, and with respect to reduction of use of laboratory animals, but also will open new horizons of risk assessment, giving the possibility of establishing levels of safety concern associated with human exposure to toxicologically uncharacterized chemicals. This is very relevant for both fast decision making (management of emergency safety issues) and priority setting (safety by design in research and development, R&D). Indeed new molecules are continuously identified and quantified in products as a consequence of the impressive improvement of analytical methods, and therefore companies need often to face and manage cases of emerging issues associated with chemicals for which no or little toxicological data are available. Moreover fast preliminary safety evaluations are increasingly required at the beginning of R&D projects for priority setting of potential new ingredients and to design intrinsically safe chemicals (safety by design).

In silico strategies are already integrated in the preclinical screening scheme of pharmaceutical discovery pipelines where an early identification of unacceptable toxicological hazard is a clear competitive advantage (Benfenati et al., 2009). Unfortunately it is difficult to directly transfer and use this expertise to food safety. Indeed the need of the food sector is different, where the most likely application of computational toxicology models would be in the establishment of the level of safety concern associated with the inadvertent/accidental presence of chemicals in finished products. This requires not only qualitative information on the potential hazardous properties of the chemical (e.g. probability that a compound is carcinogenic) but also quantitative information (e.g. carcinogenic potency) allowing a comparison with estimated exposure to establish the level of concern (Schilter et al., 2014).

Several qualitative (hazard-based) models for carcinogenicity prediction are available through commercial and free software, but only few tools are currently available for quantitative prediction. Carcinogenicity has often been considered as a too complex

* Corresponding author.
E-mail address: elena.lopiparo@rdls.nestle.com (E. Lo Piparo).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.yrtph.2014.07.010
0273-2300/

end point (many mechanisms of action involved and little structural commonality) to be adequately modelled and quantitatively predicted.

In this contest, guidance for genotoxic impurities (GTI) was developed by US-FDA. The guidance suggests calculating cancer risk based on carcinogenic potency from a structural similar known carcinogen (USFDA, 2008). In addition recent efforts in the (Q)SAR field have resulted in the development of local quantitative models for the prediction of carcinogenic potency, opening some hopes in this area. Indeed these models provide reasonable predictions with errors within the same order of magnitude than the estimated variability of experimental data. This promising approach is currently limited by the fact that the proposed models are neither publically (Bercu et al., 2010 and Toropov et al., 2009) nor freely available (Contrera, 2011).

In contrast with the use of QSAR tools, generally the application of read-across is a more *ad hoc* approach involving a range of subjective choices in terms of similarity metrics and criteria for analogues selection. In this paper we describe two quantitative models (one for rat and one for mouse) to predict carcinogenic potency of genotoxic compounds by an alternative, automated and reproducible read-across like procedure. The models have been developed using *Lazar* (shortcut for *laz*y structure–*a*ctivity *r*elationships), a modular framework for predictive toxicology (Maunz and Helma, 2008; Maunz et al., 2013). The lazar models have been compared with the recently published ones by Bercu et al. (2010) and Contrera (2011), resulting in a similar performance.

Furthermore to provide transparency and meet regulatory demands the models have been submitted to QMRF (QSAR Model Reporting Format) Database (http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/QRF) and will be made freely available online through a user friendly platform that will provide detailed supporting information to the predicted toxicity values, such as the identification of the similar compounds used to build the model and the prediction confidence.

2. Materials and methods

2.1. Lazar similarity search

Lazar searches a database with chemical structures and experimental data (training set) for compounds similar to the query structure (*neighbours*) and calculates a prediction from the experimental measurements of the neighbours. Therefore it provides predictions for a given query compound in a three-step process (Maunz et al., 2013):

- Identification of similar compounds in the training dataset (neighbours).
- Creation of a local or read-across model for predictions based on structures and experimental activities of these neighbours.
- Application of the local or read-across model to predict the activity of the query compound.

For the determination of toxicity-related chemical similarities it is important to consider only descriptors, or features, that are relevant for the toxic endpoint under investigation. The crucial task is therefore to identify these features. *Lazar* relies on data mining algorithms to identify relevant features automatically from the training data. This procedure is reproducible and saves expensive expert work.

2.2. Statistical learning

In statistical learning theory, overfitting occurs when a statistical model describes noise instead of the underlying relationship. Machine Learning (ML) algorithms, for example Support Vector Machines (SVM) and Random Forests (RF) support strategies to limit the fit to the training data.

SVMs are a class of algorithms where data points are treated as vectors. For classification and regression, the data points are usually mapped to a high-dimensional feature space through kernel functions. SVMs support regularization via an internal cost function (Vapnik and Cortes, 1995).

The RF algorithm incorporates a general strategy for regularization known as bagging (short for bootstrap aggregation) (Breiman, 2001). In bagging, the training data is not processed as a whole by the learning algorithm, but n so-called bootstrap samples are drawn with replacement and trained upon individually. For increasing n, the instances that where not selected for each sample, termed OOB (out-of-bag), will cover around 36% of the data, on average. RF builds a decision tree model for each bootstrap sample to predict the dependent variable, and predicts the OOB data with it to estimate the error rate of the model (Liaw and Wiener, 2002). A RF model consists of a set of n such trees. A prediction for an unknown data point (query compound) is derived by averaging over the individual tree predictions of the forest. RF can fit arbitrarily shaped dependent variables, especially non-linear and non-continuous ones, and is able to handle large amounts of features.

