Available online on 15 Dec, 2020 at https://ijdra.com/index.php/journal

International Journal of Drug Regulatory Affairs

Published by Diva Enterprises Pvt. Ltd., New Delhi
Associated with Delhi Pharmaceutical Sciences & Research University

Review Article

Understanding OECD Guidelines for QSAR Models-A Software development and harmonization strategy

Subba Rao Bayya*

Associate Professor, RBVRR Women's College of Pharmacy, Barkatpura, Hyderabad-500 027, Telangana State, India

Abstract

Post GATT era has turned towards harmonization processes not only relating to trade, services, and taxation but also towards research processes to some extent. The current guideline especially helps in developing software that helps end users of software in overcoming ambiguities and come to a faster acceptance of the research by the authority. The objective of harmonization is to minimize time, expenditure, experimental animals, scrutiny and approval processes. The current article is to know how the Quantitative Structure Activity Relationship research is made uniform for the research community by developing a policy/guideline as an international standard setting.

Keywords: OECD, guidelines, QSAR, QSAR Validation Process, Cooper Statistics, Receiver Operating Characteristic curve (ROC)

Article Info: Received 30 Oct. 2020; Review Completed 03 Dec. 2020; Accepted 11 Dec. 2020

Cite this article as:

Bayya SR. Understanding OECD Guidelines for QSAR Models-A Software development and harmonization strategy. Int J Drug Reg Affairs [Internet]. 15 Dec.2020 [cited 15 Dec.2020]; 8(4):25-36. Available from: http://ijdra.com/index.php/journal/article/view/431

DOI: 10.22270/ijdra.v8i4.431

*Corresponding author Tel.: +91-9440763781;
E-mail address: drbayyasubbarao@yahoo.com (Subba Rao Bayya).

1. Introduction

Quantitative Structure Activity Relationship (QSAR) is a process of developing either new chemical entities as leads or optimization of leads for desired activity while minimizing un-desired activity. This is a part of drug discovery and right from student level to the research community the fundamental concepts are well versed. The objective of the current review is to enlighten the established and accepted protocol so as to conduct research relating to QSAR and even develop software models in pharmaceutical research perspective.

2. Brief Introduction to OECD (1)

The Organization for Economic Co-operation and Development (OECD) is an intergovernmental organisation of thirty industrialized countries of North America, Europe, Asia and Pacific region. The key role of the OECD is to identify international problems, discuss, co-ordinate and frame policies to meet the objective of minimizing time, scrutiny and approval process. The administration of the OECD is at the Secretariat located in Paris. It is here the directorates and divisions are established. The OECD has about 200 specialized committees and working groups composed of delegates of member countries. Currently, there are 37 member countries, 1 candidate country, 5 key partner countries (i.e., Brazil, China, India, Indonesia, South Africa) and with 6 regional initiatives (i.e., Africa, Eurasia, Latin America, Middle East and North Africa, South East Europe and South East Asia). The current guideline is released by Environment Directorate, joint meeting of the chemicals committee and the working party on chemicals, pesticides and biotechnology. The objective in establishing the current protocol is to develop reliability and repeatability of QSAR experiments by researchers and ensure chemicals are grouped using several market available computer models with necessary validations. In November, 2004 at the 37th OECD meeting the OECD principles for the validation for regulatory purposes of QSAR models were agreed.

3. Brief Introduction to QSAR and Proposal for Validity of QSAR at OECD (1)

Earlier chemical entities were synthesized, chemically characterized and tested for biological activity based on minimal rationality. Such entities synthesized may or may not be fruitful of the researchers' objectives. Quantitative Structure Activity Relationship (QSAR) has made rational drug design more predictive before synthesizing and testing. This is due to the earlier scientists who made incremental research into computational data that is now computerized (*in-silico*) so that current researchers can use the computed data to establish predictions of activity. Historically, at the international workshop on the "Regulatory Acceptance of QSARs for Human Health and Environment Endpoints" organized by International Council of Chemical Associations (ICCA) and the European Chemical Industry Council (CEFIC), held on 4-6 March, 2002 at Setubal, Portugal, a proposal was made on set of principles for assessing the validity of QSAR model to overcome variations in acceptance. In November 2002, at the 34th joint meeting, the proposal was accepted and to establish an Expert Group. On 31st March-2nd April, 2003 the first meeting of the Expert Group hosted by European Commission's Joint Research Centre (JRC) was conducted at Ispra, Italy. In the meeting, the Expert Group has proposed a two year work plan as Work Items such as applying the validation principles agreed (Work Item 1), develop guidance documents to develop, validate and regulatory application of QSARs (Work Item 2) and identify practical approaches to enable QSAR to be readily available, accessible, including development of databases of accepted QSARs that are available at regulatory, industry and universities.

In the process of implementing the Setubal Principles, eleven case studies were considered i.e., acute fish toxicity, atmospheric degradation, mutagenicity, carcinogenicity, software models such as Multi-CASE model for in-vitro chromosomal aberrations, Multi-CASE and MDL models for human NOEL, ECOSAR, BIOWIN, Derek, the Derek Skin sensitization rule base, the Japanese METI biodegradation model, the rat oral chronic toxicity models in TOPKAT. The cases considered were found covering physicochemical, environmental, ecological and human health endpoints. A check list was also developed so as to establish the guidance.

4. OECD Agreed Setubal Principles (1)

The five principles for QSAR models that were agreed for regulatory purposes are

- **Defined endpoint:** The computer model has to predict the end point such as physico-chemical, biological or environmental effects. The model has to be developed using homogenous dataset that is using a single protocol. This indicates that the collection of experimental past data should be from literature that follows same to same procedures. However, there is a provision to use data obtained by different procedures such as different animal models etc.
- **An unambiguous algorithm:** The algorithm developed as a computer model should be unambiguous in decision making. For this the software providers either commercial or free should provide the details of the methods used for calculations.
- **A defined domain of applicability:** For a researcher, a domain is a macromolecule that is involved in disease condition. Researchers develop ligands to bind the macromolecule to treat the disease condition. The structural features of the ligands should be in close proximity and a model should be in a position to outlier those ligands that are far away in binding features to the macromolecule.
- **Appropriate measures of goodness of fit, robustness and predictivity:** To achieve this, the researcher is expected to develop or use a model that is able to assess or establish the activity by a sample of past experimental data as training data set and using a part of past experimental test data set, the model is reassessed whether working or not for expected activity as prediction.
- **A mechanistic interpretation, if possible:** In QSAR mathematical model, for a dependent variable several independent variables are used. The independent variables can be called as descriptors. When a mathematical model is established, with respect to the descriptors, the researcher should be in a position in interpretation of the mechanism of activity.

5. OECD definitions for Validation, Reliability and Relevance (1)

Validation: The process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose.

Reliability: Measures of the extent that a test method can be performed reproducibly within and between laboratories over time, when performed using the same protocol. It is assessed by calculating intra- and inter-laboratory reproducibility and intra-laboratory repeatability.

