Structure-based QSAR Models to Predict Systemic Toxicity Points of Departure

Prachi Pradeep[1,2], Richard Judson[2]

[1]ORISE, Oak Ridge, TN, USA
[2]NCCT, ORD, US EPA, NC, USA

Human health risk assessment associated with environmental chemical exposure is limited by the tens of thousands of chemicals with little or no experimental in vivo toxicity data. Data gap filling techniques, such as quantitative structure activity relationship (QSAR) models based on chemical structure information, are commonly used to predict hazard in the absence of experimental data. This study presents a set of QSAR models developed for chronic or sub-chronic in vivo points of departure (POD, the point on the dose-response that marks the beginning of a low-dose extrapolation). The in vivo data is taken from the EPA's ToxValDB, a compilation of data on ~3000 unique chemicals from a variety of public data sources. Using PubChem fingerprints and Chemistry Development Kit (CDK) descriptors as physchem descriptors (with feature selection), and support vector machines, random forests, K-nearest neighbor and gradient boosting regressor as machine learning algorithms (with hyper-parameter tuning) models were developed and evaluated using 5-fold internal cross-validation and external test validation. Quantitative POD models were developed for mouse (538 chemicals) and rat (811 chemicals), using point estimates for the experimental POD values. The best mouse model had an external test root mean square error (RMSE) = 0.86 and R^2 = 0.36. The best rat model had an external test RMSE = 0.95 and R^2 = 0.18. Since the training data for both mouse and rat models was skewed, they were re-constructed by creating bootstrap samples with 10% duplicate data (randomly selected from the long tail), and the models were re-built. Re-constructing the datasets to reduce the skewness in original data did not result in significantly improved models. A second set of models were built that accounted for the known lab to lab variability in the POD values. To do this, a POD distribution was constructed for each chemical using mean = median experimental POD value and standard deviation = 0.5 log-units, based on the typical lab to lab variability. Bootstrap models were built with random sampling of values from the pre-generated POD distribution to derive point estimates of POD values and confidence intervals for each prediction. The best mouse model had an average external test RMSE = 0.96 and R^2 = -0.31. The best rat model had an average external test RMSE = 1.15 and R^2 = -0.02. All units are $\log_{10}$ mg/kg/day. These models will inform chemical screening and prioritization efforts.

This abstract does not necessarily represent U.S. EPA policy.