Abstract # 2273b

Benchmarking Assessment of Open Source and Newly Released *Salmonella* Mutagenicity (Q)SAR Models for Potential Use Under ICH M7

Lidiya Stavitskaya, Barbara L. Minnier, and Naomi L. Kruhlak

FDA Center for Drug Evaluation and Research (CDER), 10903 New Hampshire Ave., Silver Spring, MD 20993

[The findings and conclusions in this presentation have not been formally disseminated by the FDA and should not be construed to represent any agency determination or policy.]

ABSTRACT

The current draft of the International Conference on Harmonisation (ICH) M7 guideline describes the use of (quantitative) structure-activity relationship ((Q)SAR) models during drug safety evaluation. The guideline specifies that an expert rule-based (SAR) methodology and a statistical-based (QSAR) methodology should be applied and the results should be reviewed with expert knowledge to provide additional supportive evidence on the relevance of the prediction. The guideline, however, does not specify the use of any particular model, but instead recommends that the models meet the general definition of statistical or rule-based methodologies, and allow the identification of structural alerts. In our previous studies, we reported the construction of QSAR models using commercial software with prediction accuracy ranging from 82% to 84% in cross-validation and 73% to 76% in external validation. More importantly, when the models were applied in combination, sensitivity of 91% and negative predictivity of 79% was achieved, which are key parameters in the use of these models under ICH M7. In this study, we evaluated the performance of two freely-available, open source (Q)SAR programs, Toxtree and OECD Toolbox and three newly-released, commercial (Q)SAR programs, Leadscope Expert Alert System, Sarah Nexus, and Symmetry as potential candidates for qualifying pharmaceutical impurities. To effectively assess their performance, an in-house Salmonella mutagenicity database of 3979 compounds and a highly-curated version of the Hansen dataset of 3700 compounds were used for benchmarking. These sets are comprised of drug-like and industrial chemical examples containing a variety of functional groups as well as less-common atoms. These assessments will improve and facilitate the ability of FDA/CDER to interpret the quality and reliability of *in silico* data submitted under ICH M7 for the qualification of pharmaceutical.

INTRODUCTION

Regulatory Context

- The Salmonella mutagenicity assay is used in the drug review process to assess the genotoxic potential of APIs, as well as impurities, metabolites and degradants.
- More recently, under the ICH M7 guideline draft (step 2) [1], (Q)SAR data may be submitted by sponsors in place of conventional in vitro bacterial (*Salmonella* and *E. coli*) mutagenicity assay data to qualify pharmaceutical impurities for their genotoxic potential.
- The guideline specifies that an expert rule-based methodology and a statistical-based methodology should be applied and the results of the analyses should be reviewed with expert knowledge in order to provide additional supportive evidence on the relevance of any positive or negative prediction.
- The guideline does not discriminate between the use of commercially available models or in-house proprietary models for (Q)SAR analyses.
- All (Q)SAR submissions must provide sufficient supporting documentation to assure regulators that the analyses were adequately performed and appropriately interpreted.

External Validation Dataset

- Predictive performance of a model should ideally be assessed using a combination of cross-validation, external validation, and y-scrambling techniques [2].
- An appropriate external validation set should be applied to commercially available and in-house proprietary models to assess relative performance, as well as performance in combination.
- An external validation set for Salmonella mutagenicity should be representative of a broad range of mutagenic mechanisms, as well as be balanced in its ratio of chemicals that have been shown in laboratory testing to be mutagenic (positives) to those that have been shown to be non-mutagenic (negatives).
- The use of multiple (Q)SAR models in combination has been shown [3, 4] to increase sensitivity over any one model alone. It has been shown that toxicophore performance differs across QSAR models and rule-based predictions, thus confirming the ability to have complementarity among selected approaches on a global level.