Relevance: Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of the accuracy (concordance) of a test method.

The definitions indicate reproducibility of experimental results within and between laboratories irrespective of time, scientific basis for expecting an experimental method to predict a response of end points.

6. The QSAR Validation Process (1)

The OECD guidance document indicates to excise the principles on the model or set of models and also indicates that the validation process need not be carried by an organisation, committee or formal validation body. This indicates that the guidance document helps in developing models by developers, and researchers to fulfill prior implementation conduct of further research and publication of research. Hence, validity judging needs chemical domain, end point, performance and statistics assessing goodness-of-fit, robustness and predictivity. The outcome of the process should be providing a dossier (say an equation) that ensures performance and transparency. One interesting point is that the guidance document insists on validation of model but not the software programme, indicating that highly predictive model is valid. In terms of application of principles, the models, decision trees, neural network models for an endpoint should be ensured.

7. OECD Principles

7.1 OECD Principle 1-Defined Endpoint in Detail (1)

An endpoint is defined as the measures of activity for chemicals made under specific conditions. For a Quantitative Structure Activity Relationship, it is necessary to bring out the relation between activity and the chemical structure. Here, the activity is the endpoint and the chemical structure involves with various descriptors. Table 1, illustrates the various endpoints and the descriptors.

Table 1. Endpoint and Some Descriptors

Physicochemical Properties	Environmental Fate	Ecological Effects	Human Health Effects
Melting Point, **Boiling Point,** **Vapour Pressure,** **K octanol/water,** **K organic C/water*,** **Water Solubility**	Biodegradation, Hydrolysis, Atmospheric, Oxidation, Bioaccumulation*	Acute Fish, Toxicity, Acute Daphnid, Toxicity, Alga Toxicity, Long-term Aquatic Toxicity, Terrestrial Effects	Acute Oral Toxicity, Acute Inhalation Toxicity, Acute Dermal Toxicity, Skin Irritation /Corrosion*, Eye Irritation/Corrosion *, Skin Sensitisation *, Repeated Dose, Genotoxicity (in vitro), Genotoxicity (in vitro, non bacterial), Genotoxicity (in vivo), Reproductive Toxicity, Developmental Toxicity, Carcinogenicity*, Organ Toxicity (*e.g.*, hepatotoxicity, cardiotoxicity, nephrotoxicity, etc.)

*** non-SID endpoints (i.e., Screening Information Data Set)**

Activity of a chemical has to be interpreted in broad sense. This is because, a chemical entity not only possess desired activity but also several other activities. Hence, a chemical entity may fall into one desired end point or several undesired end points. In a researcher point of view the end point of biological activity as well as toxic effects play a role. The various descriptors/end points may be in the form of induction of cytochromes, hypertrophy of hepatocytes, serum levels of aminotranferases, increased relative kidney weight, range of clinical signs, body weight, food consumption changes, clinical chemistry attributes, haematology parameters, macro and microscopical parameters, acute oral toxicity in rats, skin irritation/corrosion etc. Hence while quantification of an endpoint, it is necessary to follow standard permitted procedures. But, in case of classification and labeling of chemicals, there is a necessity of genotoxicity, repeat-dose-toxicity as No-Observed-Adverse-Effect-Level (NOAEL) to be conducted with various species such as dogs, rats or mice either administering test entity as feed/gavage /capsules from a duration of three weeks to an year or even more. Under these differences in species, administering and term of study, the so called battery of different in vitro and in vivo test protocols may even be considered for end point. Hence, principle 1 aims on clarity and transparency of the model. It is necessary to understand that each descriptor value has to be achieved with well accepted models, transparently. The guideline clearly indicates the current principle should be considered in parallel with other principles. Where, several models or heterogeneous data are necessary to build a descriptor or model, may be taken into consideration, but validity, transparency, repeatability, reproducibility has to be achieved. Especially in determining Quantitative Structure Bioactivity Relationships (QSBRs) a variation of 20% is accepted, even though standardized by OECD.

7.2 OECD Principle 2-Unambiguous Algorithms in Detail (1)

The guideline indicates algorithm as the relationship between descriptors of chemical structure and activity (i.e., end point). The algorithm may be a mathematical or knowledge based rule developed by one or more skilled in the art. Hence algorithm helps in overcoming ambiguity and validation process of QSAR model. In other words, if the algorithm is publicly available or not, the process helps other researchers achieve reproducible results, how endpoint results are achieved statistically, that may even help reviewers in drawing firm conclusions indicating accepted approach of developing the model. But, the developers of the model should give clear description how the model end points are achieved as estimates and that led to reproducibility. Role of statistics is emphasized how one can achieve transparency of the model and the interpretation of the algorithm (as predictors or coefficients) as a cause-effect relationship. Such algorithms with statistics are also found useful where there is limited chemical descriptors and large variability in data for conclusions. It is advised for a neural network where the model has been developed by learning and giving a prediction.

In order to assess an algorithm, the following elements are necessary for consideration:

i. Dataset of chemicals, descriptor and end-point values.

ii. A description of descriptors used, their way of development and measurement

iii. Description of test and training set, if any outliers are removed an explanation for justification
iv. A mathematical model between the descriptor and the end point and their relationship.
v. A statistical model describing the performance of the model
vi. Parameters and values that constitute the QSAR

The guideline emphasizes the usage of univarate, multiple linear regression, principal component analysis/principal component regression, partial least squares (a combination of multiple linear and principle component regression), artificial neural nets (ANN) (for pattern recognition, process analysis and non-linear modeling), Fuzzy clustering and regression, K-nearest neighbouring clustering, genetic algorithms (GA) i.e., artificial intelligence. Linear model is preferred to overcome ambiguous in non-linear models.

In case of neural net, it is found to be flexible when compared to statistics, but may have to need large data that leads to ambiguity. In neural net, either supervised or un-supervised learning is made. In the former system it is forced to assign an object to a specific class as training set where as in the latter, clusters are formed without any prior information. Fuzzy clustering and regression is helping using probabilities for inclusion of an object in a class rather than hardly. K-nearest neighbouring helps the closeness of an object in the class. Genetic algorithm involves with natural selection, generating formula, developing control strategies, giving several solutions, developing fitness, retaining or discarding and replacing with new population and running the procedure several times.

