A *k*-NN Algorithm for Predicting Oral Sub-Chronic Toxicity in the Rat

Domenico Gadaleta [1,2], Fabiola Pizzo [1], Anna Lombardo [1], Angelo Carotti [2], Sylvia E. Escher [3], Orazio Nicolotti [2] and Emilio Benfenati [1]

[1]Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy; [2]Dipartimento di Farmacia – Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Bari, Italy; [3]Fraunhofer Institute for Toxicology and Experimental Medicine (ITEM), Hannover, Germany

Summary

Repeated dose toxicity is of the utmost importance to characterize the toxicological profile of a chemical after repeated administration. Its evaluation refers to the Lowest-Observed-(Adverse)-Effect-Level (LO(A)EL) explicitly requested in several regulatory contexts, such as REACH and EC Regulation 1223/2009 on cosmetic products. So far in vivo tests have been the sole viable option to assess repeated dose toxicity. We report a customized k-nearest neighbors (k-NN) approach for predicting sub-chronic oral toxicity in rats. A training set of 254 chemicals was used to derive models whose robustness was challenged through leave-one-out cross-validation. Their predictive power was evaluated on an external dataset comprising 179 chemicals. Despite the intrinsically heterogeneous nature of the data, our models give promising results, with $q^2 \geq 0.632$ and external $r^2 \geq 0.543$. The confidence in prediction was ensured by implementing restrictive user-adjustable rules, excluding suspicious chemicals irrespective of the goodness in their prediction. Comparison with the very few LO(A)EL predictive models in the literature indicates that the results of the present analysis can be valuable in prioritizing the safety assessment of chemicals and thus making safe decisions and justifying waiving animal tests according to current regulations concerning chemical safety.

Keywords: k-nearest neighbor, repeated dose toxicity, lowest observed (adverse) effect level, REACH, QSAR

1 Introduction

There is growing concern about the safety assessment of chemicals and their effects on human health and the environment. In parallel, there is great attention towards alternative methods (i.e., *in vitro* and *in silico* approaches) for investigating potentially harmful substances in place of *in vivo* experiments, which are time-consuming, expensive and ethically questionable (Kroes et al., 2007).

Several regulations, such as Registration, Evaluation, Authorization and restriction of Chemicals (REACH) (European Commission, 2006) and the EC Regulation on cosmetic products 1223/2009 (European Commission, 2009), have been issued by the European Union (EU) in the last few years. Regulators explicitly require the evaluation of repeated dose toxicity (RDT), as it provides valuable information about adverse effects that may arise upon repeated exposure to a certain substance over a limited period of time (SCCS, 2012; Sakuratani et al., 2013). RDT studies give the no observed (adverse) effect level (NO(A)EL) and the lowest observed (adverse) effect level (LO(A)EL). The former is the experimental dose at which there is no significant response (Sand et al., 2008); the latter indicates the lowest dosage at which adverse effects arise compared to a control group (e.g., onset of an adverse effect) (Sakuratani et al., 2013).

The NO(A)EL and LO(A)EL are usually based on *in vivo* studies following diverse protocols that differ in the exposure period, the animal model (rodent or non-rodent species) and the exposure route (oral, inhalation or dermal) (SCCS, 2012). The present study focused on RDT data for sub-chronic (90 days) oral exposure in rats because this kind of data is easily accessible and oral exposure is the standard route of administration for assessing systemic toxicity.

The NO(A)EL serves to determine both the acceptable daily intake (ADI) by the application of uncertainty factors (World Health Organization, 1999; Sand et al., 2008) and the reference dose (RfD). The safety of a substance is thus estimated by comparing these values with the human safety threshold (Kalberlah et al., 2003). As regards the evaluation of cosmetics toxicity, NO(A)EL is useful to calculate the margin of safety (MoS), indicating the potential toxicity for human health of substances contained in a cosmetic formulation (SCCS, 2012).

Received May 9, 2014; accepted in revised form July 3, 2014; Epub July 10, 2014; http://dx.doi.org/10.14573/altex.1405091

REACH too requires information on RDT in terms of NO(A)EL and LO(A)EL for substances manufactured or imported in quantities ranging from 10 to 1,000 tons per year (annexes VIII and IX) (Lilienblum et al., 2008). As regards the Chemical Safety Assessment (CSA), the NO(A)EL is necessary to obtain DNELs (derived no effect levels) (ECHA, 2008). Where a NO(A)EL is missing, LO(A)EL can be used, applying uncertainty factors (SCCS guideline, 2012; ECHA, 2008; European Commission, 2006; Setzer and Kimmel, 2003).

NO(A)EL and LO(A)EL are strongly dependent on the number of experimental doses and the intervals between doses (Filipsson et al., 2003). Neither gives an accurate assessment of the real dose at which an effect occurs (LO(A)EL) or does not occur (NO(A)EL). They are not calculated on the basis of dose-response curves, but can be any level ranging from the lowest to the highest dose tested (Tluczkiewicz et al., 2013). As a result, NO(A)EL and LO(A)EL approaches have been harshly criticized and the benchmark dose (BMD) evaluation has been proposed to assess RDT better in place of the traditional approaches (Setzer and Kimmel, 2003; Sand et al., 2008). Although it is well known, the BMD based approach has not fully replaced those based on NO(A)EL and LO(A)EL (Kodell, 2009).

REACH strongly encourages the use of alternative methods, such as quantitative structure-activity relationship (QSAR) models and, more generally, *in silico* strategies (Gissi et al., 2014). Since March 2013, the use of animals for toxicity testing of cosmetic products has been banned (Pauwels and Rogiers, 2010). The use of toxicological evidence from QSARs can reduce or replace *in vivo* assays according to the 3Rs principle (Replace, Reduce, Refine) (Russell and Burch, 1959). Despite all efforts in this direction, reliable alternative methods still do not exist (SCCS, 2012; Adler et al., 2011). Therefore, there is a need to develop reliable and scientifically valid methods that can properly predict toxicological data.

To date, there are no models for the NO(A)EL assessment, though some attempt has been made to model LO(A)EL data. For instance, a module for modeling LO(A)EL has been implemented in the commercial software TOPKAT (http://accelrys.com/mini/toxicology/predictive-functionality.html). De Julián-Ortiz et al. (2005) and García-Domenech et al. (2006) modeled chronic LOEL data by means of different algorithms based on multiple linear regression (MLR) and linear discriminant analysis (LDA). Mazzatorta et al. (2008) developed a QSAR model based on LO(A)EL data related to chronic oral toxicity in rats. More recently, Sakuratani et al. (2013) proposed, within the Hazard Evaluation Support System (HESS), a model for predicting RDT using toxicological categories based on LO(A)EL data.