OBJECTIVE

To evaluate the performance and interpretability of two freely-available, open source (Q)SAR models and three newly-released commercial (Q)SAR programs in order to improve the ability of FDA/CDER to interpret the quality and reliability of *in silico* data submitted under ICH M7 for the qualification of pharmaceutical impurities.

MATERIALS AND METHODS

(Q)SAR Software

Performance of the following (Q)SAR software programs was evaluated: Toxtree v2.6.0, Leadscope Expert Alert System (LEAS) v1.0.0, Sarah Nexus v1.0.0, and Prous Symmetry 1.4.1.

- Toxtree is a open-source expert rule-based application that uses a decision tree approach to estimate toxic hazards [5] based on rules for mutagenicity and carcinogenicity developed by Benigni and Bossa [6].
- Leadscope Expert Alert System is a new, commercial expert rule-based system that predicts the results of a bacterial mutagenicity assay by matching a query molecule to alerts derived from the published literature with consideration of mitigating structural factors [7].
- Sarah Nexus is a new, commercial, statistical QSAR methodology developed specifically for prediction of *Salmonella* mutagenicity based on Lhasa's comprehensive, curated regulatory and public mutagenicity datasets for the prediction of mutagenicity from chemical structure.
- OECD Toolbox is a freely-available system that groups chemicals into categories to fill data gaps for a query molecule through the application of read-across, trend, and (Q)SAR analyses.
- Symmetry is new, commercial software that contains a QSAR model to predict the outcome of the Salmonella mutagenicity assay from a training set of approximately 7,300 chemicals [8].

External Validation Sets

Two independently derived datasets were used:

1) Hansen dataset [9]

2) FDA/CDER in-house dataset

- A binary scoring system was applied to encode activity for each chemical in each data set: positive = 1, negative = 0.
- The Hansen set is comprised of non-proprietary data collected by Hansen et al. from the scientific literature and made available as an SD file accompanying the published manuscript. The entire set contained 6512 compounds; however, 2680 were in the internal data set or were stereo or geometric isomers of structures already in the training set. A further 132 were perceived duplicates within the test set or unmodelable structures, which were removed, leaving a total of 3700 compounds in the final test set.
- The FDA/CDER in-house set is composed of non-proprietary data harvested from publicly available FDA approval packages and published literature. The entire set contained 3979 compounds.
- Both test sets were run against the selected (Q)SAR models and the overall performance was calculated using Cooper statistics.

➢ ***External validation studies were carried out using slightly different variations of the two test sets and therefore are not directly comparable.***.

RESULTS AND DISCUSSION

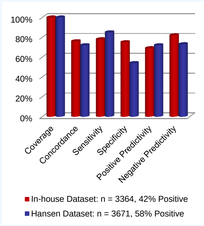
Toxtree Evaluation

- Toxtree is an expert-rule based platform which uses expert rules developed by Benigni and Bossa [6] to identify structural alerts that may be present in a test chemical.
- Figure 1. shows an example of a prediction obtained from Toxtree. Although the details of an alert, to our knowledge, are not displayed within Toxtree, they can be easily obtained from numerous Benigni and Bossa references that are displayed during model selection process.

Figure 1. Prediction Scheme

- The performance of Toxtree was assessed using the FDA/CDER in-house *Salmonella* dataset after removal of 615 compounds that overlapped with the training set and the Hansen dataset after removal of 29 compounds that also overlapped with the training set.
- The resulting in-house dataset comprised of 3364 chemicals (42% positive) yielded sensitivity and negative predictivity of 78% and 82%, respectively (shown in Table 1).
- The resulting Hansen set was comprised of 3671 chemicals (58% positive), and yielded sensitivity and negative predictivity of 85% and 73%, respectively (shown in Table 1).

Table 1. External Validation Performance of Toxtree

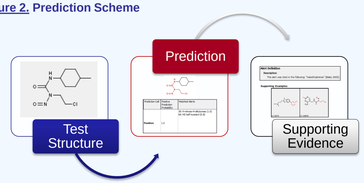


➢ ***Overall, Toxtree shows high sensitivity and sufficient transparency and interpretability for use under the ICH M7***.