Lazar was designed to handle high-dimensional, numerically unconstrained feature spaces, while maintaining its instance-based approach, i.e. a separate model is trained for each query structure in a time efficient manner. Technically, this work presents:

- Instance-based SVM learners with regularization.
- Feature selection services, controlled by bootstrapping.

These are employed for:

- Feature selection from more than 300 freely available, non-proprietary, physico-chemical descriptors (Steinbeck et al., 2006; O'Boyle et al., 2011; Wegner, 2004) using a Random Forest approach.
- Several regression and derived classification models for predicting numeric TD_{50} values and categories for potency.

2.3. Data set

A measure of carcinogenic potency is given by TD_{50}, defined by the daily dose in mg/kg/day that causes a tumor type in 50% of the exposed animals that otherwise would not develop the tumor in a standard lifetime (Gold et al., 2001). 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 set. Moreover, Bercu et al. converted TD_{50} values to pTD_{50} for data normalization by the following equation:

$$pTD_{50} = -\log\left(\frac{TD_{50}}{1000 * \text{molecular weight}}\right)$$

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 article, such as the one from Bercu et al., contains a total of 460 training set plus 51 test set compounds for rat, and 362 training set plus 40 test set compound for mouse.

2.4. Descriptors calculation

Several descriptors were calculated by OpenTox (Hardy et al., 2010) compliant descriptor calculation services. For physico-chemical descriptors the services provided by Ideaconsult (Jeliazkova and Jeliazkov, 2011) were used. Categories were formed for the available features as follows:

Category	Number of descriptors
Constitutional	16
Electronic	33
Topological	176

A textual description follows as list below:

- constitutional: Largest Chain, Aromatic Bonds Count, Longest Aliphatic Chain, Rule Of Five, Atom Count, XLogP, ALOGP, Aromatic Atoms Count, Mannhold LogP, Bond Count, Rotatable Bonds Count, Largest Pi System.
- electronic: APol, BPol, H-Bond Acceptor Count, H-Bond Donor Count, CPSA.
- topological: Chi Path, Fragment Complexity, Kier-Hall Smarts, Kappa ShapeIndices, Petitjean Number, Autocorrelation Mass, VAdjMa, Chi Path Cluster, Wiener Numbers, Autocorrelation Polarizability, Carbon Types, Eccentric Connectivity Index, Chi Chain, MDE, Petitjean Shape Index, TPSA, Chi Cluster, Zagreb Index, Autocorrelation Charge.

The numbers in the table exceed number of items for each list category, because a descriptor name from the list may produce several individual descriptors. The table numbers give the actual number of descriptors used for modelling.

2.5. Feature selection

In order to cut down on the number of features, an approach termed *Recursive Feature Elimination* (RFE) was applied. 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 (Kuhn, 2008).

The incarnation of RFE used here employed RF that provides a valuable feature, namely a ranking of feature importance (see Section 2.2). For each *k*, the bootstrap accounts for bias in selecting features for the particular training set, so there is no need for a set-aside test set. The number of bootstrap samples was set to 50 for each *k* and differences in the ranking among samples were resolved by consensus voting.

2.6. Model training

Features were generated for training and test structures in advance and stored in a local dataset, to be re-usable in each model building process.

Model training in lazar is done separately for each prediction (instance based learning). A schematic overview is given in Fig. 1.

Upon prediction time, 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.

The neighbours are found by using cosine similarity to the query structure, and the model employs the SVM approach using a radial basis function kernel. Suitable (hyper) parameters are found via grid search (Maunz et al., 2013).

2.7. Statistical analyses

Validation runs proceeded as follows: for each training dataset, the associated test dataset was predicted using each feature type in turn. In total, two validation runs were conducted with different feature selections.

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). The exact features used in each model for characterising compounds (training compounds and the test compound) were either all features, or varied according to feature selection. Feature selection was done beforehand in a single pass for all training compounds. Test compounds were of course completely ignored during feature selection.

For each model, a prediction was made only if the neighbours exceeded a certain similarity threshold to the query. Additionally, a certain training accuracy was required on the training compounds (governed by internal bootstrapping and grid-search as described in Section 2.2). The computational effort could be kept under control through parallel processing on all available CPU cores, leading to mean training times of only 2–3 s per model.

Once a dataset prediction was finished, statistics were gathered by the OpenTox validation service (Gütlein, 2013), which also produces detailed validation reports. In total, two validation runs were conducted.

2.8. Model comparison

The quality of the models was determined through statistical parameters such as:

- Coverage. The proportion of compounds in the test set that was predicted. Variations between the models indicate differences in the AD estimation, see Section 2.9 Applicability Domain.
- Specificity, also called true negative rate. Proportion of compounds correctly predicted to be not potent relative to all compounds experimentally determined not to be potent.
- Sensitivity, also called true positive rate. Proportion of compounds correctly predicted to be potent relative to all compounds experimentally determined to be potent.
- Concordance. The proportion of compounds correctly predicted to be potent and not potent relative to total number of predictions.
- Positive predictivity. Proportion of compounds correctly predicted to be potent relative to all predictions categorized as potent.
- Negative predictivity. Proportion of compounds correctly predicted to be not potent relative to all predictions categorized as not potent.