The aim of this work is to provide a new tool for predicting LO(A)EL using a simpler approach, based on the *k*-nearest neighbors (*k*-NN) algorithm that can provide toxicological information on the substances and at the same time supports the use of alternative methods. Similarity-based approaches, such as *k*-NN algorithms, have already proven effective in modeling a number of complex endpoints (Cassotti et al., 2014; Raevsky et al., 2011; Stoyanova-Slavova et al., 2014). This present study proposes a valuable *in silico* method as an option to back up animal models, provided the conditions we identified are met.

2 Materials and methods

2.1 Experimental data

Three sources were taken into account to build the training set (TS) of the *k*-NN model.

1. The Munro database (2006) was downloaded from OECD QSAR Toolbox (version 3.1) (freely available at http://www.qsartoolbox.org/). It includes LO(A)EL and NO(A)EL data related to 613 organic chemicals used in industrial, environmental, consumer and food substances likely to be encountered in commerce.
2. HESS (Hazard Evaluation Support System) database, taken from the OECD QSAR Toolbox. It contains information related to RDT (LO(A)EL and NO(A)EL) for 502 industrial chemicals.
3. EPA's Integrated Risk Information System (IRIS) database, available at http://cfpub.epa.gov/ncea/iris/index.cfm?fuseaction=iris.showSubstanceList, includes LO(A)EL and NO(A)EL related to 557 substances.

The external validation set (VS) was taken from the RepDose database, which includes 761 LO(A)EL and NO(A)EL studies on 438 different substances. The data can be consulted at http://www.fraunhofer-repdose.de/. Only fully reliable data or data with minor deficiencies was considered in the construction of the VS.

In order to have consistent data, only values gained in subchronic toxicity studies (from 84 to 98 days) of oral exposure (gavage, diet, or drinking water) were taken into account. Since interspecies differences between rats and mice in RTD data are reported in the literature (Escher et al., 2013), we selected only data related to studies on rats (*Rattus norvegicus*). LO(A)EL were for male and female rats. However, we did not consider data related to reproductive effects in females. Inorganic compounds, isomeric mixtures, metal complexes and data related to mixtures of chemicals were rejected. Ionized structures were neutralized and counterions eliminated. The numerical values were converted to a logarithmic scale.

Canonical SMILES for each retained chemical were retrieved from public online databases (ChemIDplus (http://chem.sis.nlm.nih.gov/chemidplus/), PubChem (https://pubchem.ncbi.nlm.nih.gov/)). Duplicates were detected through a multiple check of canonical SMILES and CAS number of each compound. In case of multiple data per compound, the lowest value was retained as a precautionary approach. Finally, the data were checked and combined to give a TS and a VS of 254 chemicals (138 from HESS, 99 from Munro and 17 from U.S. EPA's IRIS) and 179 chemicals, respectively.

2.2 *k*-Nearest neighbors

A *k*-NN algorithm was applied to obtain LO(A)EL predictive models. This algorithm is the simplest among those used in machine learning and can estimate the outcome (i.e., a continuous or categorical value) of a query point (i.e., a target compound)

on the basis of read-across accounting for its most similar objects (i.e., nearest neighbors) among a set of examples for which the outcomes are known (i.e., chemicals within the model's TS) (Altman, 1992).

We obtained similarity values between a target compound and nearest neighbors using in-house software (istSimilarity developmental version). The software computes an integrated similarity index (SI), ranging from 1 (maximum similarity) to 0 (minimum similarity), resulting from a weighted combination of a binary fingerprint array and three non-binary structural keys (Durant et al., 2002) based on molecular descriptors. The fingerprints are the extended fingerprints, which comprise Daylight notation (http://www.daylight.com/dayhtml/doc/theory/theory.finger.html) and additional bits accounting for ring features. The structural keys are based on a) 35 constitutional descriptors; 2) 11 counters for different hetero-atoms; 3) 154 functional group counts (as defined by the software Dragon) (Talete srl: Milano, Italy).

The software istSimilarity is freely available upon request at the VEGA website (http://www.vega-qsar.eu). On the basis of this index, *k*-NN models can predict LO(A)EL for a target compound as the average (arithmetical mean) of the experimental values (outcomes) of the *k*-nearest neighbors.

The robustness of each model was evaluated using leave-one-out (LOO) cross-validation. The LO(A)EL of each chemical in the TS was predicted on the basis of the *k*-nearest neighbors among those remaining in the TS. Then we calculated the cross-validated coefficient of determination (q^2) and the root mean square error (RMSE). The real prediction power of each model was evaluated on the VS by reporting the coefficient of determination (r^2) and RMSE values:

$$q^2 or\ r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y_{avg})^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

where y_i is the experimental value of the *i*-th chemicals in the dataset; $\hat{y}_i$ is the predicted value of the *i*-th query compound in the VS for the determination of r^2, or the predicted response resulting from LOO cross-validation for the *i*-th neighbor in the TS for the determination of q^2; y_{avg} is the mean of the experimental values of the compounds in the dataset; and *n* is the number of compounds (Nicolotti and Carotti, 2006).

2.3 Refined algorithm

The algorithm was refined by setting additional conditions, and a target chemical must fulfill all those rules to be considered reliably predicted. Nearest neighbors among the *k* selected for the prediction of the query compound's LO(A)EL should have a SI ≥ 0.85 (Willett, 1998) otherwise they do not participate in the prediction stage. If there are no neighbors (i.e., chemicals in the TS) matching at least this similarity threshold, the model does not provide a prediction value for the target compound. If two or more neighbors fulfill this condition, the difference between the maximum and minimum experimental values among retained neighbors is considered. If this difference is < 1 log unit (all the neighbors have similar LO(A)ELs), the target is predicted as the

