17th International Conference on QSAR in Environmental and Health Sciences, Miami, Fl, June 13-17, 2016

Proposed abstract for Oral Presentation in Track 5: Using (Q)SARs as part of hazard and risk assessment

(Submission Deadline: November 1, 2015)

Title: The importance of data curation on QSAR Modeling: EPISuite data as a case study.

Authors: Kamel Mansouri[1], Christopher M Grulke[2], Ann M Richard[3], Richard Judson[3], Antony J Williams[3]

[1] ORISE Postdoctoral Fellow, US EPA, Research Triangle Park, NC 27711
[2] Lockheed Martin – Contractor to the US EPA, Research Triangle Park, NC 27711
[3] National Center for Computational Toxicology, US EPA, Research Triangle Park, NC 27711

Many QSAR models and tools developed at the US EPA, such as the widely used EPISuite, go back a few decades ago. Since then, the arsenal of computational capabilities supporting cheminformatics has broadened dramatically with multiple software packages. These modern tools implement advanced techniques for chemical structure representation and storage, as well as automated data-mining and standardization approaches to examine and fix data quality issues.

This presentation will highlight how data curation impacts the reliability of QSAR models. As part of this work we disentangled the influence of data quality versus quantity in the Syracuse PHYSPROP database partly used by EPISuite software. We examined key datasets related to EPISuite to validate across chemical structure representations (e.g., mol file and SMILES) and identifiers (chemical names and registry numbers), and approaches to standardize data into QSAR-ready formats prior to modeling procedures. This allowed us to quantify and segregate data into quality categories. This improved our ability to evaluate the resulting models that can be developed from these data slices, and to quantify to what extent efforts developing high-quality datasets have the expected pay-off in terms of predicting performance. The most accurate models that we build will be accessible via our public-facing platform and will be used for screening and prioritizing chemicals for further testing. This abstract does not reflect U.S. EPA policy.