7.3 OECD Principle 3-Defined Domain of Applicability in Detail (1)

The principle, using the training set establishes the scope and limitations of the model using structural, physicochemical and response information. Those chemicals that are close to the training set will be predicted and those chemicals that are outside the applicability domain (AD) are extrapolated and are less reliable. The model helps in developing confidential interval in ensuring degree of similarity of the test chemical with the training set. Applicability domain is the mechanical structural requirements that are derived from interactive hypothesis generation and testing in design of the training set. In other words, with defined physico-chemical properties, similarities are achieved that in turn are helpful to define applicability domain (AD) that in turn develops a model for predictions with reliability. There are possibilities that a chemical which is of interest not being captured by the model and the vice versa, may or may not be acting by a different mechanism. Under such circumstances, the model has to be refined by inclusion and exclusion rule with respect to training set. Hence, in defining applicability domain (AD), one has to establish the limits of every descriptor for activity keeping in view of the mechanism/s of action. For this purpose, even though un-reliable, every descriptor is assumed to follow normal distribution and a range is established. In order to achieve the interpolation, the training set molecules when plotted one descriptor versus the other descriptor may fall well apart or in a circular region. But, still there exists spaces between molecules. To overcome these setbacks, several other approaches like the distance between the query chemical and a defined point in the descriptor space of the model, iso-distance contours in the interpolation space can also use. Hotelling's test with leverage statistics helps in assessing the leverage of a chemical. Chemicals in the training set have leverage values between 0 and 1. Warning leverage (h*) is fixed 3p/n, where p is number of descriptors plus one and n is number of training chemicals. A well-recognized and commonly used William's plot, Figure 1, comprising of standardized residuals (R) vs. leverages (or hat values, h) helps in better visualization of outliers both in the descriptor and response space and, Figure 2, helps in understanding the correlation among the parameters. With respect to QSAR AD acceptability, statistics either in the form of parametric (like normal, poisson) or non-parametric (kernel density estimation function) are found robust than range, distance and leverage approaches. Tanimoto coefficient arithmetic, ranging 0 to 1, is a ratio of shared substructures, to the number of all substructures appears in the training set. The value of zero indicates no similarity where as one indicates identical. Such arithmetic or descriptors as an example are emphasized so as to finalize, describe applicability domain (AD).

In decision making, it may be necessary to monitor the endpoint on one or more QSAR models which follow different approaches. In such circumstance, if the query molecule falls in intersection of the AD of different models, the predictions of different models have to be averaged. It is indicated that addition of two QSARs have high specificity and low or moderate sensitivity resulting in high overall specificity and sensitivity.

A four stage approach was proposed for determining the model AD that includes:

Stage 1: Identifying whether chemical falls in the range of variation in physicochemical properties of the model.

Stage 2: Defining the model and prediction of similarity of the query chemical and chemicals.

Stage 3: Mechanistic check of the model by assessing whether the chemical contains specific reactive groups hypothesized to cause the effect.

Stage 4: Based on the assessment of probability, the model has to check whether the chemical is metabolically activated.

The proposed four sequence approach of multiple AD is found to increase reliability of prediction for chemicals that satisfy all the conditions for inclusion in the acceptability domain (AD).

7.4 OECD Principle 4-Measures of Goodness-of-fit, Robustness and Predictivity in Detail (1)

The principle mainly speaks the necessity of statistical validation of the model established. The statistical validation helps in ensuring the model performance internally (goodness-of-fit and robustness) and externally (predictivity).