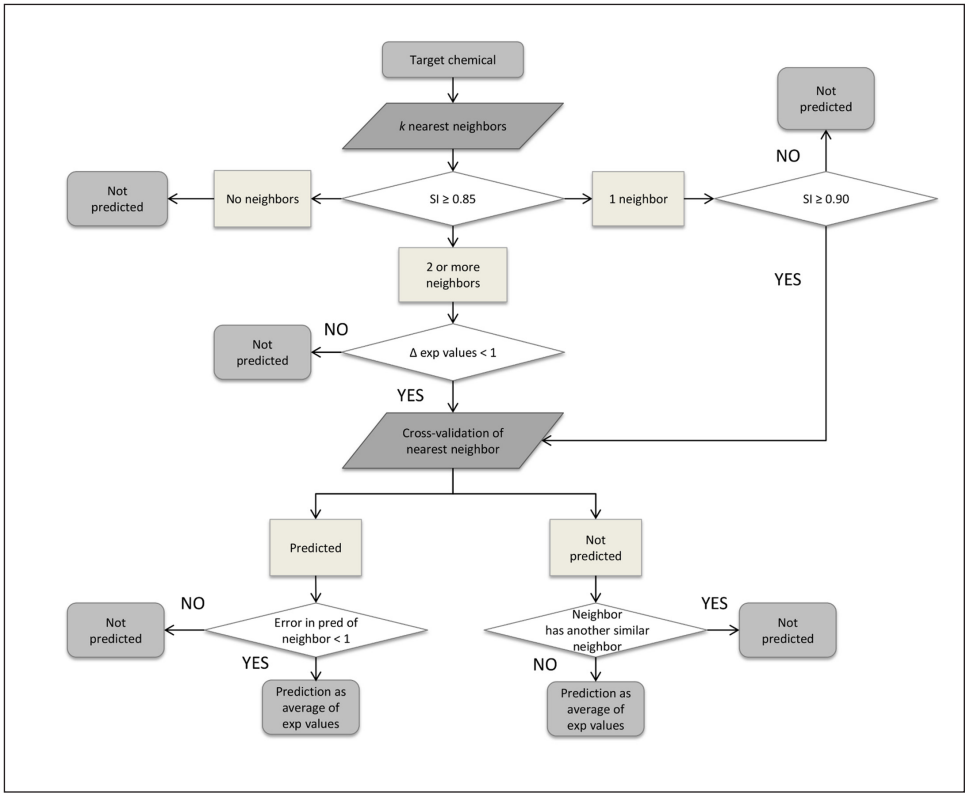

Fig. 1: Flowchart for the selection of the output predictions
SI is similarity index between the target chemical and its nearest neighbors; Δ exp values are the differences between experimental values of nearest neighbors; error in pred is the error in prediction returned in cross-validation of a neighbor in the TS.

average of the neighbors' values, otherwise the model does not return any prediction.

If the prediction of the target is based on a single neighbor, the SI must be ≥ 0.90 for obtaining a prediction (which is equal in this case to the experimental values of the neighbor). In addition, the algorithm verifies how the target's nearest neighbor is predicted in LOO cross-validation. Reasonably, if the model fails in predicting the neighbor, we cannot consider a prediction for the query reliable. If the error in internal prediction (resulting from LOO cross-validation) of the query's nearest neighbor is 1 log unit or more, the query's prediction is rejected.

It may happen that the neighbor is not predicted in LOO cross-validation because it does not fulfill one of the above conditions. If it is not predicted because it lacks other nearest neighbors (similarity index ≥ 0.85) for the read-across, the query's prediction is retained. Otherwise, if the neighbor has at least one other similar compound within the training set, but it is not predicted because it does not fulfill other conditions (e.g., differences among experimental values for its similar compounds), the query prediction is rejected. The flow chart of the algorithm applied to derive LO(A)EL models is shown in Figure 1. All statistical analyses were done using Microsoft Excel 2010 (Microsoft Corp., Redmond, WA).

3 Results

The basic *k*-NN models give poor results, with values of q^2 resulting from LOO cross-validation between 0.057 and 0.112. However, the application of the additional conditions strongly improves model performance, at the cost of refusing a number of predictions. Table 1 summarizes the performance of the refined *k*-NN models. The internal predictivity of each model was confirmed by $q^2 \geq 0.632$ and RMSE ≤ 0.478. The results in LOO cross-validation were further confirmed in external validation. All models returned $r^2 \geq 0.543$ and RMSE ≤ 0.659.

As shown in Table 1, the good performance results at the cost of a low prediction rate. Among the applied conditions, the criterion that results most often by far in refusing prediction is the first similarity threshold of 0.85 applied in the selection of nearest neighbors: it causes the refusal of 123 predictions for the TS (48% of total) for every model. Table S1 in the supplemental file at http://dx.doi.org/10.14573/altex.1405091s shows the percentage of refused predictions for the TS after the application of each further parameter.

We assumed as outliers those compounds with an error in prediction ≥ 1 log unit. As shown in Figure 2, only few chemicals have such a large error, for each model in cross-validation (from 2 to 0 outlier for the TS) and in external validation (from 2 to 3 outliers, about 10% of predicted compounds, for the VS). No chemicals returned an error in prediction ≥ 2 log units for all the models, considering both TS and VS.

Table S2 in the supplemental file at http://dx.doi.org/10.14573/altex.1405091s shows that even better statistics for both internal and external validation can be obtained by applying more restrictive rules in the algorithm. However, this happens at the cost of a further decrease in the percentage of predicted compounds, from both the TS and VS. Lowering the maximum accepted difference among the experimental values for similar substances from an initial value of 1 log unit to a more restrictive margin of 0.75 log units involves an increase of $q^2 \geq 0.741$

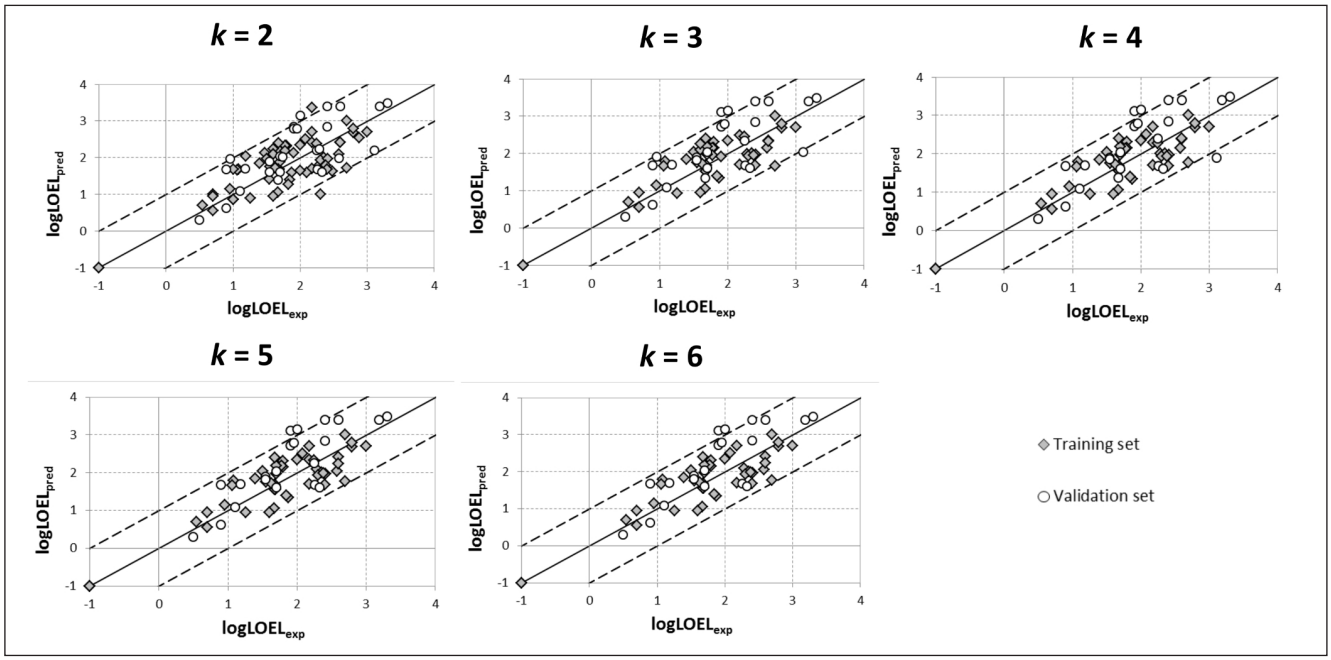

Fig. 2: Experimental and predicted logLO(A)EL for the five *k*-NN models

TS and VS chemicals are represented by grey diamonds and white circles. The continuous lines indicate the case of ideal correlation. The dashed lines delimit the chemicals with an error in prediction lower than 1 log unit.

