DIAGNOSTIC OF A QSPR MODEL: AQUEOUS SOLUBILITY OF DRUG-LIKE COMPOUNDS

SORANA D. BOLBOACĂ[a,b], LORENTZ JÄNTSCHI[b]

ABSTRACT. A diagnostic test for a qSPR (quantitative Structure-Property Relationship) model was carried out using a series of statistical indicators for correctly classifying compounds into actives and non-actives. A previously reported qSPR model, able to characterize the aqueous solubility of drug-like compounds, was used in this study. Eleven statistical indicators like those used in medical diagnostic tests were defined and applied on training, test and overall data sets. The associated 95% confidence interval under the binomial distribution assumption was also computed for each defined indicator in order to allow a correct interpretation. Similar results were obtained in the training and test sets with some exceptions. The prior probabilities of active and non-active compounds proved not to be significantly different in the training and test sets. However, the probability of classification as active compounds proved to be significantly smaller in the training set as compared to the test set (p = 0.0042). The total fraction of correctly classified compounds proved to be identical in the training and test sets as well as in the overall set. Nevertheless, the overall model and the model obtained in the test set show a higher ability to correctly assign the non-active compounds to the non-active class while the model obtained in the training set has a higher ability to correctly assign the active compounds to the active class.

Keywords: *quantitative Structure-Property Relationships (qSPR), diagnostic parameters, 2×2 contingency table, solubility, drug-like compounds*

INTRODUCTION

Quantitative structure-property relationships (qSPRs) procedures able to quantitatively correlate the chemical structure with a defined property [1], are widely used in drug design [2,3], drug classification [4,5] and screening [5,6].

A series of studies were drawn in order to establish the validation methods of a qSPR model [7,8], including the principle of parsimony, selection of the simplest model, cross-validation, Y scrambling and external predictability [9]. Various procedures for variable selection have been created [10-13] and statistical

[a] *"Iuliu Hațieganu" University of Medicine and Pharmacy Cluj-Napoca, 13 Emil Isac, RO-400023 Cluj-Napoca, Romania, sbolboaca@umfcluj.ro*

[b] *Technical University of Cluj-Napoca, 28 Memorandumului, RO-400114 Cluj-Napoca, Romania, lori@academicdirect.org*

analysis of molecular similarity matrices was developed in order to identify the best quantitative structure-activity relationships [14]. Reliability and accuracy have also been introduced for the validation of QSPR models [15,16]. The information criteria (Akaike's information criteria - AIC [17], corrected AIC [18], Schwarz (or Bayesian) Information Criterion – BIC, Amemiya Prediction Criterion – APC, and Hannan-Quinn Criterion - HQC) and Kubinyi's function [19, 20] are the parameters used to compare different qSPR/qSAR models [21-23].

The aim of this study was to carry out a diagnostic test on a qSPR (quantitative Structure-Property Relationships) model, by using a series of statistical indicators for correctly classifying compounds into actives and non-actives.

RESULTS AND DISCUSSION

Eleven statistical indicators were proposed as diagnostic parameters for qSPR models. The contingency tables used to calculate these parameters are presented in Table 1. The statistical indicators computed for the training, test and overall data sets are presented in Table 2 – 4.

Table 1. 2×2 contingency tables for the investigated qSPR model

Generic Table	Observed		
Estimated	+	-	Σ
+	TP	FP	
-	FN	TN	
Σ			n

Test Set	Observed		
Estimated	+	-	Total
+	26	10	36
-	4	29	33
Total	30	39	69

Training Set	Observed		
Estimated	+	-	Total
+	28	7	35
-	12	48	60
Total	40	55	95

Overall	Observed		
Estimated	+	-	Total
+	54	22	76
-	11	77	88
Total	65	99	164

+ = active class; - = non-active class;
Estimated = aqueous solubility estimated by Duchowitz's et al. qSPR model

The chi-squared test was applied on contingency tables in order to test the null hypotheses that the estimated class (active and non-active) is independent from the observed class (active and non-active). The value of the chi-squared statistics and associated significance level, presented at the bottom of Tables 2 - 4, supported the rejection of the null hypotheses that the estimated classification into active and non-active compounds is unrelated to the observed classification. These results sustain the ability of the qSPR model to classify compounds as actives and non-actives. The degree of association between the estimated and the observed classification of compounds proved to be a positive and moderate one, in all the investigated sets (training, test and overall set of studied compounds). The moderate association, expressed as the Φ contingency correlation coefficient, revealed that the reported qSPR [24] is not a perfect model.