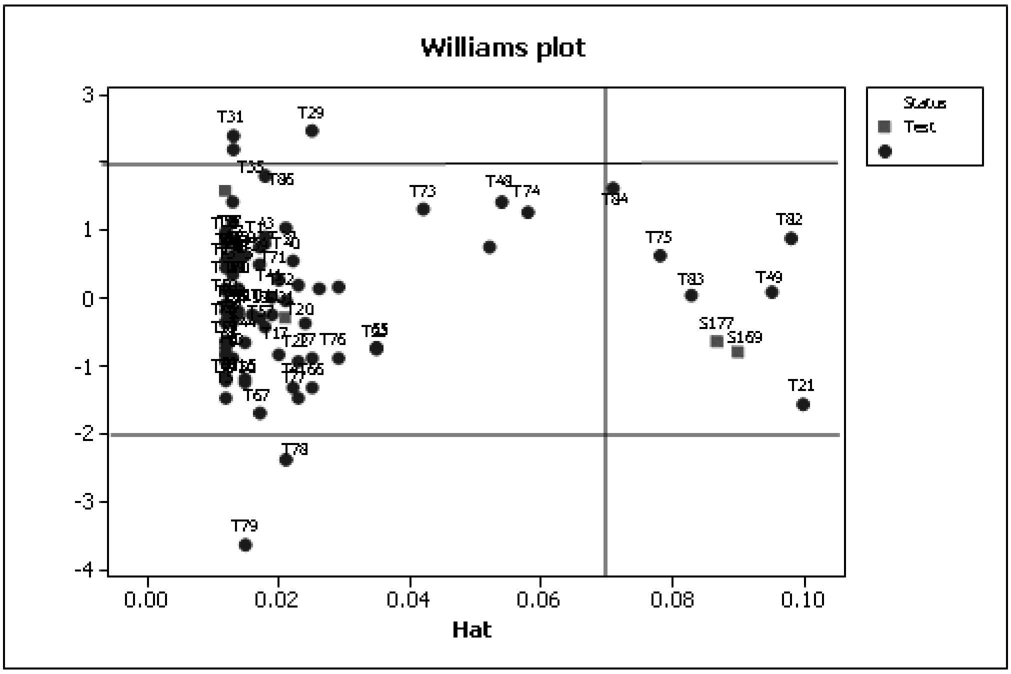

Figure 1. William's Plot

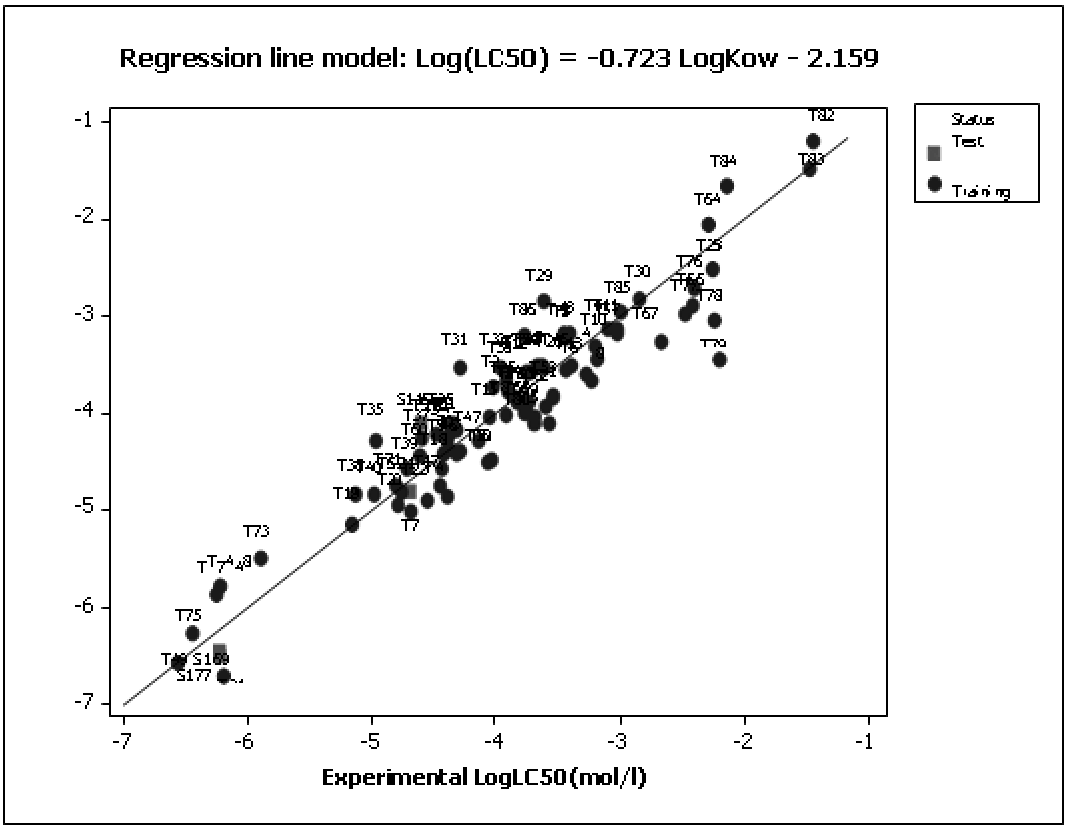

Figure 2. Regression line between Theoretical and Practical

Statistical validation helps in developing a model that overcomes "under fitted", "over fitted", "too simple", "too complex", providing "modeling noise", identifying spurious models, comparing among models etc. Statistics and acceptability domain limits have to be kept in mind for a better performance and robustness. During model validation, statistics helps a chemical outside the acceptability domain (AD) unlikely to be predicted with desired level of reliability.

As the previous principles indicate, one has to identify various end points, compile all data sets comprising chemicals. The data set is divided into a training set and a test set. The training set helps in deriving the model and the test set helps in testing the model. Hence test set chemicals are not used in derivation of the model. Predictors (i.e., molecular descriptors) called as variables of the model helps in optimizing model complexity and testing hypothesis regarding mechanism. This helps in judging whether more dataset of chemicals necessary.

Once a model is developed using training set, one can predict/estimate endpoint either using training or test set. Accuracy can be ensured when the chemical is predicted/ estimated and its closeness to the reference value. Greater the proportion of accurate predictions, the more the model is reliable.

In order to assess the goodness-of-fit, the model should detect the variance in the response of the various training set chemicals. When a query chemical's predicted value if within the range it is called as interpolation and if outside the range is called as extrapolation. Hence, a model involving mechanistic should develop range, able to interpolate or extrapolate query chemical and cross checked whether the model to be hold.

Model robustness is achieved, by achieving stability of the model parameters (predictor coefficients). This can be ensured by perturbation (deletion of one or more chemicals). This means that when chemical is removed from training/test set and model developed, and when another chemical removed and model developed with remaining set should give same model with permissible variation or so.

Predictive ability (or power or capacity or predictivity) of a model is ensured by using test set data on the model, which in turn is developed by training set.

For model developers or users, the guideline indicates multiple linear regression analysis (MLR), where, in pharmaceuticals 'y' is biological activity (dependent variable) and x is the predictor (independent variable or molecular descriptors). Regression coefficient is achieved using least square method and by minimizing the sum of the squared residuals. If the descriptors (or variables) used in the model have mean of zero and standard deviation of one, then the regression coefficients in the model are called beta coefficients. Beta coefficient helps to understand the relationship between one independent variable with the dependent variable and if the value is higher, then greater the variance among the independent and dependent. Goodness-of-fit of the model is ensured by R^2 (regression coefficient), which in turn helps to know the proportion of variation of 'y'. If $R^2 = 0$, $R^2 =1$ and $R^2 >$ 0.5 then it indicates no linear, linear and explained variance is higher than the unexplained relating dependent and independent variables respectively. A model to be accepted should have R^2 at least 0.8. In order to avoid over fitting, if necessary, "$R^2{}_{adj}$" is used instead of "R^2". In the "$R^2{}_{adj}$", the residual sum of squares and the total sum of squares is divided by respective degrees of freedom. Standard error helps in understanding dispersion of observed value from regression line and is calculated from observed and observed dependent variable values. If the standard error of estimate is smaller than the experimental error of the biological data, it indicates over fitted model. Yet statistical significance of the regression model can be ensured by F-value which in turn is the ratio of explained and un-explained variance for a given number of degrees of freedom. The higher the F-value, the greater probability that the equation if significant. Student 't' test helps to know the significance of regression coefficient (R^2). It is used to test the hypothesis on regression coefficient is zero. If the hypothesis is true, then the predictor variable does not contribute to explain the dependant variable. Higher the 't' value, greater the significance of the regression coefficient. The calculated 't' value is compared with standard Tabled 't' value to ensure significance.

Partial Least Squares (PLS) is used in Multiple Linear Regression (MLR) X-block of 'p' predictors and a single 'y' response (PLS1) or Y-block or 'r' responses. The advantage with PLS is that it tolerates up to 20 percent of missing data. PLS also helps in generating matrix of scores relating summarized 'x', 'y' variables, matrix of weights relating 'x' and 'y' and matrix of weights relating 'y' and 'x'. The scores help in understanding similarities/ dissimilarities among compounds and weights help in bringing relationship between 'x' and 'y'. Quantitatively, R^2 helps in measuring goodness-of-fit, contrary to PLS. PLS model is characterized by $R^2(Y)$, $R^2(X)$, $R^2(Y)_{adj}$, $R^2(X)_{adj}$. In order to avoid over fitting, usage of large 'x' variables leading to high $R^2(Y)$ may not be sufficient criteria for validity of PLS model. Hence, a cross-validation procedure for $Q^2(Y)$, which in turn reduces complexity, parameter has to be calculated so as to get highest predictive ability. The difference between $R^2(Y)$ and $Q^2(Y)$ should not exceed 0.3. To identify the outliers, residual standard deviation (RSD) is calculated i.e., RSD for 'x' and 'y' variable separately. The former helps the relevance and the latter help how well the response to and by the PLS model respectively.

Classification Models

Chemicals are classified (Classification Model-CM) into active/inactive, pre-defined categories for scientific and regulatory purposes. As these chemicals may possess biological variability, the data may fall in one or more categories. The regulatory guideline relating to labeling of chemicals utilizes one or more symbols etc., as necessary. With relating to QSAR classification models, a variety of statistical linear methods may be employed such as Multivariate Discriminant Analysis (MDA), Logistic Regression (LR), Decision or Classification trees (CT), rule based models using if ….. then, and non-linear methods may be employed using Embedded Cluster Modeling (ECM), Neural Networks (NN), k-Nearest Neighbour (k-NN). To assess goodness-of-fit of the classification model (CM), Cooper statistics using Bayesian approach, may be employed for judgment either single or using combined results.