Tab. 1: Performance of *k*-NN models
Statistics and predicted compounds are reported for cross-validation on the TS (containing 254 compounds) and external validation on the VS (containing 179 compounds).

Training set (TS)					
k	2	3	4	5	6
q^2	0.632	0.747	0.769	0.764	0.762
RMSE	0.478	0.405	0.396	0.400	0.407
Predicted compounds	68	51	49	49	47
Prediction rate	27%	20%	19%	19%	19%
Validation set (VS)					
k	2	3	4	5	6
r^2	0.543	0.552	0.554	0.675	0.682
RMSE	0.612	0.659	0.654	0.627	0.642
Predicted compounds	27	24	23	21	20
Prediction rate	15%	13%	13%	12%	11%

for all the models. The result is best for $k = 5$ or $k = 6$ with $q^2 = 0.831$ on the TS. The percentage of compounds that comply with this more restrictive margin ranges from 19% to 13% of the TS. In the VS, $r^2 \geq 0.564$ reaches a maximum of 0.707 for k ranging from 4 to 6 on 8% of the data set. An even more restrictive margin (i.e., 0.5 log unit) further increases the performance. In this case, the results are best for $k = 4$, with $q^2 = 0.859$ on 11% of the TS, and $r^2 = 0.779$ on the 7% of the VS.

Application of more restrictive rules in the algorithm greatly improves performance, although at the cost of a substantial drop in the number of predicted compounds. We thus suggest a maximum difference among neighbors' experimental values of 1.0 log unit for a good compromise between the statistical performance and the number of predicted compounds.

4 Discussion

4.1 Endpoint complexity

To the best of our knowledge, this study is the first attempt to develop a predictive tool for modeling sub-chronic toxicity (90 days) in rats, an endpoint that is explicitly required for the risk assessment of chemicals in several regulations (SCCS, 2012; European Commission, 2006, Annex IX). Based on a straightforward *k*-NN algorithm, our approach proved effective in modeling LO(A)EL. In addition, a series of user-adjustable rules in the selection of the best nearest neighbors allowed even higher trustworthy predictive values. The gain in prediction and confidence was obtained for a given percentage of the dataset (see Tab. 1 and S2), while a number of compounds were left unpredicted as a precautionary measure. The use of restrictive conditions in modeling such a complex endpoint, based on heterogeneous data in terms of quality, experimental sources and organisms used (Tsakovska et al., 2008), meets both the scientific and regulatory purposes established by international bodies for the protection of human health. Providing few but highly reliable predictions constitutes a valuable attempt to better prioritize chemicals and thus reduce the number of animals needed for *in vivo* testing.

Moreover, predictions that meet the more restrictive parameters set in the algorithm (i.e., maximum difference among experimental values of similar compounds 0.75 or 0.50, see Section 3) can be considered more reliable than those that meet only the basic parameters. This may serve as an alert to tune the confidence in the predictions.

Unlike other work focusing on QSAR models for the prediction of LO(A)EL (Mazzatorta, 2008; Sakuratani, 2013), the choice of a simpler algorithm, such as *k*-NN, avoids some pitfalls. One is related to the wide variety of toxic effects against different organs and tissues, as well as the different mechanisms of action that determine LO(A)EL. In a QSAR model these mechanisms are parameterized in specific chemical attributes (descriptors) describing structural, biological and physico-chemical properties of molecules. QSAR models try to create a correlation (in the form of a mathematical equation) between an optimal set of descriptors and toxicity, or other relevant endpoints. For endpoints that refer to a single mechanism of action it is more or less simple to find a number of properties (e.g., log P for bioaccumulation potential or aquatic toxicity) or some structural motives (e.g., for mutagenicity) that result effectively in the explanation and in the prediction of a wide range of compounds. On the other hand a description of a complex endpoint such as RTD by only a small set of descriptors alone is insufficient. The attributes for single-effect specific QSARs may differ from those derived from combined toxicity data, leading to contradictory, hence unreliable, toxicity predictions (Venkatapathy et al., 2004). It is therefore clearly difficult to build up a mechanistically transparent structure-activity model for an endpoint accounting for whole-body assessment (Sakuratani et al., 2013; Tilaoui et al., 2007). On the other hand, read-across methods, such as the *k*-NN algorithm presented herein, predict toxicity by simply comparing the target compound to a restricted pool of structurally similar molecules, which plausibly generate toxicity with similar mechanisms. In this case descriptors (i.e., fingerprints) are applied only to find such similar molecules, not to explain the endpoint. In this respect, a read-across approach seems more adequate and reliable for this kind of endpoint than a QSAR model.

Our strategy is not a mere application of the traditional *k*-NN methodology, which gave poor results. Better results are obtained thanks to the information on the small set of similar compounds. If the similar substances have dissimilar experimental values, we can expect greater uncertainty in the prediction of the value for the target compound, and this was in fact shown. In this way, we shifted from a blind, unsupervised *k*-NN to a supervised approach. It is worth noting that the definition of each condition and the refinement of thresholds was performed to obtain a good prediction on the TS. However, the risk of overfitting is avoided by the good performance obtained in the external validation, giving confidence that models are not biased, but actually predictive.

Another limitation in the LO(A)EL modeling is that available data are often obsolete and their quality is closely related to differences in the protocols used and the purity of the tested chemicals

(Mazzatorta et al., 2008). An unprecedented advantage of the algorithm presented here is that the models can be easily improved by updating the TS with new, more accurate toxicity data, without any need for model retraining. Our attempt offers an important step forward to obtain more useful and effective tools for the prediction of the toxicity of chemicals (Venkatapathy et al., 2004).

4.2 Mechanistic interpretation of the RDT

The inspection of predicted chemicals disclosed the frequent occurrence of some structural chemotypes. This information can constitute a valuable basis to shed light on the causative relationships linking the occurrence of a given structural motif and the observation of a toxic effect.