For classification, the ROC (Receiver Operating Characteristic) score was calculated to provide an additional measure of the predictive performance of the models. On a graph sensitivity (true positive rate) was plotted versus one minus the specificity (false positive rate). Where a poor model with random predictions will yield points on the diagonal line (ROC = 1) and the best possible models, with high true positive rates and low false positive rates, yield points in the top left-corner of the plot. For classification analysis TD_{50} values were categorized as potent or not potent, or as falling in between these two categories and considered to be

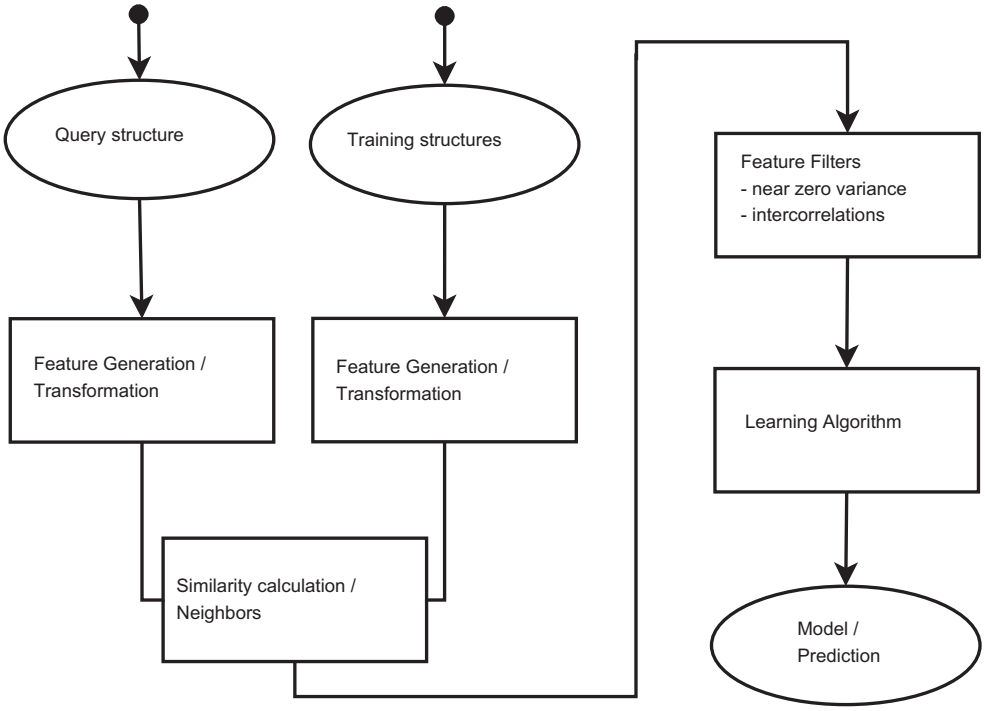

Fig. 1. Lazar model workflow.

indeterminate. As proposed by Bercu the boundaries for the classes were chosen using the TTC value for genotoxic compounds equal to 1.5 μg/day for a 70 kg person (Müller et al., 2006). It was linearly extrapolated to a TD_{50} value of 1 mg/kg/day and converted in pTD_{50} values of 4.53 (using the lowest molecular weight in the set of compounds with TD_{50} values $\leqslant$ 1 mg/kg/day). In order to clearly separate the potent and not potent categories, the not potent category was selected to start a TD_{50} level ten times higher than the potent category (10 mg/kg/day). The TD_{50} value of 10 mg/kg/day was associated with a pTD_{50} of 3.75, using the lowest molecular weight in the set of compounds with TD_{50} values $\geqslant$ 10 mg/kg/day.

Summarizing:

- The potent category contains compounds having $pTD_{50} \geqslant 4.53$ (corresponding to $TD_{50} \leqslant 1$ mg/kg/day).
- The not potent category contains compounds having $pTD_{50} \leqslant 3.75$ (corresponding to $TD_{50} \geqslant 10$ mg/kg/day).
- The indeterminate compounds contain compounds having $4.53 \geqslant pTD_{50} \geqslant 3.75$.

Since pTD_{50} cutoff was based on the lower bound for all molecular weights, many compounds that would have been categorized as non-potent using TD_{50} were labelled potent using pTD_{50}, reflecting the model conservatism.

2.9. Applicability domain

Applicability domain estimation is a core model of the lazar algorithm, and it is closely tied to the prediction algorithm, subject to the same validation procedures as predictions. Conceptually, the following factors affect the applicability domain of an individual prediction:

- Number of neighbours.
- Similarities of neighbours.
- Coherence of experimental data within neighbours.

Consequently, a prediction based on a large number of neighbours with high similarity and concordant experimental data will be more reliable than a prediction based on a low number of neighbours 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 consider 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. Moreover, our models cannot handle certain structures such as inorganic compounds, organometallics and macromolecules (e. g. polymers, proteins and DNA).

3. Results

3.1. Feature selection

The first validation run was conducted using all available descriptors. Acceptable performance was only achieved for constitutional descriptors, where mouse data were better predicted than rat ones.

In order to cut down on the number of features, especially for categories electronic and topological, an approach termed recursive feature elimination (RFE) was applied (see Section 2.5). The RFE used here employed Random Forests, providing a ranking of feature importance (see Section 2.2). The high number of bootstrapped samples gave stable results, i.e. when repeated they differed only very slightly for constitutional and electronic features.

