2019 ACMT Annual Scientific Meeting
MedTox Shark Tank Research Forum

Principal Investigator: Daniel Nogee, MD
Research Mentor: Anthony Tomassoni, MD, MS
Co-investigator: Adrian Haimovich, MD, PhD

Title of Presentation:

Machine learning enhanced diagnosis of toxic exposures

Aims:

1) Develop and benchmark machine learning models to assist with diagnosis of toxic exposures using large toxicology data sets
2) Compare accuracy of machine learning models to a previously created software tool for toxic exposure diagnosis
3) Develop a software platform for prospective validation and distribution of machine learning models to assist diagnosis of toxic exposures

Significance and Innovation:

Toxicology may be an underrepresented discipline in many training programs and practices where esoteric toxicants are rarely encountered. Access to poison center (PC) resources, including medical toxicology consultations, is an important resource available to treating clinicians faced with challenging toxicology cases. In order to access this resource, clinicians must suspect a toxic exposure in the differential diagnosis when faced with nebulous cases and have enough confidence in their differential to call a consult and subject their diagnostic skills to review. A decision support tool may help non- toxicologists expand or increase confidence in their differentials, facilitating early toxicology consultation for improved reporting, care and outcomes. Diagnosis of causative agents may also depend on laboratory testing of clinical samples which may be accomplished too late to implement effective clinical treatment.

Machine learning, a general term for a variety of computational approaches for solving complex statistical problems, has been applied to analytical toxicology[1] and multiple other clinical fields[2-3], but has not yet made a large impact on clinical toxicology.

Through application of machine learning algorithms to large-scale toxicology data sets, we plan to create models that can correctly identify the causative agent based primarily on early clinical presentation features (history of present illness, vital signs, symptoms, etc.). With further refinement and incorporation into software platforms, we can set the stage for prospective validation of these machine learning models, with the eventual goal of creating a software tool readily available to clinicians that can accurately identify potential toxic exposures and lead to more rapid clinical diagnosis, facilitate outbreak recognition and promote effective treatment. Eventually, this may lead to automated integration into Electronic Medical Records (EMRs) or

even the National Poison Data System (NPDS), allowing for real-time improvement in initial clinical recognition of potentially toxic exposure patients and automated feedback of data to PCs.

Research Approach & Timeline:

We propose to approach the diagnosis of toxic exposures as a "multi-class classification" problem, with different potential toxic agents that patients could have been exposed to representing different unique "classes". Multiple machine learning approaches, involving different computational algorithms, have been developed and used for multi-class classification problems. The success of either approach is dependent on the quality and quantity of input data, which are used to algorithmically "learn" statistical rules that define a given outcome class. Machine learning models use the data variables supplied to help predict which class (toxic agent) each data point (patient) belongs to, with different models varying by the exact methods in which the variables (clinical features) are used to calculate a prediction.

The ToxIC Registry and the National Poison Data System (NPDS) are both potentially usable databases for developing machine learning models. Both encompass large numbers of clinically-significant toxic exposure cases and provide associated clinical data that can be used to train machine learning models. The total number of cases of individual toxic agents and the quality of associated label data input may significantly impact the success of the models at predicting outcome classes.

The "Scikit-learn" module[4] for the Python computer programming language[5] will be used to create machine learning models. Specifically, two separate independent approaches will be used to generate and compare two models: 1) the "Support Vector Machines" (SVM) approach, and 2) the "Naive Bayes" (NB) approach. These two approaches have the advantage of being relatively fast and require similar data pre-processing; they have also been used with similar data sets before with success[1,3]. Accuracy of models can be tested by generating each model on a subset of the available data, then testing if it correctly predicts the outcome class (toxic agent) of cases in the remaining portion of the data.

"ToxDiff", a rudimentary online educational tool we developed for toxidrome diagnosis support was previously presented at the NACCT 2018 Annual Meeting[6]. ToxDiff employs user-provided clinical data to calculate relative scores for each toxic agent in the database, based on a weighted matrix of point values for presence/absence of each clinical feature, and presents the user with a list of highest-scoring toxic agent matches. While it performed well on a small number of simulated cases, validation on retrospective human cases is still needed and its accuracy in real cases remains to be tested. Results from these machine learning models may help create variable-weighting matrices and decision-tree pathways to improve future similar software tools to the point of clinical usefulness. The current iteration of ToxDiff will be tested against a subset of data from the dataset and its predictive accuracy compared to that of the two machine- learning approaches.