In this respect, we observed that halogenated aromatic chemicals are often well predicted by our models. Generally, such compounds are predicted as highly toxic compounds. In particular, considering the model with $k = 2$, 11 out of 15 halogenated aromatic compounds included in the TS have a predicted log LO(A)EL < 2, and 10 out of 15 have a predicted log LO(A)EL ≤ 1. As shown in Figure 3A, the toxicity is likely due to the metabolic action of CYP450 enzymes, which transform such compounds into highly reactive and electrophilic species (i.e., epoxides) exposed to the nucleophilic attack of tissue proteins for covalent binding. Moreover, the spontaneous conversion of the epoxide to phenol and then the secondary oxidation of phenols by CYP450 enzymes lead to the formation of hydroquinones, which can be subsequently oxidized to quinones. Again, quinones are electrophilic species and can also bind tissue proteins or lead to the generation of reactive oxygen species (Sakuratani et al., 2013; Chan et al., 2007). The only halogenated aromatic compounds with a predicted log LO(A)EL > 2 are: pentachloranisole (CAS number: 1825-21-4), because the hydrolysis of its methoxy group leads to the formation of a phenol derivative that could be easily conjugated and eliminated without generating more toxic metabolites; o-chlorotoluene (CAS number: 95-49-8), because the steric and electronic effects of the methyl group hamper the oxidative process of the aromatic ring; chlorobenzene (CAS number: 108-90-7) and 1,2-DCB (1,2-dichlorobenzene, CAS number: 95-50-1).

Actually, the latter two chemicals are under-predicted by the model. Indeed both chemicals have among their nearest neighbors 1,4-DCB. It was observed that the reactivity of halobenzenes is strongly related to their dipole moment (Chan et al., 2007). The presence of a permanent dipole moment in 1,2- DCB is likely responsible for its higher reactivity and higher toxicity. Conversely, the higher molecular symmetry of 1,4-DCB switches off the dipole moment and also reduces its toxicity (log LOEL = 2.477). It is thus expected that symmetrical halobenzenes are poorly metabolized by CYP450 and therefore are not readily metabolized to their active metabolites compared to asymmetrical halobenzenes (Chan et al., 2007).

Other structural moieties often recurring in predicted molecules are nitrobenzene and aniline rings. As shown in Figure 3B, the metabolism of those compounds is strictly related. The biotransformation of nitrobenzene involves both oxidation and reduction reactions. Oxidation products of nitrobenzene include nitrophenols, while reduction products include nitrosobenzene, phenylhydroxylamine and, indeed, aniline itself (Rickert, 1987). The metabolism of both nitrobenzenes and anilines is important because many of the toxicological effects of these compounds are caused by their metabolites. For example, there is evidence that methemoglobinemia is caused by the interaction of hemoglobine with metabolites of nitrobenzenes (i.e., phenylhydroxylamine and the iminoquinone form an electrophilic adduct that could bind tissue proteins). Moreover, the formation of ROS during the reduction of nitrobenzenes may lead to oxidative stress, compromising the functionality of cells such as hepatocytes (Gutteridge, 1995).

As already explained, LO(A)EL experimental data are biased by several issues concerning the low reproducibility of experimental protocols and the wide range of diverse biological mechanisms responsible for the overall toxicity. For clarity, our reasoning mostly addressed the mispredicted compounds in the attempt to find a chemical as well as a biological rationale.

Methyl tert-butyl ether (CAS number: 1634-04-4) is an example of a recurrent misprediction through our five models. A log LOEL = 2.0 is reported for this compound within the RepDose database. However, the prediction made on the basis of two nearest neighbors, i.e. diethyl ether (SI = 0.924; log LOEL = 3.301) and 2-methyl-1-propanol (SI = 0.894, log LOEL = 3.00), leads to an error greater than 1 log unit. From a metabolic point of view, diethyl ether could be considered a safer chemical because it degrades into less toxic products (Booth and McDonald, 1982). On the other hand, tert-butyl ether is slowly metabolized to tert-butyl alcohol, which is eliminated from blood more slowly and which increases its half-life with dose (McGregor, 2010). 2-Methyl-1-propanol, despite being bulkier than diethyl ether, is an alcohol that does not require preliminary biotransformation (e.g., hydrolysis) in order to be conjugated in phase II metabolism, so it could be eliminated more easily from the body.

Phenol (CAS number 108-95-2, log LOEL = 3.10) is well predicted by the model with $k = 2$. However, it is mispredicted by models with $k = 3$ and 4 because of his third neighbor, hydroquinone (SI = 0.912, log LOEL = 1.699), which is more toxic than the other ones. Indeed, hydroquinone can easily convert to quinone, a well-known highly redox active molecule, and to semiquinone radical, leading to formation of reactive oxygen species (ROS), including superoxide, hydrogen peroxide, and the hydroxyl radical. Production of ROS can cause severe oxidative stress within cells through the formation of oxidized cellular macromolecules, including lipids, proteins and DNA. Furthermore, ROS can activate a number of signaling pathways leading to several toxic effects (Bolton et al., 2000).

As shown from those examples, metabolism is often the cause of the conversion of apparently less toxic into more dangerous substances. Such aspects cannot be clearly taken into account by the *k*-NN models. It will be desirable that future models could make an estimation of possible metabolite chemicals in order to obtain more accurate predictions.

4.3 Comparison with other LO(A)EL predictive models

A model for the prediction of LO(A)EL was reported by Mazzatorta et al. (2008), who applied an approach that integrated

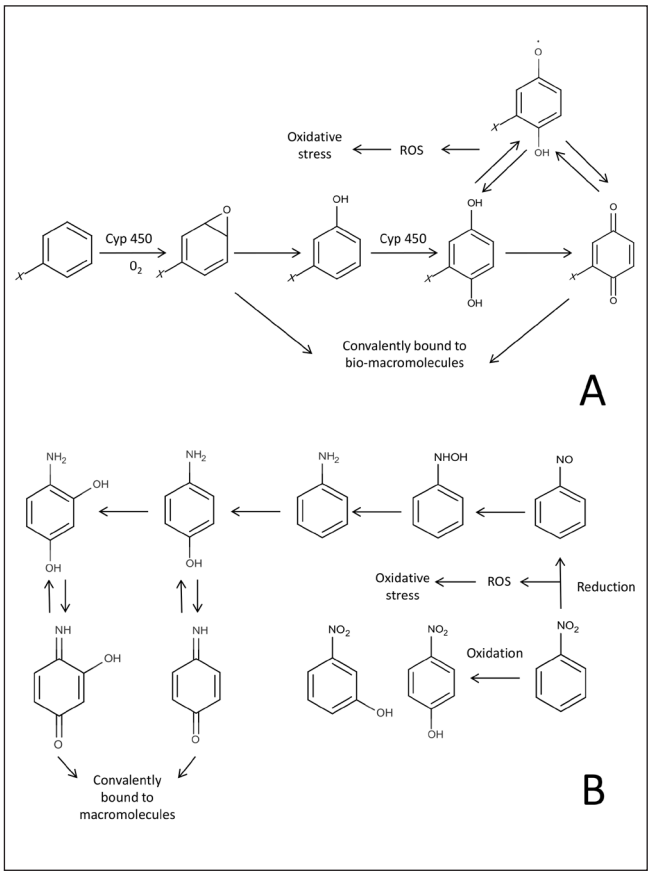

Fig. 3: Mechanism of action of toxicity of halobenzenes and nitroaromatic compounds