Table 2. Statistical indicators for assessing the qSPR model: training set

Parameter (Abbreviation)	Value	95%CI
Concordance / Accuracy / Non-error Rate (CC/AC)	80.00	[71.07-87.02]
Error Rate (ER)	20.00	n.a.
Prior proportional probability of an active class	0.4211	[0.3254-0.5215]
Prior proportional probability of a non-active class	0.5789	n.a.
Sensitivity (Se)	70.00	[54.76-82.39]
False-negative rate (under-classification, FNR)	30.00	[17.61-45.24]
Specificity (Sp)	87.27	[76.39-93.96]
False-positive rate (over-classification, FPR)	12.73	[6.04-23.61]
Positive predictivity (PP)	80.00	[64.55-90.44]
Negative predictivity (NP)	80.00	[68.52-88.49]
Probability of classification		
- as active (PCA)	0.3684	[0.2766-0.4682]
- as non-active (PCIC)	0.6316	[0.5318-0.7234]
Probability of a wrong classification		
- as active compound (PWCA)	0.2000	[0.0956-0.3545]
- as non-active compound (PWCI)	0.2000	[0.1151-0.3148]
Odds Ratio (OR)	16.0000	[5.7090-45.0262]

95% CI = confidence interval at a significance level of 5%; n.a. = not available;
χ^2 = 30.2305 ($p < 0.0001$) (Chi-squared statistics); Contingency correlation coefficient Φ = 0.5641

Table 3. Statistical indicators for assessing the qSPR model: test set

Parameter (Abbreviation)	Value	95%CI
Concordance / Accuracy / Non-error Rate (CC/AC)	79.71	[69.04-87.79]
Error Rate (ER)	20.29	n.a.
Prior proportional probability of an active class	0.4348	[0.3225-0.5524]
Prior proportional probability of a non-active class	0.5652	n.a.
Sensitivity (Se)	86.67	[70.96-95.08]
False-negative rate (under-classification, FNR)	13.33	[4.92-29.04]
Specificity (Sp)	74.36	[59.21-85.91]
False-positive rate (over-classification, FPR)	25.64	[14.09-40.79]
Positive predictivity (PP)	72.22	[56.25-84.67]
Negative predictivity (NP)	87.88	[73.29-95.52]
Probability of classification		
- as active (PCA)	0.5217	[0.4050-0.6367]
- as non-active (PCIC)	0.4783	[0.3633-0.5950]
Probability of a wrong classification		
- as active compound (PWCA)	0.2778	[0.1533-0.4375]
- as non-active compound (PWCI)	0.1212	[0.0448-0.2671]
Odds Ratio (OR)	18.8500	[5.4919-64.5994]

95% CI = confidence interval at a significance level of 5%; n.a. = not available;
χ^2 = 22.9206 ($p < 0.0001$) (Chi-squared statistics); Contingency correlation coefficient Φ = 0.5764

The accuracy of the qSPR model proved to be almost 80% in all the investigated sets of compounds. The accuracy of the model in the training set proved not to be statistically different from the accuracy of the model in the test set (the confidence intervals overlap, see Tables 2 and 3). A similar interpretation is true when the values and associated confidence intervals of other statistical indicators are analyzed (see Tables 2 -4).