If the machine learning models prove to be more accurate than ToxDiff, we will attempt to integrate either one or both into a new prototype software tool of similar design.

Further validation would still be necessary before clinical use, likely through a prospective trial in collaboration with a regional PC. A future goal may be integration with an EMR system that can push data to the machine learning model in real time,

allowing for true automated diagnostic assistance for clinicians. This goal is still far in the future, but the above work is a necessary step on that pathway.

Timeline:

1. April-June 2019: Obtain access to ToxIC and/or NPDS data
2. June-July 2019: Data preprocessing to create usable dataset
3. August-December 2019: Develop machine learning models:
    a. Support vector machines model
    b. Naive Bayes model
4. January-February 2020: Test accuracy of machine learning models against dataset
    a. Publish initial results of accuracy of machine-learning approaches
    b. Compare to ToxDiff accuracy on same sample dataset
5. March 2020-June 2020: Develop new diagnostic-assistance software tool incorporating best machine learning model into ToxDiff framework

Major Limitations/Questions:

1. Success of machine learning approaches is entirely dependent on the quality and quantity of the underlying datasets. If the data sets do not contain accurate outcome classification and variable values, then the created machine learning models will be significantly less accurate. Uncommon toxic agents may not have enough presence in the data sets for the machine learning models to use, and therefore will not be predictable by the final models. Given the total number of toxic agents that should be included for meaningful clinical use, a very large dataset will be needed to generate an accurate model, likely on the order of tens to hundreds of symptomatic cases per toxic agent and thousands of total cases. As some cases were likely included in both the ToxIC and NPDS data sets, and the identifying information necessary to distinguish them may not be present, we will not be able to merge the two datasets, limiting us to utilizing them independently.

2. Even with a large volume of data, machine learning models may not "work" in that they may not be able to accurately predict outcome class to a clinically useful margin. Other machine learning algorithms do exist and could be used if these approaches prove unsuitable.

3. Machine learning models can be susceptible to "over-fitting" the datasets used to generate them, in that they may accurately identify cases within the datasets but not externally. Prospective validation on future cases will be needed.

4. Multi-class classification focuses on sorting patients into unique *non-overlapping* toxic agent "classes" and does not directly account for the common situation presented by patients with exposure to multiple toxic agents or with complex underlying medical diseases. Separate machine learning methods can be used for "multi-label classification" where objects could have multiple labels (toxic agents) simultaneously.

However, with a large enough dataset to create the multi-class model, it may still turn out to be accurate at predicting toxic exposure with less computational expense. If the multi-class approach fails at accurate prediction, a multi-label approach would be a logical future direction.

5. We aim to create a clinically useful tool to assist with diagnosis of toxic exposures. This initial investigation is a critical step on the path to creating such a tool, but much future work will be needed to reach future goals of EMR integration and automated reporting to PCs. Creation of similar tools in other disciplines is an active research area in medical informatics. This project will help recognize medical toxicology as a potential area for future investment of informatics expertise and funding.

References:

1. Lodhi H, Muggleton S, Sternberg MJ. Multi-class Mode of Action Classification of Toxic Compounds Using Logic Based Kernel Methods. Mol Inform. 2010;29(8-9):655- 64.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
3. Gunčar G, Kukar M, Notar M, et al. An application of machine learning to haematological diagnosis. Sci Rep. 2018;8(1):411.
4. Pedregosa et al. Scikit-learn: Machine Learning in Python. JMLR. 2011;12(Oct):2825-2830. https://scikit-learn.org/.
5. Python 3.7.2. Python Software Foundation. https://www.python.org/.
6. Nogee D, Tomassoni A. Development of a prototype software tool to assist with toxidrome recognition. Clinical Toxicology. 2018;56(10)1049.