Some highly electrophilic species, such as quinone and imminoquinone derivatives, can covalently bind tissue proteins. Contextually, the formation of Reactive Oxygen Species (ROS) could lead to oxidative stress. Marvin (version 5.9.2, 2012 ChemAxon, http://www.chemaxon.com) was used for drawing chemical structures.

a genetic algorithm (GA) and partial least squares (PLS). The model comprised 19 descriptors selected among those available in the Dragon software and was trained on more than 445 chronic toxicity data sets. Selected descriptors were then used to derive a LO(A)EL predictive model through a leave-one-out stepwise multiple linear regression (LOO-SMLR).

De Julián-Ortiz et al. (2005) modeled a dataset of chronic LO(A)EL data for 234 compounds compiled from different sources using MLR. The model was based on 15 topological descriptors selected by a Furnival-Wilson algorithm among those included in the DESCRI program. Both MLR and the Furnival-Wilson algorithm were further applied to a smaller (86 compounds) but more homogeneous dataset. Moreover, De Julián-Ortiz et al. also derived a classification model based on 12 variables from the same pool of 86 chemicals. The classification takes three categories into account (i.e., high, medium and low LO(A)EL values). A LDA for model derivation and a stepwise procedure for variable selection were applied. Then, the model was validated on 17 external chemicals.

García-Domenech et al. (2006) applied the same described algorithms (Furnival-Wilson for descriptor selection and MLR for model derivation) on the same dataset of 86 molecules, the feature selection was however performed on a larger pool of descriptors. The model, based on 6 descriptors, was validated on 16 external chemicals.

TOPKAT, originally developed by Health Design Inc., is a program for the prediction of a wide variety of endpoints. It includes a module for quantitative LO(A)EL prediction, implementing a 44-descriptor MLR model initially based on chronic oral rat data of 234 chemicals (Mumtaz et al., 1995), subsequently increased to 393 chemicals. Descriptors were selected from an initial pool of electronic, topological, symmetry descriptors and molecular connectivity indices.

Sakuratani et al. (2013), from a training set of 500 chemicals, defined a series of 33 chemical categories related to a type of toxicity on the basis of mechanistic knowledge. Chemicals were assigned to a given category, then the LO(A)EL was derived as the result of a data gap filling approach by read-across on other chemicals within the category. The category library is incorporated in the Hazard Evaluation Support System (HESS) integrated computational platform.

Compared to our *k*-NN models, the above approaches are specifically aimed at long-term toxicity prediction (up to two years), whereas the models here are based on data accurately compiled referring only to subchronic toxicity (84-98 days). A great effort was spent in data selection because better LO(A)EL models can be derived using homogeneous and more reliable sources (de Julián-Ortiz et al., 2005). It is important to note too that the work described above used data related to different administration periods, without any restriction. Mazzatorta et al. (2008) selected studies conducted for more than 180 days. Sakuratani et al. (2013) considered periods from 28 to 120 days. Studies lasting for 12 or more months were considered for the TOPKAT database.

Importantly, previous models were not validated against a set of external compounds or, if they were, the validation returned poor results. However, our *k*-NN models returned solid statistics even on an external dataset. For the sake of clarity, additional details are reported for a fair comparison. In addition, performance of literature-based models also has been summarized in Table 2.

The model described by Mazzatorta et al. (2008) gave r^2 = 0.570 and RMSE = 0.700. A LOO cross-validation was done (q^2 = 0.500 and RMSE = 0.727, compared to the q^2 greater than 0.600 and RMSE about 0.400 of the *k*-NN models). However, external validation was not done, so the real model's predictive power was not known. Mumtaz et al. (1995) assessed the performance of the TOPKAT LO(A)EL model on the initial database, reporting that over 93% of predictions fell within a factor of 5 of the relative experimental LO(A)EL, and all predictions fell within a factor of 10. Further validation was done by Venkatapathy et al. (2004) on two databases, of 343 and 313 chemicals, containing a significant number of non-training compounds. However, this analysis, which gives a more accurate representation of TOPKAT predictive power, returned different results from those reported by Mumtaz. For the first database

Tab. 2: Literature-based QSAR models for RDT in rodents
The applied methods, the number of descriptors, the size of TS and VS and the performance on both datasets are reported for each model.

Model	Methods	Descriptors	Training set size	Test set size	Performance on training set	Performance on test set
k-NN models for LO(A)EL	*k*-NN	Fingerprint + 3 structural keys	254	179	q^2 =0.632-0.769 and RMSE =0.396-0.478 (LOO cross-validation)	r^2 =0.543-0.682 and RMSE =0.612-0.659
Mazzatorta et al., 2008	GA-PLS for descriptor election; LOO-SMLR for model derivation	19	445	none	r^2 =0.570 and RMSE =0.700 q^2 =0.500 and RMSE =0.727 (LOO cross-validation)	none
De Julián-Ortiz et al., 2005	Furnival-Wilson algorithm for descriptor selection; MLR for model derivation	15	234	none	r^2 =0.524 and RMSE =0.74	none
		6	86	none	r^2 =0.647; RMSE =0.66	none
De Julián-Ortiz et al., 2005	Stepwise procedure for descriptor selection; LDA for model derivation	12	86	17	86.2% successful classification	70.6% successful classification
Garcia-Domenech et al., 2006	Furnival-Wilson algorithm for descriptor selection; MLR for model derivation	11	86	16	r^2 =0.795 and RMSE =0.517 q^2 =0.719 and RMSE =0.564 (LOO cross-validation)	r^2 =0.712 and RMSE =0.853
TOPKAT	MLR	44	234	343*	within a factor of 5 and 10: 93% and 100% (Mumtaz et al., 2005)	within a factor of 5 and 10: 60% and 80% Venkatapathy et al., 2004
				313*		within a factor of 5 and 10: 69 and 79% Venkatapathy et al., 2004
Sakuratani et al., 2013	Read-across, data gap filling	33 chemical categories	500	none	none	none

*Some test set chemicals are already included in the training set

the percentages of chemicals whose prediction fell within a factor of 5 and 10 were respectively 60% and 80%. For the second database the percentages were 69 and 79% (compared with the 93 and 100% reported by Mumtaz et al., 1995). A further validation of TOPKAT was done by Tilaoui et al. (2007) on 340 substances typically occurring in food and not included in the TOPKAT training set. It was estimated that TOPKAT returned reliable predictions (i.e., with an error lower than 1 log unit) only for 33% of these chemicals (Tsakovska et al., 2008). Finally, the model presented by Sakuratani et al. (2013) does not make a quantitative assessment of the LO(A)EL but can be used to identify the target organ most likely affected by the chemical.