In a classification model, the results of the classification can be arranged, in confusion or contingency matrix, Table 2, where rows represent reference classes (Ag) and columns represent predicted classes (Ag'). With respect to interpretation, the main diagonal ($C_{Gg'}$) represents the cases where true classes coincide with the assigned classes. The non-diagnosized cells represent the misclassifications. Over predictions are to the right and above the diagonal and under predictions are to the left and below the diagonal. The right hand column represents the number of objects belonging to each class (n_g) and the below row represents the total number of objects assigned to each class according to the CM ($n_{g'}$).

Table 2. Confusion or Contingency Matrix {C_{GG}}for general case with G classes

		Assigned Class				
		A1’	**A2’**	**A3’**	**Ag’**	**Marginal Total**
True Class	**A1**	**$C_{11'}$**	$C_{12'}$	$C_{13'}$	$C_{1g'}$	n_1
	A2	$C_{21'}$	**$C_{22'}$**	$C_{23'}$	$C_{2g'}$	n_2
	A3	$C_{31'}$	$C_{32'}$	**$C_{33'}$**	$C_{3g'}$	n_3
	Ag	$C_{G1'}$	$C_{G2'}$	$C_{G3'}$	**$C_{Gg'}$**	n_g
	Marginal Totals	$n_{1'}$	$n_{2'}$	$n_{3'}$	$n_{g'}$	

In order to set misclassifications, (2,3) some classification errors may be worse and to quantify such errors, Loss Matrix, Table 3, is used. The matrix comprises of different weights for different types of classification errors. Here, the non-diagonal elements quantify the type of error in the classification. With respect to interpretation, classification error with classes A1, A3, Ag is more significant (loss value of 2) than with classes A1 and A2 (loss value of 1).

Table 3. Example of loss matrix {$l_{GG'}$} where the loss function has been arbitrarily defined in an integer scale

		Assigned Class			
		A1’	**A2’**	**A3’**	**Ag’**
True Class	**A1**	**0**	1	2	2
	A2	1	**0**	1	1
	A3	2	1	**0**	2
	Ag	2	1	2	**0**

In order to assess the goodness-of-fit of classification model (CM), several mathematical calculation are conducted i.e., concordance or accuracy (Non-error rate), error rate, No-Model error rate (NOMER %), prior probability of a class, prior proportional probability of a class, sensitivity of a class, specificity of a class, misclassification risk. The result after evaluation have to be compared with a reference and this is taken to be the one all objects are assigned to the class that is most represented. The reference condition refers to no model and hence called No-Model. The No-Model value is unique and independent from the classification method adopted. Goodness-of-fit values close to the ones of the No-Model condition give evidence of a poor result of the classification method.

The performance of a classification model (CM) is understood by measuring sensitivity (ability to detect known active compounds), specificity (ability to detect non-active compounds) and accuracy (ability to detect all chemicals). To understand the performance, Copper statistics, Table 5, are applied on 2 x 2 contingency, Table 4. The complement of specificity and sensitivity, one can calculate and predict false positive and false negative, if any. A positive predictivity indicates that a molecule predicted active is really active and the vice versa. High sensitivity indicates higher true positive rate. A high specificity indicates higher true negative rate and low false positive rate. It is also indicated that classification model (CM) should not only be assessed by statistics because of the positive and negative predictivities vary with proportion of active chemicals in the population, i.e., [(a + b)/N]. To achieve maximum classification performance; one has to keep in mind the model purpose, quality of predictor and response data. For a stand-alone classification model, the Cooper statistics should be greater than 50% and where classification model identifies active and inactive chemicals using a battery of models, a lower performance by Cooper statistics may be acceptable.

Table 4. 2 x 2 Contingency Table

		Assigned Class		
		Toxic	**Non-Toxic**	**Marginal Totals**
Observed (in vivo) Class	**Active**	a	b	a + b
	Non-Active	c	d	c + d
	Marginal Totals	a + c	b + d	a + b + c + d

Table 5. Definition of Cooper Statistics

Statistic	Formula	Definition
Sensitivity (True Positive Rate)	a/(a + b)	fraction of active chemicals correctly assigned
Specificity (True Negative Rate)	d/(c + d)	fraction of non-active chemicals correctly assigned
Concordance or Accuracy	(a + d)/(a + b + c + d)	fraction of chemicals correctly assigned

Positive Predictivity	a/(a + c)	fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative Predictivity	d/(b + d)	fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False Positive (Over classification) rate	c/(c + d) 1-specificity	fraction of non-active chemicals that are falsely assigned to be active
False Negative (Under classification) rate	b/(a + b) 1-sensitivity	fraction of active chemicals that are falsely assigned to be non-active

A classification model which has been developed by a data set of chemicals may vary significantly with another data set of chemicals. Hence, Cooper statistics confidence interval has to be established and this can also be achieved by bootstrap re-sampling technique. To compare number of classification models, Receiver Operating Characteristic curve (ROC) is suggested, which in turn is a plot of y-axis sensitivity (true positive rate) vs. x-axis 1-specificity (false positive rate). A good classification model indicates a point on the left top corner of the ROC space (i.e., high true positive and low false positive rates). If there is no discrimination in the model, a straight line is observed at an angle 45 degrees to the horizontal indicating equal rates of true and false positive. With respect to area under the curve, if the area is 1.0, it indicates perfect goodness of classification model where as a non-discriminating has an area of 0.5, which falls on the diagonal.

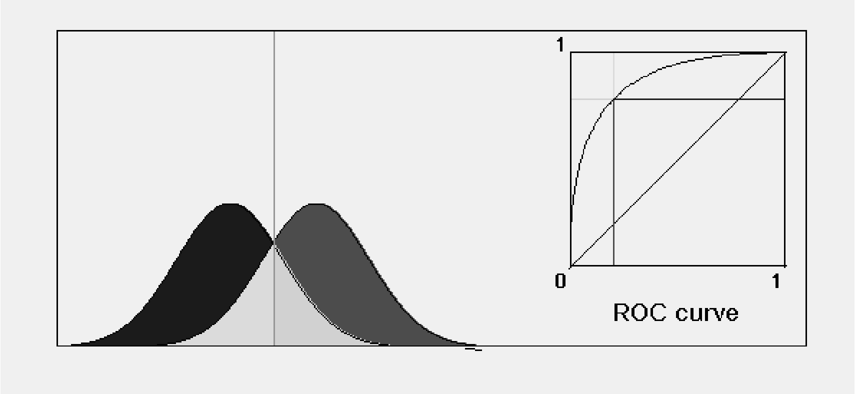

Figure 3. ROC Curve

If the classification model is expected to give continuous predictors (i.e., continuous values), the relationship between sensitivity and specificity have corresponding different thresholds. Hence an optimal threshold has to be established using ROC curve. As we know that both sensitivity and specificity equal to one, increasing one threshold leads to decrease in another, if the point is greater than or less than threshold, indicates positive and negative respectively. In a ROC curve, Figure 3, one curve represents distribution of true negatives and the other curve represents the true positives. If the threshold is increased (from left to right), false positive rate is decreased. As the false positive rate decreases, the true positive rate also decreases and in ROC curve it is represented at the bottom left of the curve. If the threshold is decreased, the true positives increases gradually corresponding top right of the curve. This entire description in common terms are usually called as type I and II errors.