Table 4. Statistical indicators for assessing the qSPR model: overall set

Parameter (Abbreviation)	Value	95%CI
Concordance / Accuracy / Non-error Rate (CC/AC)	79.88	[73.22-85.43]
Error Rate (ER)	20.12	n.a.
Prior proportional probability of an active class	0.3963	[0.3238-0.4725]
Prior proportional probability of an non-active class	0.6037	n.a.
Sensitivity (Se)	83.08	[72.50-90.55]
False-negative rate (under-classification, FNR)	16.92	[9.45-27.50]
Specificity (Sp)	77.78	[68.82-85.05]
False-positive rate (over-classification, FPR)	22.22	[14.95-31.18]
Positive predictivity (PP)	71.05	[60.19-80.30]
Negative predictivity (NP)	87.50	[79.26-93.06]
Probability of classification		
- as active (PCA)	0.4634	[0.3883-0.5398]
- as non-active (PCIC)	0.5366	[0.4602-0.6117]
Probability of a wrong classification		
- as active compound (PWCA)	0.2895	[0.1970-0.3981]
- as non-active compound (PWCI)	0.1250	[0.0694-0.2074]
Odds Ratio (OR)	17.1818	[7.7989-38.1475]

95% CI = confidence interval at a significance level of 5%; n.a. = not available
χ^2 = 83.6385 (p < 0.0001) (Chi-squared statistics); Contingency correlation coefficient Φ = 0.5761

The Z test was applied in order to compare the statistical indicators expressed as probabilities obtained in training and test sets. The prior probabilities of active and non-active compounds proved not to be statistically different in training and test sets. The absence of statistically significant differences between prior probabilities of active and non-active compounds in training and test sets supports the correct assignment of compounds to the active/non-active sets. However, the probability of classification as active compounds proved to be statistically smaller in the training set compared to the test set (p=0.0042); thus, the classification model proved to perform better in terms of correct classification of active compounds when applied on test set.

The objective of this study is to propose a series of statistical indicators as diagnostic tools for the qSPR model. In achieving this, various aspects are considered:

- Analyzing the correct assignment of compounds to training and test sets: prior proportional probability of an active class & prior proportional probability of a non-active class
- Analyzing the correct classification of active and non-active compounds: all the other statistical indicators (see Table 2-4).

The proposed statistical indicators have to assess the qSPR model in training and test sets: as the indicators have similar performances in training and test sets, it could involve the model has similar classification abilities, thus being considered as a good model. The best model is the one with the highest possible accuracy and the smallest possible error rate. The best model is also the one with the highest sensitivity and specificity and the smallest false-negative and false-positive rates. In this respect, it can be observed that sensitivity is smaller than specificity in the training set while sensitivity is higher than specificity in the test set (see Tables 2 and 3). In other words, the investigated qSPR model has a higher ability to correctly assign active compounds to the active class in the test set and a higher ability to correctly assign non-active compounds to the non-active class in the training set. An excellent classification model should also have the best possible positive and negative predictability values while the probability values of a wrong classification into active and non-active compounds should have the smallest possible values.

Similar statistical parameters are used to assess the performances of machine learning classification models: accuracy, recall (true positive rate, false positive rate, true negative rate, false negative rate, and precision) [25, 26]. These parameters are calculated based on the confusion matrix [27]. Note that the confusion matrix is the same as the generic contingency table presented in Table 1.

The present study is aimed to introduce a series of statistical indicators in order to diagnose a qSPR model. Useful information related to the assignment of compounds in the training and test sets could be obtained by using prior proportional probability of an active class & prior proportional probability of a non-active class. All the other proposed statistical indicators allow the characterization of a qSPR model in terms of total fraction of correctly classified compounds (accuracy), correct assignment to active or non-active class (sensitivity and specificity, false positive and false negative rates), etc. Statistical indicators were applied on a 2×2 confusion matrix but the same approach could also be applied on r×c confusion matrices when compounds are classified into more than two groups (e.g., non-active, active, and very active). The usefulness of this approach in diagnosing qSPR/qSAR models is currently investigated in our laboratory.