De Julián-Ortiz et al. (2005) obtained poor results for the MLR model (i.e., r^2 = 0.524 and RMSE = 0.74) on the first dataset. Results obtained for the second dataset (r^2 = 0.647 and RMSE = 0.66) were better but external validation was not done in either case. The LDA model returned good results for training (86.2% average successful classification) and external compounds (70.6% average successful classification). However, as it is a classification model, LO(A)EL continuous values cannot be predicted.

Better results have been obtained by the model described by García-Domenech et al. (2006) (i.e., r^2 = 0.795 and RMSE = 0.517). The model gave good results both in LOO cross-validation (i.e., q^2 = 0.719 and RMSE = 0.564) and in external validation (r^2 = 0.712 and RMSE = 0.853).

The models presented here return acceptable quantitative results in both internal and external validation. This suggests that they are more reliable than the others reported.

4.4 Uncertainty of data

The experimental determination of LO(A)EL is closely dependent on the employed protocol (e.g., the doses used). The test substance is administered daily at different doses to several groups of animals, one dose level per group, for a given period (in case of sub-chronic toxicity, 90 days). During this period the animals are observed for signs of toxicity. Then, the highest dose at which no adverse effects are noted (NO(A)EL) and the lowest dose at which an adverse effect is noted (LO(A)EL) are determined (OECD, 1998). Hence, there is intrinsic uncertainty in the LO(A)EL experimental data. Indeed, the "true" LO(A)EL (i.e., the real dose of the chemical that starts to generate any effect) may be anywhere between the NO(A)EL and the LO(A)EL. This uncertainty will be implicitly transferred into the predicted data generated by a model.

We analyzed the databases used for TS compilation (HESS, Munro and EPA's IRIS databases) in order to verify the spacing (i.e., the difference between two doses) for each chemical.

The spacing differed even in the same experiment and did not always seem to follow any specific pattern. However, most of the spacing was close to 0.30 and 0.70 logarithmic units, with a mean of about 0.50 for all the databases examined. Therefore, if the user prefers a conservative approach, we suggest subtracting 0.50 logarithmic units from the predicted values in order to take into account the difference between reported LO(A)ELs and true LO(A)ELs.

5 Conclusion

We presented here a *k*-NN algorithm developed for the prediction of RDT sub-chronic toxicity (LO(A)EL) in rats. Despite the complexity of the endpoint and the quality of the starting data, the models gave encouraging performance in internal and external validation. Due to the very restrictive conditions imposed on the algorithm, the models were able to predict only a small number of compounds. However, the results give us a firm belief that these models could be considered reliable, useful tools in a prioritization context for the prediction of toxicity data required by several regulations (European Commission, 2006; SCCS, 2012). The models presented here will be implemented within the VEGA platform (http://www.vega-qsar.eu) and will be freely available.

Supplementary data

Supplementary data may be found at http://dx.doi.org/10.14573/altex.1405091s. The percentage of refused prediction by each model for each algorithm criterion is reported in Table S1. The improvement in performance derived from the modulation of the maximum admitted error among nearest neighbors is shown in Table S2. The QMRF documents of the models are included in the supplementary data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Adler, S., Basketter, D., Creton, S. et al. (2011). Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. *Arch Toxicol 85*, 367-485. http://dx.doi.org/10.1007/s00204-011-0693-2

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat 46*, 175-185. http://dx.doi.org/10.1080/00031305.1992.10475879

Bolton, J. L., Trush, M. A., Penning, T. M. et al. (2000). Role of quinones in toxicology. *Chem Res Toxicol 13*, 135-160. http://dx.doi.org/10.1021/tx9902082

Booth, N. H. and McDonald, L. E. (eds.) (1982). *Veterinary Pharmacology and Therapeutics*. Ames, IA, USA: Iowa State University Press.

Cassotti, M., Ballabio, D., Consonni, V. et al., (2014). Prediction of acute aquatic toxicity toward Daphnia magna by using the GA-kNN method. *Altern Lab Anim 42*, 31-41.

Chan, K., Jensen, N. S., Silber, P. M. et al. (2007). Structure-activity relationships for halobenzene induced cytotoxicity in rat and human hepatocytes. *Chem Biol Interact 165*, 165-174. http://dx.doi.org/10.1016/j.cbi.2006.12.004

De Julián-Ortiz, J. V., García-Domenech, R., Gálvez, L. et al. (2005). Predictability and prediction of lowest o.bserved adverse effect levels in a structurally heterogeneous set of chemicals. *SAR QSAR Environ Res 16*, 263-272. http://dx.doi.org/10.1080/10659360500036927

Durant, J. L., Leland, B. A., Henry, D. R. et al. (2002). Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci 42*, 1273-1280. http://dx.doi.org/10.1021/ci010132r

ECHA (2008). Guidance on information requirements and chemical safety assessment. Chapter R.8: Characterisation of dose (concentration)-response for human health. http://echa.europa.eu/documents/10162/17224/information_requirements_r8_en.pdf (accessed April 2014)

Escher, S. E., Batke, M., Hoffmann-Doerr, S. et al. (2013). Interspecies extrapolation based on the RepDose database – a probabilistic approach. *Toxicol Lett 218*, 159-165. http://dx.doi.org/10.1016/j.toxlet.2013.01.027

European Commission (2006). Regulation (EC) of No 1907/2006 of the European parliament and of the council 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC.

European Commission (2009). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products.

Filipsson, F. A., Sand, S., Nilsson, J. et al. (2003). The benchmark dose method – review of available models, and recommendations for application in health assessment. *Crit Rev Toxicol 33*, 505-542. http://dx.doi.org/10.1080/10408440390242360

García-Domenech, R., de Julián-Ortiz, J. V. and Besalú, E. (2006) True prediction of lowest observed adverse effect levels. *Mol Diversity 10*, 159-168. http://dx.doi.org/10.1007/s11030-005-9007-z

Gissi, A., Gadaleta, D., Floris, M. et al. (2014). An alternative QSAR-based approach for predicting the bioconcentration factor for regulatory purposes. *ALTEX 31*, 23-36. http://dx.doi.org/10.14573/altex.1305221

Gutteridge, J. M. (1995). Lipid peroxidation and antioxidants as biomarkers of tissue damage. *Clin Chem 41*, 1819-1828.