The second run was conducted using the features found by RFE. The feature selection procedure using RFE yielded improved results, mainly for the electronic and topological descriptors, the effect for rat model being even greater than for mouse. Also, the

number of unpredicted compounds has decreased drastically for these descriptor types. These findings indicate that feature selection using RF is useful for SVM learners to cut down on the number of features.

3.2. Model comparison

Predictions for this study were calculated using constitutional descriptors, for mouse using all, and for rat using the ones selected by RFE. Classification analysis was done, such as by Bercu et al., according to two different dichotomizations (see Section 2.8):

- *hard cutoff*: compounds with experimental $pTD_{50} \geqslant 4.53$ were classified as potent, the rest as not potent.
- *without indeterminate compounds*: compounds with experimental $pTD_{50} \geqslant 4.53$ were classified as potent, and with experimental $pTD_{50} < 3.75$ as not potent. Compounds with experimental or predicted $pTD_{50} \geqslant 3.75$ and < 4.53 were disregarded (indeterminate).

In his publication, Contrera et al. gives single prediction values (but no validation statistics) for the test set predictions. Therefore it was possible to derive the statistic performance of the models from the single values. This explains the reason why we concentrate mainly on Contrera's study for comparison, because no single predictions were reported by Bercu et al. The numbers used were calculated based on the individual predictions given in the paper (predicted vs. experimental pTD_{50}). Consequently Fig. 2 compares scatterplots only for the Contrera models. A linear fit is superimposed on top of the plots (dashed) to highlight systematic deviations from the diagonal. The shaded regions indicate false positive (lower right) and false negative (upper left) predictions according to the hard cutoff value of 4.53. Clearly, according to this criterion, lazar made quite a lot of false negative predictions for rat, where for mouse it did well for false positives and false negatives. These findings seem contrary to the much reduced scatter in the lazar models, as compared to the SciQSAR ones (Contrera et al.). This example shows that the result depends very much on threshold location, but not on numerical fit of the model. A hard cutoff is also problematic due to the fact that a single value has no weight in on a continuous scale (only intervals have), lacking rationale.

In Table 1 we analysed the percentage of compounds with ratios between predicted and experimental less than or equal to 1, 2, 5 and 10-fold. For Contrera's study, we converted pTD_{50} values to TD_{50} values.

For rat, SciQSAR (Contrera model) performs better than our model (78% of compounds within predictions $\leqslant$5-fold the experimental value), while it is the other way around for mouse (86% of compounds have been predicted from our model within $\leqslant$5-fold the experimental value). Using VISDOM (Bercu model) a majority of compounds had TD_{50} predictions that were less than or equal to 5-fold the experimental value.

Table 2 gives classification results according to the hard cut-off. The numbers for Contrera were re-calculated from his numeric predictions.

Table 3 gives classification results with indeterminate compounds left out. The numbers for Contrera were re-calculated from his numeric predictions.

As reported in Tables 2 and 3, excluding indeterminate compounds improved the results and the sensitivity of each model

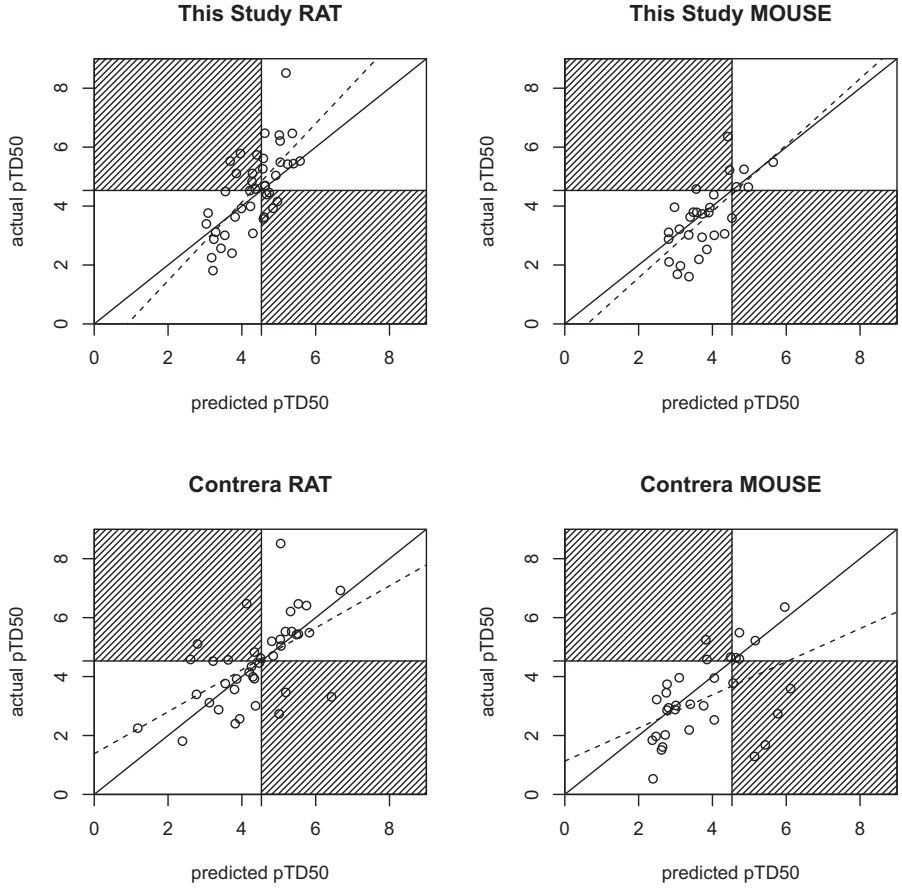

Fig. 2. Scatterplot comparison.