In order to achieve good classification model, there is a necessity to set the importance of misclassification. Minimizing the errors will increase accuracy. Unequal errors and uneven class distributions leads to high costs and there is a necessity to compensate. This is achieved by modifying prevalence in the data set. Leave-one-out method helps in understanding the total number of misclassifications, which in turn assess the robustness of the classification model. The losses by various models help in developing a best fitted model so that resulting models are robust and accurate than by one single model.

Similar to human brain, Artificial Neural Network (ANN) is development of a model based on learning data. (4, 5) In the process several layers are formed (i.e., input, hidden, output) and upon repeated learning, connection weights are adjusted to minimize errors. Based on architecture and in learning, Artificial Neural Network (ANN) is categorized into two groups i.e., unsupervised and supervised self-organizing maps; and supervised back propagation ANN. Based on only descriptors or both descriptors (input variables) and biological activities (output variables) one classifies unsupervised or supervised training of ANN. Artificial Neural Network (ANN) helps in developing relations as non-linear, trends, modeling between continuous and categorized responses, among multiple responses and helps in tackling mathematical problems like data exploration, pattern recognition. To assess the goodness-of-fit of neural networks, it is suggested not only to use recall ability test but also to use leave-one-out, leave-many-out, Y-scrambling, and assessment with independent test set. In recall ability test, the activity values are calculated for the objects of training set and later provide an indication how the model recognizes the objects of training set.

As multiple regression coefficient ‘R^2’ and standard error of estimate ‘s’ helps in understanding the model fitness, even though reliable with training set, the parameters are not sufficient enough for new data. As the complexity of the model increases, even though explained variance ‘R^2’ is fitting, if the model is not well supervised, the explained prediction variance ‘Q^2’ may be found decreasing. Figure 4, illustrates a plot of Number of predictors (x-axis) vs. ‘R^2 and Q^2’ (y-axis). Upon comparison of explained variances in ‘fitting’ with ‘prediction’, it has been observed that, as the number of predictors increases, the explained variance of fitting ‘R^2’ improving. But, it has been observed that up to five numbers of predictors, explained variance of prediction ‘Q^2’ is increasing and later decreasing with the increase in the number of predictors. Hence, it was suggested that the first condition of model validity is fulfilling Topliss ratio i.e., the ratio of number of chemicals (objects) to the number of selected variable. It is recommended for a Topliss ratio of at least five. Hence, a model established should be inspected by validation techniques, able to detect over fitting due to variable multi-collinearity, noise, sample specificity, and unjustified model complexity.

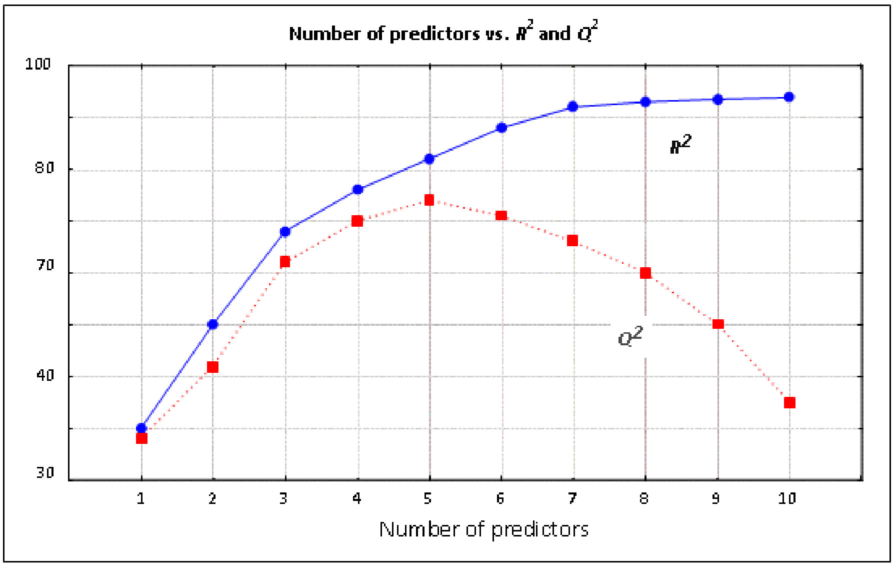

Figure 4. Comparison of explained variances of fitting with prediction

Validation of a model is done by internal and external methods. In internal validation, (6-9) the data set is modified by removing one or more objects (say chemicals) leading to a training and test set, Figure 5. The training set develops the model and the test set is used to test the model. Hence the training set and test set involves in the development of model. In case of external validation, an entirely new set of test chemicals that are not involved in the development of model are used to check the capability of the model.