Kalberlah, F., Schneider, K. and Schuhmacher-Wolz, U. (2003). Uncertaintynty in toxicological risk assessment for non-carcinogenic health effects. *Regul Toxicol Pharmacol 2003*, 92-104. http://dx.doi.org/10.1016/S0273-2300(02)00032-6

Kodell, R. L. (2009). Replace the NOAEL and LOAEL with the BMDL01 and BMDL10. *Environ Ecol Stat 16*, 3-12. http://dx.doi.org/10.1007/s10651-007-0075-3

Kroes, R., Renwick A. G., Feron, V. et al. (2007). Application of the threshold of toxicological concern (TTC) to the safety evaluation of cosmetic ingredients. *Food Chem Toxicol 45*, 2533-2562. http://dx.doi.org/10.1016/j.fct.2007.06.021

Lilienblum, W., Dekant, W., Foth, H. et al. (2008). Alternative methods to safety studies in experimental animals: role in the risk assessment of chemicals under the new European Chemicals Legislation (REACH). *Arch Toxicol 82*, 211-236. http://dx.doi.org/10.1007/s00204-008-0279-9

Mazzatorta, P., Estevez, M. D., Coulet, M. et al. (2008). Modeling oral rat chronic toxicity. *J Chem Inf Model 48*, 1949-1954. http://dx.doi.org/10.1021/ci8001974

McGregor, D. (2010). Tertiary-Butanol: A toxicological review. *Crit Rev Toxicol 2010*, 697-727. http://dx.doi.org/10.3109/10408444.2010.494249

Mumtaz, M. M., Knauf, L. A., Reisman, D. J. et al. (1995). Assessment of effect levels of chemicals from quantitative structure-activity relationship (QSAR) models. I. Chronic lowest-observed-adverse-effect level (LOAEL). *Toxicol Lett 79*, 131-143. http://dx.doi.org/10.1016/0378-4274(95)03365-R

Nicolotti, O. and Carotti, A. (2006). QSAR and QSPR studies of a highly structured physicochemical domain. *J Chem Inf Model 46*, 264-276. http://dx.doi.org/10.1021/ci0502931

OECD (1998). Test No. 408: Repeated Dose 90-day Oral Toxicity Study in Rodents, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing. http://dx.doi.org/10.1787/9789264070707-en

Pauwels, M. and Rogiers, V. (2010). Human health safety evaluation of cosmetics in the EU: A legally imposed challenge to science. *Toxicol Appl Pharmacol 243*, 260-274. http://dx.doi.org/10.1016/j.taap.2009.12.007

Raevsky, O. A., Grigor'ev, V. Y., Liplavskaya, E. A. et al. (2011). Prediction of acute rodent toxicity on the basis of chemical structure and physicochemical similarity. *Mol Inf 30*, 267-275. http://dx.doi.org/10.1002/minf.201000145

Rickert, D. E. (1987). Metabolism of nitroaromatic compounds. *Drug Metab Rev 18*, 23-53. http://dx.doi.org/10.3109/03602538708998299

Russell, W. M. S. and Burch, R. L. (1959). *The Principles of Humane Experimental Technique*. London, UK: Methuen.

Sakuratani, Y., Zhang, H., Nishikawa, S. et al. (2013). Hazard evaluation support system (HESS) for predicting repeated dose toxicity using toxicological categories. *SAR QSAR Env Res 24*, 351-363. http://dx.doi.org/10.1080/1062936X.2013.773375

Sand, S., Victorin, K. and Filipsson, A. F. (2008). The current state of knowledge on the use of the benchmark dose concept in risk assessment. *J Appl Toxicol 28*, 405-421. http://dx.doi.org/10.1002/jat.1298

SCCS – Scientific Committee on Consumer Safety (2012). The SCCS's notes of guidance for the testing of cosmetics substances and their safety evaluation 8th revision. http://ec.europa.eu/health/scientific_committees/consumer_safety/docs/sccs_s_006.pdf (accessed April 2014).

Setzer, R. W. and Kimmel, C. A. (2003). Use of NOAEL, benchmark dose, and other models for human risk assessment of hormonally active substances. *Pure Appl Chem 75*, 2151-2158. http://dx.doi.org/10.1351/pac200375112151

Stoyanova-Slavova, I. B., Slavov, S. H., Pearce, B. et al. (2014). Partial least square and *k*-nearest neighbor algorithms for improved 3D quantitative spectral data-activity relationship consensus modeling of acute toxicity. *Env Tox Chem 33*, 1271-1282. http://dx.doi.org/10.1002/etc.2534

Tilaoui, L., Schilter, B., Tran, L. et al. (2007). Integrated computational methods for prediction of the lowest observable adverse effect level of food-borne molecules. *QSAR Comb Sci 26*, 102-108. http://dx.doi.org/10.1002/qsar.200610060

Tluczkiewicz, I., Batke, M., Kroese, D. et al. (2013). The OSIRIS weight of evidence approach: ITS for the endpoints repeated-dose toxicity (RepDose ITS). *Regul Toxicol Pharmacol 67*, 157-169. http://dx.doi.org/10.1016/j.yrtph.2013.02.004

Tsakovska, I., Lessigiarska, I. and Netzeva, T. (2008). A mini review of mammalian toxicity (Q)SAR models. *QSAR Comb Sci 27*, 41-48. http://dx.doi.org/10.1002/qsar.200710107

Venkatapathy, R., Moudgal, C. J. and Bruce, R. M. (2004). Assessment of the oral rat chronic lowest observed adverse effect level model in TOPKAT, a QSAR software package for toxicity prediction. *J Chem Inf Comput Sci 44*, 1623-1629. http://dx.doi.org/10.1021/ci049903s

Willett, P., Barnard, J. M. and Downs, G. M. (1998). Chemical similarity searching. *J Chem Inf Comput Sci 38*, 983-996. http://dx.doi.org/10.1021/ci9800211

World Health Organization (1999). Principles for the Assessment of Risks to Human Health from Exposure to Chemicals. Environmental Health Criteria 210. Geneva. WHO. http://www.inchem.org/documents/ehc/ehc/ehc210.htm (accessed April 2014).

Acknowledgements

We acknowledge the financial support of the European Commission and Cosmetics Europe, within the project ToxBank.

Correspondence to

Emilio Benfenati
Laboratory of Environmental Chemistry and Toxicology
IRCCS-Istituto di Ricerche Farmacologiche Mario Negri
Via La Masa 19
20156 Milano, Italy
Phone: +39 02 3901 4420
e-mail: emilio.benfenati@marionegri.it

Orazio Nicolotti
Dipartimento di Farmacia – Scienze del Farmaco
Università degli Studi di Bari "Aldo Moro"
Via E. Orabona 4
70125 Bari, Italy
Phone: +39 080 544 2551
Fax: +39 080 544 2230
e-mail: orazio.nicolotti@uniba.it