Table 1
Percentage of compounds with ratios between predicted and experimental less than or equal to 1, 2, 5, 10-fold.

Author (method)		⩽1-fold (%)	⩽2-fold (%)	⩽5-fold (%)	⩽10-fold (%)
Bercu et al. (VISDOM)	Rat	59	64	86	86
	Mouse	66	81	88	97
Contrera (SciQSAR)	Rat	48	65	78	85
	Mouse	56	75	78	97
This study (Lazar)	Rat	43	57	71	76
	Mouse	48	76	86	93

Table 2
Classification results (hard cutoff).

Measure	This study (Lazar)		Contrera (SciQSAR)		Bercu et al. (*Consensus*)	
	Rat	Mouse	Rat	Mouse	Rat	Mouse
Coverage	82%	73%	78%	80%	98%	98%
Specificity	71%	100%	85%	79%	40%	75%
Sensitivity	67%	57%	70%	63%	71%	53%
Concordance	69%	90%	78%	75%	62%	67%
Positive predictivity	70%	100%	82%	50%	74%	57%
Negative predictivity	68%	88%	74%	86%	38%	72%
ROC	2.33	$+\infty$	4.67	3.00	1.18	2.12

Table 3
Classification results (without indeterminate compounds).

Measure	This study (Lazar)		Contrera (SciQSAR)		Bercu et al. (*Consensus*)	
	Rat	Mouse	Rat	Mouse	Rat	Mouse
Coverage	43%	40%	37%	50%	63%	70%
Specificity	80%	100%	63%	79%	36%	85%
Sensitivity	93%	80%	82%	100%	86%	88%
Concordance	88%	94%	76%	83%	69%	86%
Positive predictivity	88%	100%	82%	56%	72%	70%
Negative predictivity	89%	92%	63%	100%	57%	94%
ROC	4.65	$+\infty$	2.2	4.76	1.34	5.87

increased. Moreover, Receiver Operating Characteristic (ROC) curve (Provost and Fawcett, 2001), as visualized by Fig. 3, was used to provide an additional measure of the predictive performance of the models and to visualise the relationship between sensitivity and false positive rate. In a ROC curve, a model on the diagonal is a poor model, having predictions no better than chance, whereas a model located in the top left corner is the ideal model, having a perfect (100%) prediction of positives and a perfect (0%) false positive rate. Typically, the ability to predict positives is made at the expense of the false positive rate.

The ROC plot in Fig. 3 summarizes the situation (hollow points: leaving out indeterminate compounds, filled points: hard cutoff, dashed lines indicate ROC levels of 2.0, 3.0 and 4.0). When leaving out indeterminate compounds, i.e. classes are separated by a distance of $4.53 - 3.75 = 0.78$, both Lazar models have ROC > 4.0 with both sensitivity and specificity $\geqslant 0.8$. This should be the case for all studies, but it is not. For example, the rat models of the other two studies perform less well (Contrera) especially (Bercu et al.) in terms of specificity. They also perform worse in terms of sensitivity. For the Contrera rat model, the specificity is even much worse than in the hard cutoff scenario. The flaw is in using the hard cutoff for classification: it hides the fact that the models perform less well in the easier scenario (where classes are separated), and thus the gross errors made for low and high pTD_{50} values. Both lazar models, on the other hand, clearly profit from the latter, as it should be the case.

Clearly, leaving out indeterminate compounds reveals the better lazar numeric predictivity in the form of very good values for all measures (the "infinity" value for ROC is due to the fact that lazar had perfect specificity, so a division by zero occurred). SciQSAR, on the other hand, even gets worse on some measures.

3.3. Model uncertainty

Fig. 4 plots lazar Applicability Domain estimation, or confidence, against RMSE (Root Mean Squared Error) (Wikipedia, 2014). Confidence here is an uncalibrated index, not a probability. It is provided for every single prediction and is defined as the median neighbour similarity. Fig. 4 is interpreted as follows: the leftmost point indicates RMSE of the very first prediction only, the second point indicates RMSE of the first two predictions, and so forth, where predictions are ranked in descending confidence order. There is a clear trend for both datasets: the more similar the neighbours, the better the predictions. Therefore, it is possible to estimate the accuracy of a given prediction based on its confidence.

4. Discussion

The results presented in this paper show the feasibility of using automated and reproducible read-across like models for the prediction of carcinogenic potency. Previously approaches to make carcinogenic potency predictions were reported in 2009 by Toropov et al., in 2010 by Bercu et al. and in 2011 by Contrera. The latter study employed a procedure similar to this study's system to obtain local models for a prediction. Given that the last two studies used exactly the same experimental dataset that we did, it was