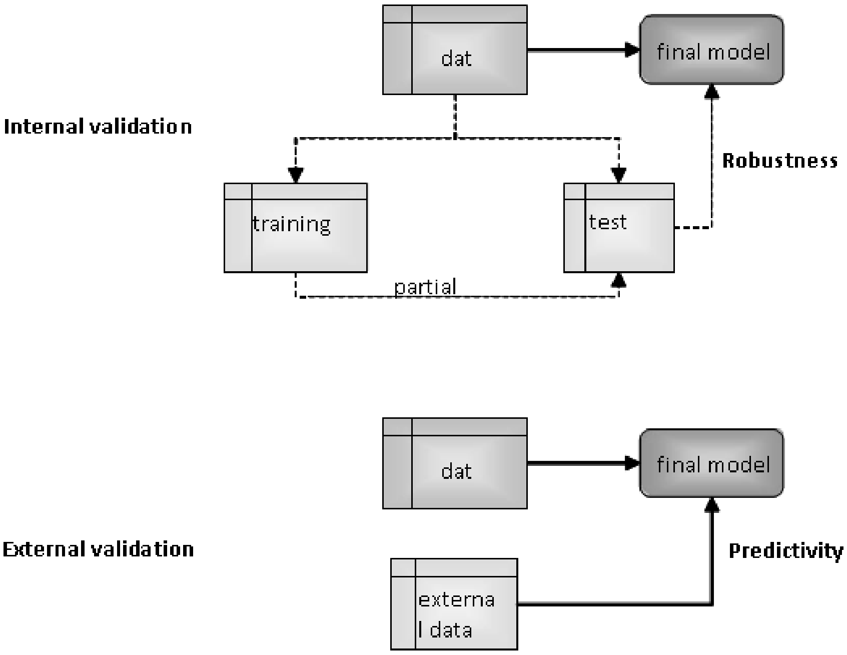

Figure 5. Internal and External Validation

Other than training/test set splitting, several other internal validation methods used to ensure the predictive ability of the model are cross validation [Leave-one-out (LOO), Leave-many-out (LMO)], bootstrapping, Y-scrambling or response permutation testing. In the cross validation method, from the original dataset, one or a group of objects are removed once and only once and the remaining reduced data set is converted into training and test set and model is developed and tested, similar to internal and external validation. The deleted object/s (chemical) response is predicted. The squared differences between the true response and the predicted response for each compound left out are added to the predictive residual sum of squares (PRESS). From the final predictive residual sum of squares, the Q^2 (or R^2_{cv}) and the standard deviation error of prediction (SDEP) are calculated. It is found to be more realistic with Leave-many-out (LMO) than Leave-one-out (LOO). In the Leave-many-out method, if number of objects are 120 ('n' chemicals) and the cancellation group 'G' ranges say 2, 3, 5, 10, then at each time "m=n/G" objects are left in the test group (i.e., 60, 40, 24, 12 chemicals respectively). It is necessary to frame rule in order to select group of chemicals for the test set keeping in mind leaving one chemical only once. The LOO method is equivalent to LMO method when G = n indicating number of cancellation groups is equal to number of compounds (chemicals/objects).

In yet another, internal validation technique i.e., Bootstrap re-sampling, data set is withdrawn from a population, representing the population. For a size of 'n', corresponding number of groups 'K' are generated. The 'n' numbers of chemicals are selected randomly from the original data set, forming a training set and the remaining as test set. Hence, when 'n' chemicals are selected randomly from original dataset, there is a possibility that same chemical being selected several times into training set or not at all being selected. The squared differences between the true response and predicted response of the test set are expressed in PRESS statistic for deriving conclusions. The procedure of grouping training set and test set from population is done thousands of times. In LMO, a high average 'Q^2', in bootstrap validation indicates good robustness, referred some times as internal predictivity.

In order to ensure robustness of the model, Y-scrambling is used as another internal validation technique. The test indicates any chance correlations, i.e., models where the independent variables are randomly correlated to the response variables. To ensure that one has to calculate R^2 or Q^2. The method, Figure 6, involves with randomly modifying the sequence of response vector 'y' by assigning to each compound a response randomly selected from the true set of responses. If the model has no chance correlation, Figure 7, then there is a significant difference in the quality of the original model and that associated with a model obtained with random responses. The procedure has to be repeated several hundred times.

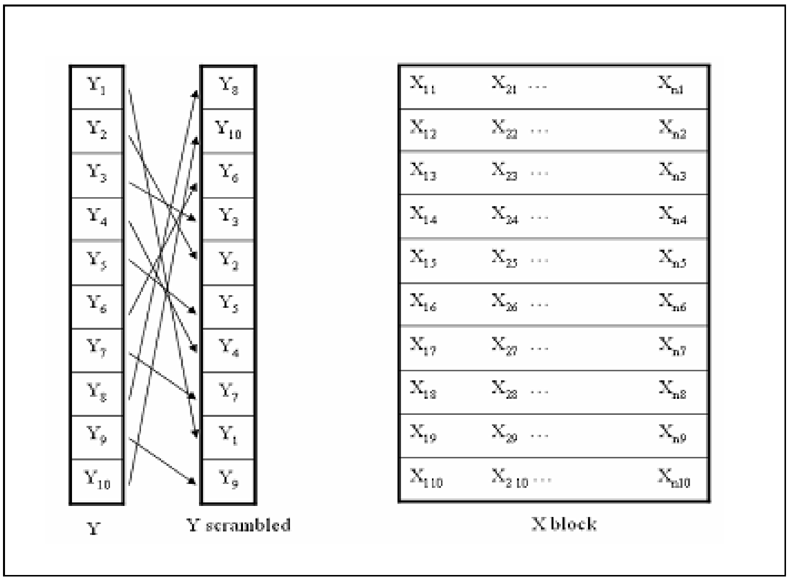

Figure 6. Y-scrambling by random permutations of activity values (Y)

QUIK rule (10) involves with a criteria that helps in making decision of rejection of a model that has high predictor collinearity, indicating chance correlation. (11) The rule is based on multivariate correlation index 'K' that measures the total correlation of set of variables. If the total correlation in the set given by model predictors 'X' plus the response 'Y (K_{xy})' i.e., [X + Y] is greater than calculated 'K_x' only in the set of predictor [X], the model is accepted. For instance, in a model, $y = -8.40 + 8.35\ \pi_N - 1.70\ \pi^2_N + 1.43\ E_s$ with calculated statistics $R^2 = 91.8$, $Q^2_{LOO} = 81.5$, $Q^2_{LMO} = 67.0$ the model was suspected for its ability to predict. Based on QUIK rule, i.e., $K_{xy} = 47.91 < K_x = 54.87$, the same conclusion was confirmed and hence rejected.

Yet another method of ensuring chance correlation of a model is by intentionally adding a certain percentage of artificial noise variables for the original variables. (12) This is found to decide the model size. When noise variables are found to appear, it is an indication that the model size no longer can be increased.

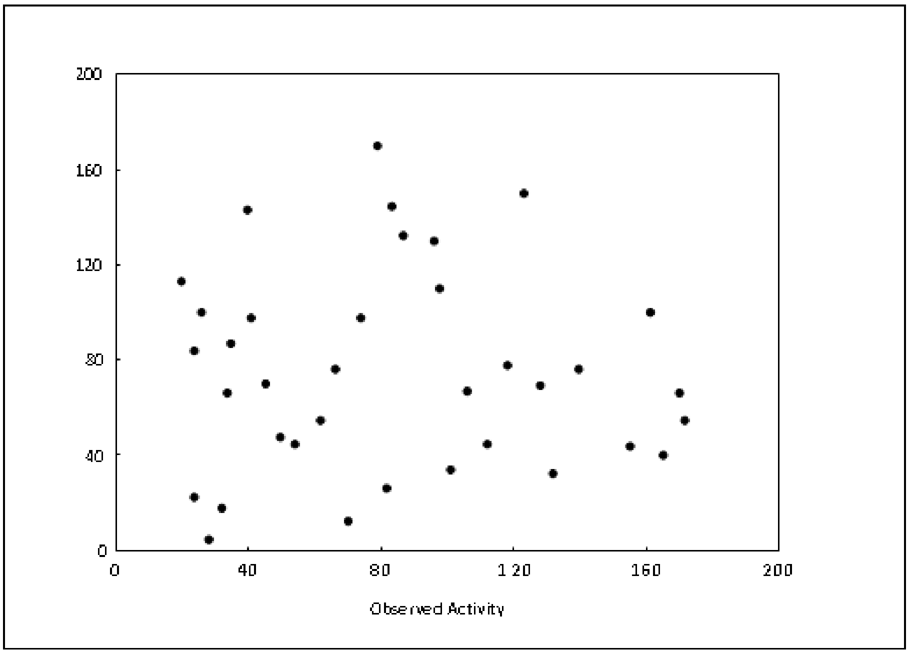

Figure 7. Plot of predicted versus observed activity values (Y) –random scatter plot indicates that the model is not probably due to chance correlations

In order to ensure the predictivity of a regression model, (13-16) one has to compare predicted and observed values and for this sufficiently large external data set which was not used in the development of the model is used. The responses are calculated and the external explained variance 'Q^2_{ext}' is calculated as the sum of the predictive residual sum of squares (of the external data set) to the reference total sum of squares (i.e., calculated by comparing predicted response of the external test chemicals with the average response of the training set). Several times it is not possible to conduct experiments on animals due to various reasons; hence the development of training set and ensuring predictivity by external data set is overcoming the issue. It was also suggested that representative points of the test sets should be closer to training set and vice versa, and additionally the training set should be diverse.