CONCLUSIONS

The total fraction of compounds correctly classified by the qSPR model proved to be identical in the training and test sets as well as in the overall set. However, the overall model and the model obtained in the test set showed a higher ability to correctly assign the non-active compounds to the negative class while the model obtained in the training set had a higher ability to correctly assign the active compounds to the active class.

EXPERIMENTAL SECTION

A previously reported qSPR model [28] able to characterize the aqueous solubility of drug-like compounds was herein used. The experimental aqueous solubility measured at 298K and expressed in mg/ml (values taken from Merck Index 13th [28]) was modeled using molecular descriptors [24].

The best model obtained in the training set (n=97) proved to be a model with 3 descriptors and the following characteristics [24]:

$R^2 = 0.871$; $S = 0.903$
$R^2_{loo} = 0.849$; $S_{loo} = 0.971$
$R^2_{val} = 0.848$; $S_{val} = 0.899$

where R^2 = determination coefficient; S = standard deviation of the model; R^2_{loo} = determination coefficient on leave one out analysis; S_{loo} = standard deviation on leave-one-out analysis; R^2_{val} = determination coefficient on validation set; S_{val} = standard deviation on validation set.

A series of statistical indicators similar with those used in medical diagnostic tests [29, 30] were defined as diagnostic parameters for the qSPR model (Table 5).

The experimental and estimated aqueous solubility of the studied compounds was transformed as dichotomial variables in order to calculate the defined statistical indicators (Table 5) using the following criteria: if experimental data ≥ 0, the compound was considered active, if experimental data < 0, the compound was considered non-active.

Table 5. Statistical indicators calculated on the 2×2 contingency table

Indicator (Abbreviation)	Formula	Definition
Accuracy / Non-error Rate (AC)	100*(TP+TN)/n	Total fraction of correctly classified compounds
Error Rate (ER)	100* (FP+FN)/n = 1-CC	Total fraction of misclassified compounds
Prior proportional probability of a class (PPP)	n_i/n	Fraction of compounds belonging to class *i*
Sensitivity (Se)	100*TP/(TP+FN)	Percentage of active compounds correctly assigned to the active class
False-negative rate (under-classification, FNR)	100*FN/(TP+FN) = 1-Se	Percentage of active compounds falsely assigned to the non-active class
Specificity (Sp)	100*TN/(TN+FP)	Percentage of non-active compounds correctly assigned to the non-active class
False-positive rate (over-classification, FPR)	100*FP/(FP+TN) = 1-Sp	Percentage of non-active compounds falsely assigned to the active class
Positive predictivity (PP)	100*TP/(TP+FP)	Percentage of compounds correctly assigned to the active class out of all compounds assigned to the active class
Negative predictivity (NP)	100*TN/(TN+FN)	Percentage of compounds correctly assigned to the non-active class out of all compounds assigned to the non-active class
Indicator (Abbreviation)	**Formula**	**Definition**
Probability of classification - as active (PCA)	(TP+FP)/n	- Probability to classify a compound as active (true positive & false

- as inactive (PCIC)	(FN+TN)/n	positive) - Probability to classify a compound as non-active (true negative & false negative)
Probability of a wrong classification - as active compound (PWCA) - as non-active compound (PWCI)	FP/(FP+TP) FN/(FN+TN)	Probability of a false positive classification Probability of a false negative classification
Odds Ratio (OR)	(TP*TN)/(FP*FN)	The odds of correct classification in the group of active compounds divided to the odds of an incorrect classification in the group of non-active compounds

The associated 95% confidence interval under the binomial distribution assumption [31] was also computed for the correct interpretation of the indicators [32].

ACKNOWLEDGMENTS

Financial support is gratefully acknowledged to CNCSIS-UEFISCSU Romania (project PNII-IDEI458/206/2007).