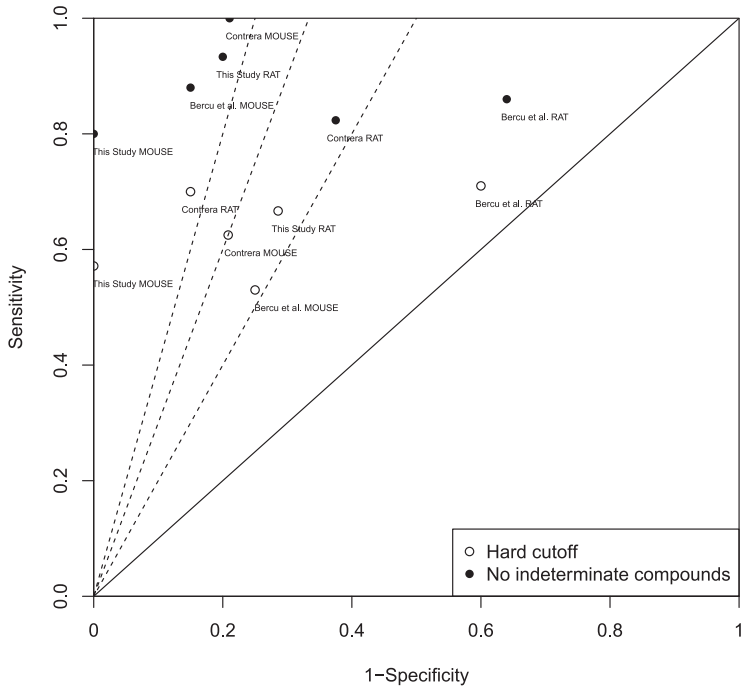

Fig. 3. ROC plots for the twelve different models.

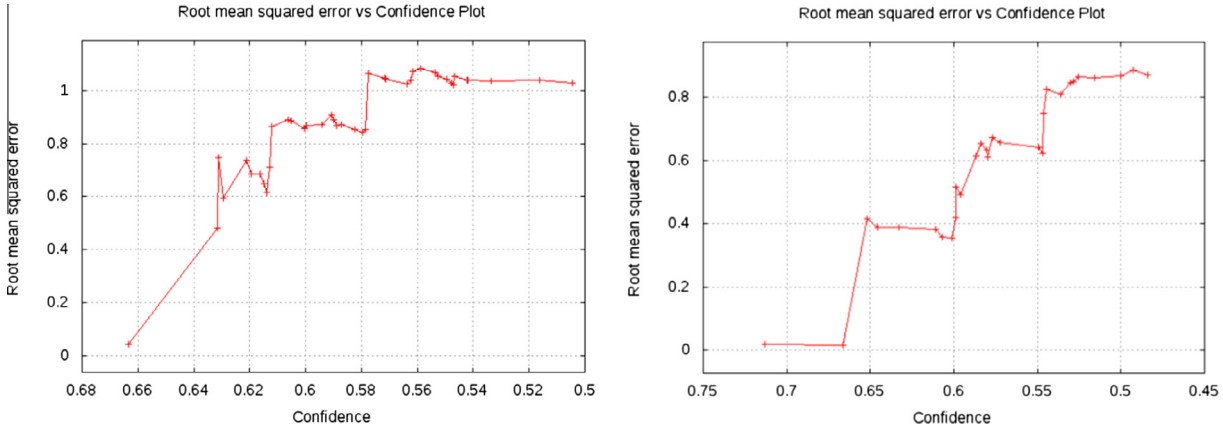

Fig. 4. Lazar applicability domain estimation: mouse (left) and rat (right).

decided to directly compare them, without any changes, neither on compounds nor on activity values. We can summarize the differences/similarities between the Contrera, Bercu et al. and our studies in few points:

- The models have been compared resulting in similar performance.
- The primary goal of both Bercu and Contera studies was to classify test compounds successfully into several potency classes, but both predicted the two test datasets numerically as well, as we did.
- Both feature a domain of applicability (AD) estimation. Similarly to this study, they omit predictions if query structures are estimated not to lie in the AD of the model.
- Methodologically this study is closer to the work of Contrera, who also used instance-based learning (a dedicated model is being learned for each prediction), and nearest neighbour selection from training data.
- All models provide predictions with errors within the same order of magnitude.

The majority of the predictions of our models are within a factor of $\leqslant$10-fold of the available experimentally-derived carcinogenic potency values. To compare this factor with experimental variability is a way to get insight into the quality of the model and its potential applicability to establish level of safety concern. Indeed, by definition prediction errors cannot be better than experimental variability. Reproducibility of carcinogenesis bioassay was examined in 70 "near-replicate" comparisons consisting of 2 of more studies applying the same experimental protocol (same route of administration, same sex and strain of rodent). For 35 comparisons of chemicals tested positive, the TD_{50} values were within a factor of 2, 5 and 10 of respectively 40, 80 and 90% of the comparisons (Gold et al., 1987). In another, similar, study approximately 95% of TD_{50}s were estimated within a factor of 4 of the mean. Between strains, about 95% of the TD_{50}s were covered by a factor of 11 of their mean (Gaylor et al., 1993). Using the same database, Gottmann et al. (2001) assessed the variability of 121 replicate rodent carcinogenicity assays from the literature part of the CPDB (Carcinogenic Potency Database) and the NCI/NTP (National Cancer Institute

and the National Toxicology Program) part of the CPDB (Gold et al., 1999), it was found that the concordance among them was 57% (that is within the concordance of our models). Taken together, these limited data converge to indicate that available experimental data are rather variable. Interestingly, the errors observed in our models (and others) are within the same order of magnitude than the experimental variability. This suggests that for the models, an important source of limitations is the quality and scarceness of the experimental data. Similar conclusion was drawn for models predicting rat chronic toxicity (Mazzatorta et al., 2008).