Several approaches were suggested for splitting the population of chemicals into training and test sets such as straight forward random selection, activity sampling, various systematic clustering techniques [using K-algorithms, dissimilarity (i.e., Kennard stone algorithm), hierarchal], Kohonen's self-organizing maps, formal statistical experimental design (factorial and D-optimal), modified sphere exclusion algorithm, binning range of experimental values and randomly selecting from each bin for even distribution.

In knowledge driven QSAR models, the inclusion of chemicals in classes involves with experts knowledge as well as data driven, hence called as two fold. For instance, in ECOSAR system more than 100 classes of chemicals are used and the QSAR models were developed using public and proprietary data. (17)

7.5 OECD Principle 5- Guidance on the principle of mechanistic interpretation in Detail (1)

The principle indicates that QSAR models should be validated for mechanistic interpretation, if possible. The clause 'if possible' is used because the QSAR models were developed in iterative process involving statistically collecting data, hypothesis generation and testing. The iterative process also involves with collecting descriptors, developing the model and several times refining at each and every stage. The mechanistic interpretation involves with molecular descriptors (parameters) such as intrinsic chemical interactions, hydrophobic, electronic, steric attributes, connectivity indexes, biological interactions, ionic constants, HOMO, LUMO, polar surface area, molar refractivity, polarisability, charges, van der Waals radii etc., that leads to final endpoint. For instance, several models developed are as follows:

$$\mathbf{RAI = Log\ D + a\ Log\ k + b\ Log\ P}$$

where, RAI is Relative Alkylation Index which is used for skin sensitization. Here, D, k and P are dose, relative rate constant and octanol/water partition coefficient respectively. Similarly,

$$\mathbf{Log\ (LC_{50}) = -0.846\ log\ K_{ow} - 1.39}$$

Where, LC50 is the concentration (moles/litre) causing 50% lethality in *Pimephales promelas*, after exposure to 96 hours, K_{ow} is octanol-water partition coefficient.

Developing such models involves with chemicals data from classes or expert data, proprietary data involving artificial intelligence as iterative processes.

8. Conclusion

In the initial days of forward research, experiments are empirically designed and conducted. As time proceeds, concepts, indications, indexes are developed. Complexities arise and lead to backward research so that uniformity is achieved. The current guideline is also obtained by forward and backward approaches by raising a problem and solving. Several organisations that are either private or government have developed softwares using multidisciplinary approaches. The guideline is the basis for these softwares development, which itself is tedious process. Several international guidelines indicate to use available well accepted resources and if necessary, new software has to be development to overcome the drawbacks. Several tissues, disease specific softwares are available for screening chemical entities.

Acknowledgements

The author is thankful to the management of the college where he is currently working. The author declares that there is no conflict of interest and is a part of research.

Financial Disclosure statement: The author received no specific funding for this work.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

1. France. OECD, Environment Directorate, OECD Series on Testing and Assessment, Guidance Document on the Validation of (Quantitative) Structure Activity Relationship [(Q)SAR] Models [Internet]. OECD iLibrary; 2007 [cited 2020 Oct.26]. Available from: https://www.oecd-ilibrary.org/environment/guidance-document-on-the-validation-of-quantitative-structure-activity-relationship-q-sar-models_9789264085442-en
2. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees, Wadsworth International Group, Belmont, CA, USA; 1984.
3. Hand D. Discrimination and Classification. New York: Wiley & Sons; 1981
4. Anzali S, Gasteiger J, Holzgrabe U, Polanski J, Sadowski J, Teckentrup A, Wagener M. The use of self organizing neural networks in drug design. Perspectives in Drug Discovery and Design. 1998;9–11:273–299.
5. Zupan J, Gasteiger J. Neural Networks for Chemistry and Drug Design. Weinheim, Germany: Wiley-VCH; 1999.
6. Diaconis P, Efron B. Computer Intensive Methods in Statistics. Scientific American. 1983;248:96-108.
7. Cramer RD, Bunce JD, Patterson DE, Frantk IK. Cross Validation, Bootstrapping and Partial Least Squares Compared with Multiple Regression in Conventional QSARs Studies. Quant. Struct-Act Relat. 1988;7:18-25.
8. Wehrens R, Putter H, Buydens LMC. Bootstrap: a Tutorial. Chemometrics and Intelligent Laboratory Systems. 2000;54:35-52.
9. Lindgren F, Hansem B, Karcher W, Sjostrom M, Eriksson L. Model Validation by Permutation Tests: Applications to Variable Selection. Journal of Chemometrics. 1996;10:421-532.
10. Todeschini R, Consonni V, Maiocchi A. The K Correlation Index: Theory Development and Its Applications in Chemometrics. Chemometrics and Intelligent Laboratory Systems. 1999;46:13-29.
11. Todeschini R, Consonni V, Mauri A, Pavan M. Detecting Bad Regression Models: Multicriteria Fitness Functions in Regression Analysis. Analytica Chimica Acta. 2004;515(1):199-208.
12. Jouan-Rimbaud D, Massart DL, de Noord OE. Random Correlation in Variable Selection for Multivariate Calibration with a Genetic Algorithm. Chemometrics and Intelligent Laboratory Systems. 1996;35:213-220.
13. Golbraikh A, Tropsha A. Beware of q^2! Journal of Molecular Graphics and Modelling. 2002;20:269-276.
14. Gramatica P, Pilutti P, Papa E. Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training-Test Sets and Consensus Modeling. Journal of Chemical Information and Computer Sciences. 2004;44:1794-1802.
15. Gramatica P, Papa P. An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors. QSAR and Comb. Sci. 2005;24:953-960.
16. Pavan M, Netzeva TI, Worth AP. Validation of a QSAR Model for Acute Toxicity. SAR and QSAR in Environmental Research. 2006;17:147-171.
17. Zeeman M, Auer CM, Clements RG, Nabholz JV, Boethling RS. U.S. EPA Regulatory Perspectives on the Use of QSAR for New and Existing Chemical Evaluations. SAR and QSAR in Environmental Research. 1995;3:179-201.