REFERENCES

1. L.P. Hammett, *Chemical reviews,* **1935**, *17*, 125.
2. I.M. Kapetanovic, *Chemico-Biological Interactions,* **2008**, *171(2)*, 165.
3. C.H. Andrade, K.F. Pasqualoto, E.I. Ferreira, A.J. Hopfinger, *Molecules,* **2010**, *15(5)*, 3281.
4. V. Potemkin, M. Grishina, *Drug Discov Today,* **2008**, *13(21-22)*, 952.
5. J. Li, P. Gramatica, *Journal of Chemical Information and Modeling,* **2010**, *50(5)*, 861.
6. M.C. Hutter, *Current Medicinal Chemistry,* **2009**, *16(2)*, 189.
7. M. Pavan, T.I. Netzeva, A.P. Worth, *SAR and QSAR in Environmental Research,* **2006**, *17(2)*, 147.
8. S. Wold, *Quantitaive Structure-Activity Relationship,* **1991**, *10*, 191.
9. R.D. Cramer III, J.D. Bunce, D.E. Patterson, I.E. Frank, *Quantitaive Structure-Activity Relationship,* **1988**, *7*, 18; Erratum **1988**, *7*, 91.
10. H. Kubinyi, *Quantitaive Structure-Activity Relationship,* **1994**, *13*, 285.
11. H. Kubinyi, *Quantitaive Structure-Activity Relationship,* **1994**, *13*, 393.
12. K. Héberger, *TrAC - Trends in Analytical Chemistry,* **2010**, *29(1)*, 101.
13. P. Ghosh, M.C. Bagchi, *Current Medicinal Chemistry,* **2009**, *16(30)*, 4032.
14. A.C. Good, S.J. Peterson, W.G. Richards, *Journal of Medicinal Chemistry,* **1993**, *36(20)*, 2929.

15. B.H. Su, M.Y. Shen, E.X. Esposito, A.J. Hopfinger, Y.J. Tseng, *Journal of Chemical Information and Modeling,* **2010**, *50(7)*, 1304.
16. L.G. Valerio Jr., *Toxicology and Applied Pharmacology,* **2009**, *241(3)*, 356.
17. H. Akaike, *Annals of the Institute of Statistical Mathematics,* **1969**, *21*, 243.
18. C.M. Hurvich, C. Tsai, *Biometrika,* **1989**, *76*, 297.
19. H. Kubinyi, *Quantitative Structure-Activity Relationships,* **1994**, *3*, 393.
20. H. Kubinyi, *Quantitative Structure-Activity Relationships,* **1994**, *13*, 285.
21. S.D. Bolboacă, L. Jäntschi, *TheScientificWorldJOURNAL*, **2009**, *9(10)*, 1148.
22. S.D. Bolboacă, M.M. Marta, C.E. Stoenoiu, L. Jäntschi, *Applied Medical Informatics*, **2009**, *25(3-4)*, 65.
23. S.D. Bolboacă, M.M. Marta, L. Jäntschi, *Folia Medica,* **2010**, *52(3)*, 37.
24. P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, *Bioorganic & Medicinal Chemistry,* **2008**, 16(17), 7944.
25. A. Lombardo, A. Roncaglioni, E. Boriani, C. Milan, E. Benfenati, *Chemistry Central Journal,* **2010**, *4 Suppl 1*, S1.
26. N. Fjodorova, M. Vracko, M. Novic, A. Roncaglioni, E. Benfenati, *Chemistry Central Journal,* **2010**, Jul 29; 4 Suppl 1:S3.
27. M. Kubat, S. Matwin, Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186, Nashville, Tennesse. Morgan Kaufmann, **1997**.
28. The Merck Index An Encyclopedia of Chemicals, Drugs, and Biologicals; Merck & Co.: NJ, **2001**.
29. S.D. Bolboacă, L. Jäntschi *Electronic Journal of Biomedicine,* **2007**, *2*, 19-28.
30. S.D. Bolboacă, L. Jäntschi, A. Achimaş Cadariu, *Applied Medical Informatics,* **2004**, *14*, 27.
31. S.D. Bolboacă, L. Jäntschi, *International Journal of Pure and Applied Mathematics*, **2008**, *47(1)*, 1.
32. L. Jäntschi, S.D. Bolboacă, *TheScientificWorldJOURNAL*, **2010**, *10*, 865.