Leadscope Expert Alert System Evaluation

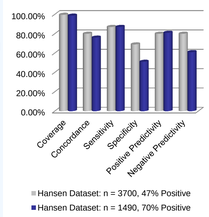
- Leadscope Expert Alert System (LEAS) is an expert rule-based platform which applies SAR rules derived from the published literature and implemented by Leadscope Inc. to identify structural alerts that are present in a test chemical.
- An example of a prediction in LEAS is displayed in Figure 2. Supporting evidence includes references and supporting examples from reference set.

Figure 2. Prediction Scheme



- The performance of LEAS was assessed using the Hansen dataset comprised of 3700 chemicals (47% positive).
- In this evaluation 85% sensitivity and 78% negative predictivity were obtained as shown in Table 2.
- In a subsequent evaluation, we removed the 2210 chemicals which were present in the model's reference set and the resulting Hansen set, comprised of 1490 chemicals (70% positive--highly skewed), yielded sensitivity and negative predictivity of 86% and 61%, respectively (shown in Table 2). This evaluation illustrates the performance statistics for chemicals that are not in the model's reference set.

Table 2. External Validation Performance of LEAS



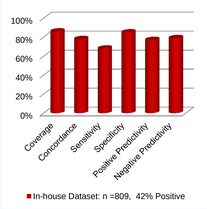
➢ ***Overall, the expert alert system showed high sensitivity and provided adequate supporting evidence to understand a given prediction, making it suitable for application under ICH M7. Additionally, it should be noted that this expert system generates negative predictions using a reference set of 7112 number of chemicals.***

Sarah Nexus Evaluation

- Sarah Nexus is a fragment based statistical methodology which uses a self-organizing hypotheses network (SOHN) approach to facilitate the identification of potential genotoxic impurities [10].
- Figure 3 shows an example of a prediction obtained from the Sarah Nexus application. Each alert contains a large number of supporting examples and each prediction is assigned a confidence score.
- For high sensitivity, optimal settings were determined to be 10% equivocal and 10% sensitivity. The equivocal setting adjusts the threshold below which there is a lack of strong evidence to support a confident prediction. The sensitivity threshold increase the number of true positive predictions at the slight expense of increasing false positives without significantly compromising overall accuracy.

Figure 3. Prediction Scheme

Table 3. External Validation Performance of Sarah Nexus



- Performance of Sarah Nexus was assessed using the FDA/CDER in-house *Salmonella* dataset.
- A total of 3170 compounds that overlapped with the training set were removed resulting in an external validation comprised of 809 chemicals (42% positive).
- Table 3 shows performance for the optimal settings with sensitivity and negative predictivity of 68% and 79%, respectively.

➢ ***Overall, Sarah Nexus showed comparable performance and favorable prediction interpretability consistent with ICH M7***

OECD Toolbox Evaluation

- OECD Toolbox is a system that groups chemicals into categories to fill data gaps for a query molecule through the application of read-across, trend, and (Q)SAR analyses.
- This application requires in-depth training and expert knowledge to allow the conduct and interpretation of an assessment. Furthermore, multiple layers of user input and parameterization are required at different stages of the analysis, resulting in significant variability in predictions obtained by different users.

➢ ***Insufficient training was available for a full evaluation of this software; however, a preliminary analysis indicates a lack of standardization of parameters and reproducibility of predictions, which have challenging regulatory implications under ICH M7.***

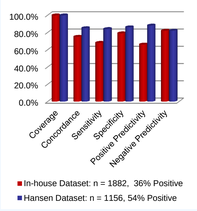
Symmetry Evaluation

- Symmetry is a molecular descriptor-based methodology, which uses either logistic regression or distance-weighted k-nearest neighbor predictive algorithms to predict genotoxic potential.
- Figure 4 shows an example of a prediction obtained from Symmetry. The output identified the algorithm used to make a prediction, the confidence score for a prediction, and one example based on the similarity of the molecular descriptors relevant to the model.