In contrast to read-across conducted by experts, which is time consuming and subjective in terms of analogues selection, our read-across models are automated and reproducible. Similar compounds are chosen automatically considering not only chemical structure but also the features (descriptors) relevant for the toxic endpoint studied. This provides the possibility to check the strength and plausibility of predictions.

To provide transparency and to allow external use and assessment, our models will be made freely available online by a user friendly interface that will convert back pTD_{50} to TD_{50}. Detailed supporting information will be provided, such as the analysis of similar compounds used and the prediction confidence.

The models proposed require a range of similar chemicals in the training set and if the models are used outside applicability domain, the reliability of the prediction will be lower. Moreover as it is encouraged by REACH, the best practise is to use more than one model and more than one platform, whenever possible, and the comparison of the results obtained will strengthen the confidence of the prediction (Schilter et al., 2014). It's important to keep in mind that generally, like in any scientific field, data interpretation requires knowledge and expertise and using *in silico* models is not an exception. *Lazar* derives computational models from objective, traceable and reproducible statistical criteria. Therefore the predictions obtained are statistically derived and the toxicological expert is a key part of the process. Toxicologists should always review and interpret the output and its belonging to the applicability domain, in order to identify errors, chance correlations and results that contradict with current knowledge and discard results if necessary.

5. Conclusions

Increasing pressure to reduce or eliminate animal testing, together with the need of fast decision making for management of emergency safety issues, brought us to think about approaching toxicity predictions in a different way. Instead of considering each new chemical as an unknown entity, the toxicity of the chemical under investigation can be directly inferred deriving all information available from similar compounds whose activities are known. For this purpose we developed automated and reproducible read-across models for the quantitative prediction of carcinogenic potency providing an automatic process for analogues selection and prediction, as well as enabling interpretation of the results obtained thought visual inspection of the similar compounds. Indeed our models don't involve any subjective choices in terms of analogue selection, that is very often very time consuming, but the similar compounds are chosen automatically, considering not only the similarity based on chemical structure but also the features relevant for the toxic endpoint studied. The models have been validated and they provide predictions with errors within the same order of magnitude than the estimated variability of experimental data. Moreover through the user friendly platform, soon freely available online, the analysis and visualization of the similar compounds selected, together with the prediction confidence, will enable the toxicologists to review the results obtained.

Conflict of interest

All authors declared that no conflict of interest exists and signed a conflict of interest policy form.

References

Arvidson, K.B., Chanderbhan, R., Muldoon-Jacobs, K., Mayer, J., Ogungbesan, A., 2010. Regulatory use of computational toxicology tools and databases at the United States Food and Drug Administration's Office of Food Additive Safety. Expert Opin. Drug Metab. Toxicol. 6 (7), 793–796.

Benfenati, E., Benigni, R., Marini, D., Helma, C., Kirkland, D., Martin, T.M., Mazzatorta, P., Meunier, J.-R., Ouédraogo-Arras, G., Richard, A., Schilter, B., Schoonen, W.G.E.J., Snyder, R., Yang, C., Youne, D.M., 2009. Predictive models for carcinogenicity and mutagenicity: frameworks, state-of-the art and perspectives. J. Environ. Sci. Health C Carcinog. Ecotoxicol. Rev. 27, 57–90.

Bercu, J.P., Morton, S.M., Deahl, J.T., Gombar, V.K., Callis, C.M., van Lier, R.B., 2010. In silico approaches to predicting cancer potency for risk assessment of genotoxic impurities in drug substances. Regul. Toxicol. Pharmacol. 57 (2), 300–306.

Breiman, L., 2001. Random forests. Machine Learning.

Contrera, J.F., 2011. Improved in silico prediction of carcinogenic potency (TD_{50}) and the risk specific dose (RSD) adjusted threshold of toxicological concern (TTC) for genotoxic chemicals and pharmaceutical impurities. Regul. Toxicol. Pharmacol. 59 (1), 133–141.

ECHA, 2011. The Use of Alternatives to Testing on Animals for the REACH Regulation. Available at: http://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2011_en.pdf.

EFSA, 2010. Applicability of QSAR Analysis to the Evaluation of the Toxicological Relevance of Metabolites and Degradates of Pesticides Active Substances for Dietary risk assessment. Available at: http://www.efsa.europa.eu/it/supporting/doc/50e.pdf.

Gaylor, D.W., Chen, J.J., Sheehan, D.M., 1993. Uncertainty in cancer risk estimates. Risk Anal. 13, 149–154.

Gold, L.S., Wright, C., Bernstein, L., de Veciana, M., 1987. Reproducibility of results in "near-replicate" carcinogenesis bioassays. J. Natl. Cancer Instit. 78, 1149–1158.

Gold, L.S., Manley, N.B., Slone, T.H., Rohrbach, L., 1999. Supplement to the carcinogenic potency database (CPDB): results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996. Environ. Health Perspect. 107 (4), 527–600.

Gold, L.S., Slone, T.H., Ames, B.N., Manley, N.B., 2001. Pesticide residues in food and cancer risk: a critical analysis. In: Krieger, R. (Ed.), Handbook of Pesticide Toxicology, second ed. Academic Press, San Diego, CA, pp. 799–843.