Figure 4. Prediction Scheme

- The performance was assessed using the FDA/CDER in-house *Salmonella* dataset, after removal of 2096 compounds that overlapped with the training set, and the Hansen dataset, after removal of 2578 compounds that overlapped with the training set. The resulting in-house dataset comprised 1882 chemicals (36% positive--highly skewed) and yielded sensitivity and negative predictivity of 68% and 82%, respectively. The resulting Hansen set was comprised of 1156 chemicals (54% positive) and yielded sensitivity and negative predictivity of 84% and 82%, respectively.

Table 4. External Validation Performance of Symmetry



- Model parameters are sufficiently standardized within the software that predictions obtained with this model are reproducible by different users.
- Limited information about structurally similar analogs in the training set supporting a positive or negative prediction is currently provided by the software.
- The use of calculated molecular descriptors in the model makes interpretation of a positive prediction from a structural alert standpoint challenging.

➢ ***Overall, Symmetry showed good performance but does not allow for prediction interpretation consistent with ICH M7. However, Prous Institute has informed us that the structural interpretability functionality will be incorporated in the upcoming release***.

CONCLUSIONS

- The overall performance of Toxtree, Leadscope Expert Alert System, and Sarah Nexus compare favorably to the most widely-used commercial model systems tested previously with the same data sets, and provide sufficient transparency and interpretability for the qualification of pharmaceutical impurities under ICH M7.
- Additionally, the performance of Symmetry was comparable to all the software evaluated here and in previous efforts, however, the predictions provided insufficient supporting evidence for the structural alert-based expert analysis required under ICH M7.
- OECD Toolbox could not be fully evaluated due to the complexity involved in making a prediction, which may also potentially leads to significant variability in predictions obtained by different users.
- The complementarity assessment for the selected QSAR models and rule-based models s currently underway.

ACKNOWLEDGEMENTS

Leadscope Expert Alert System, Sarah Nexus, and Symmetry software were used under Research Collaboration Agreements between FDA/CDER and Leadscope Inc., Lhasa Limited, and Prous Institute for Biomedical Research, respectively.

REFERENCES

[1] ICH M7 Draft Consensus Guideline – Assessment and control of DNA reactive impurities in pharmaceuticals to limit potential carcinogenic risk (2013).
[2] OECD (2007b). Guidance document on the validation of (Quantitative) structure activity relationships [(Q)SAR] models.
[3] Kruhlak N.L. et al., Poster Abstract # 791, Society of Toxicology Annual Meeting, San Francisco, CA. 2012.
[4] Hillebrecht A. et al., Chem. Res. Toxicol. 2011, 24:843-854.
[5] Contrera, J. (2013). Validation of Toxtree and SciQSAR *in silico* predictive software using a publicly available benchmark mutagenicity database and their applicability for the qualification of impurities in pharmaceuticals. Regulatory Toxicology and Pharmacology 67:285-293
[6] Benigni R., Bossa, C. (2011) Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology. J. Chem. Rev. 111:2507-2536.
[7] A new ICH M7 compliant expert alert system to predict the mutagenic potential of impurities, Leadscope® Genetox Expert Alerts White paper, March 2014, http://www.leadscope.com/white_papers/ICHM7-WhitePaper-0314.pdf.
[8] Valencia, A., Prous, J., Mora, O., Sadrieh, N., Valerio, L. Jr., A novel QSAR model of Salmonella mutagenicity and its application in the safety assessment of drug impurities; Toxicology and Applied Pharmacology, 273: 427–434.
[9] Hansen K. et al., J. Chem. Inf. Model. 2009, 49:2077-2081.
[10] Self Organising Hypothesis Networks: A new approach for representing and structuring SAR knowledge. T. Hanser et. al., J. Cheminformatics, 2014, in press.