Gottmann, E., Kramer, S., Pfahringer, B., Helma, C., 2001. Data quality in predictive toxicology: reproducibility of rodent carcinogenicity experiments. Environ. Health Perspect. 109 (5), 509–514.

Gütlein, M., 2013. A large-scale empirical evaluation of cross-validation and external test set validation in (Q)SAR. Mol. Informatics 32 (5–6), 516–528.

Hardy, B., Douglas, N., Helma, C., Rautenberg, M., Jeliazkova, N., Jeliazkov, V., Nikolova, I., Benigni, R., Tcheremenskaia, O., Kramer, S., Girschick, T., Buchwald, F., Wicker, J., Karwath, A., Guetlein, M., Maunz, A., Sarimveis, H., Melagraki, G., Afantitis, A., Sopasakis, P., 2010. Collaborative development of predictive toxicology applications. J. Cheminformatics 2, 7–8.

ICCR, 2012. Applicability of Animal Testing Alternatives. Available at: http://www.fda.gov/downloads/Cosmetics/InternationalActivities/ConferencesMeetingsWorkshops/InternationalCooperationonCosmeticsRegulationsICCR/UCM320464.pdf.

Jeliazkova, N., Jeliazkov, V., 2011. AMBIT RESTful web services: an implementation of the OpenTox application programming interface. J. Cheminformatics 3, 18.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 1–26.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Lo Piparo, E., Worth, A., Manibusan, M., Yang, C., Schilter, B., Mazzatorta, P., Jacobs, M.N., Steinkellner, H., Mohimont, L., 2011. Use of computational tools in the field of food safety. Regul. Toxicol. Pharmacol. 60 (3), 354–362.

Maunz, A., Helma, C., 2008. Prediction of chemical toxicity with local support vector regression and activity-specific kernels. SAR QSAR Environ. Res. 19, 413–431.

Maunz, A., Gütlein, M., Rautenberg, M., Vorgrimmler, D., Gebele, D., Helma, C., 2013. Lazar: a modular predictive toxicology framework. Front. Pharmacol. 4, 38.

Mazzatorta, P., Estevez, M.D., Coulet, M., Schilter, B., 2008. Modeling oral rat chronic toxicity. J. Chem. Inf. Model. 48 (10), 1949–1954.

Müller, L., Mauthe, R.J., Riley, C.M., Andino, M.M., Antonis, D.D., Beels, C., DeGeorge, J., De Knaep, A.G., Ellison, D., Fagerland, J.A., Frank, R., Fritschel, B., Galloway, S., Harpur, E., Humfrey, C.D., Jacks, A.S., Jagota, N., Mackinnon, J., Mohan, G., Ness, D.K., O'Donovan, M.R., Smith, M.D., Vudathala, G., Yotti, L., 2006. A rationale for determining, testing, and controlling specific impurities in pharmaceuticals that possess potential for genotoxicity. Regul. Toxicol. Pharmacol. 44 (3), 198–211.

NAS, 2007. Toxicity in the 21st Century: A Vision and a Strategy. Available at: http://dels.nas.edu/resources/static-assets/materials-based-on-reports/reports-in-brief/Toxicity_Testing_final.pdf.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: an open chemical toolbox. J. Cheminformatics, 33.

Provost, F., Fawcett, T., 2001. Robust classification of imprecise environment. Mach. Learn. 42, 203–231.

Rusyn, I., Daston, G., 2010. Computational toxicology: realizing the promise of the toxicity testing in the 21st century. Environ. Health Perspect. 118 (8), 1047–1050.

Schilter, B., Benigni, R., Boobis, A., Chiodini, A., Cockburne, A., Cronin, M.T.D., Lo Piparo, E., Modi, S., Thielh, A., Worth, A., 2014. Establishing the level of safety concern for chemicals in food without the need for toxicity testing. Regul. Toxicol. Pharmacol. 68 (2), 275–296.

Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., Willighagen, E., 2006. Recent developments of the chemistry development kit (CDK) – an open-source java library for chemo- and bioinformatics. Curr. Pharm. Des. 12, 2111–2120.

Toropov, A.A., Toropova, A.P., Benfenati, E., 2009. Additive SMILES-based carcinogenicity models: probabilistic principles in the search for robust predictions. Int. J. Mol. Sci. 10 (7), 3106–3127.

USFDA, 2008. Guidance for Industry: Genotoxic and Carcinogenic Impurities in Drug Substances and Products: Recommended Approaches. Available at: http://www.fda.gov/downloads/Drugs/.../Guidances/ucm079235.pdf.

U.S. EPA, 2008. Science and decision: advancing risk assessment. Free Executive Summary. Available at: https://www.law.upenn.edu/institutes/regulation/papers/BurkeScienceAndDecisions.pdf.

U.S. EPA, 2012, Science and decision: advancing risk assessment. Committee on Improving Risk Analysis Approaches. Available at: http://www.epa.gov/region9/science/seminars/2012/advancing-risk-assessment.pdf.

Vapnik, C., Cortes, C., 1995. Support-vector networks. Machine Learning.

Wegner, J.K., 2004. JOELib – an open source chemoinformatics library for data mining and graph mining on molecular structures. Invited Presentation at eCheminfo 2004, 8–19 November, Applications of Cheminformatics and Modelling to Drug Discovery.

Wikipedia, 2014. Root-Mean-Square Deviation. Available at: https://en.wikipedia.org/wiki/Root-mean-square_deviation (accessed 15.03.14).