NATIONAL ACADEMIES Sciences Engineering Medicine

# Workshops to Support EPA's Development of Human Health Assessments: Artificial Intelligence and Open Data Practices in Chemical Hazard Assessment

## May 25th and 26th, 2022

## Posters

1. *Measuring the Health Effects of Severe Health Pollution Incidents Using Spatiotemporally Tagged Tweets*

Zach Calhoun, Duke University

Tracking air pollution's impact on human health is expensive and time-consuming: hospital data can be difficult to collect, and linking health outcomes to unhealthy air pollution may be difficult or impossible.

Prior research has demonstrated that Twitter can be used to monitor local air quality conditions, and that negative sentiment among twitter users can be correlated with air quality. However, the methods used in previous research used targeted term searches to capture responses to air quality, resulting in a method that is less generalizable to other environmental hazards, and that risks missing useful tweets containing terms omitted from the researchers' search terms.

We propose a new method for using social media to monitor the health impacts of air pollution from a much broader collection of tweets that could be applied to monitoring other chemical hazards. In this new method, Poisson Factor Analysis is used to decompose a spatiotemporally aggregated set of count-vectorized tweets into a set of latent topics. These latent topics are formed with the auxiliary goal of predicting the air quality index (AQI) associated with that set of tweets, and this system is efficiently learned and estimated with deep neural networks. The result of this training method is a set of interpretable topics that correlate with AQI, thus revealing the search terms most useful for monitoring air quality. We hope to further develop this method to focus the topics generated on health-related outcomes, thus linking local air quality with the expected physical health impact.

2. *Statistical Design Considerations for Lexicon-Based Information Extraction*
   Mireya Diaz, Western Michigan University
   A large proportion of information extraction (IE) via natural language processing (NLP) is based on lexicons. Although IE is considered a low-level task it is a fundamental block for higher-level NLP tasks, and its performance determines the performance of the NLP system. There are two main statistical design considerations for this framework. These pertain to the issues of existence of a lexicon, and building high-quality training, test, and validation data. For these two tasks it is imperative to know how large should the sample size of the required datasets be. The short answer to this question is the more the better. However, logistics, validation, and resource constraints limit that "infinite" size.

   To develop a useful lexicon among the design considerations are: size of the lexicon, the document's word length, prevalence of non-stop words, distribution of appearance of unique tokens, distribution of words' length. Complexity is added if categorization of words is sought, and if so whether these categories are mutually exclusive or if they exhibit some overlap. For building the training, test and validation datasets we need to consider tokens' capture probability, and the desired pipeline performance characteristics. This work illustrates the pertinent steps to estimate these sizes and their application in real-world scenarios.

3. *Exposure Health Informatics Ecosystem: An Infrastructure for Generating and Utilizing Exposomes for Translational Research*
   Ram Gouripeddi, The University of Utah

   Quantifying effects of the modern environment on health requires taking into account data from all contributing environmental exposures (exposome) which can span endogenous processes within the body, biological responses of adaptation to environment, and socio-behavioral factors. Exposomic research is translational in nature as the exposome includes direct biological pathway alterations as well as mutagenic and epigenetic mechanisms of environmental influences on the phenome. Generating exposomes requires integration of data from wearable and stationary sensors, environmental monitors, physiology, medication use and other clinical data, genomic and other biospecimen-derived, person-reported and computational models. This aggregation and integration requires to support variable spatio-temporal resolutions due to differences in study, experimental and analytical designs. Gaps in measured data may need to be filled with modeled data along with characterization of uncertainties.

We are developing a scalable computation infrastructure, the Exposure Health Informatics Ecosystem (EHIE) to address these needs. EHIE is a comprehensive, standards-based, open-source informatics platform that provides semantically consistent, metadata-driven, event-based management of exposomic data. Using an event-driven architecture allows the modeling and storage of all activities related to the study itself and its operations in their primitive form on a timeline as events that can be transformed to higher/analytical models based on use-cases. It is aligned with the goals of modern environmental health research supporting meaningful integration of sensor and biomedical data. It consists of the following components:

1. Data acquisition pipeline: Hardware and software tools, wireless networking, and protocols to support easy system deployment, robust sensor data collection, and feedback to study participants.

2. Participant facing tools: Collect and annotate various patient reported and activity data, as well as inform participants on their current clinical and environmental status.

3. Computational modeling: Generate comprehensive spatio-temporal data in the absence of measurements and for recognition of activity signatures from sensor measurements using artificial intelligence methods.

4. Central big data federation/integration platform: Standards-based (including ontologies), open-access infrastructure that integrates measured and computationally modeled data with biomedical information along with characterizing uncertainties associated with using these data.

5. Researcher facing platforms: Tools and processes for researchers undertaking exposomic studies for a variety of experimental designs and for clinical care.

In this presentation, we discuss the architecture of EHIE, and the generalizability of this multi-scale and multi-omics platform for providing robust pipelines for reproducible exposomic research using results from pilot projects using real-time, low-cost air quality sensors to provide spatio-temporal records of particulate matter exposures.

4. *Improving Expert Elicitation on Non-Traditional Chemical Threat Agents Using a Combined Risk Assessment and Bayesian Truth Serum Methodology*
Froggi Jackson, Gryphon Scientific

Historically, expert elicitations have informed strategies to address chemical threats. However, expert elicitations face many hurdles, including identifying who among a set

of potential experts are truly the most qualified. In response, we undertook a novel approach to assess which non-traditional chemical threat agents (i.e., non-chemical warfare) are of greatest urgency to address. Our approach used supervised machine learning (ML) augmented with a statistical method called Bayesian Truth Serum (BTS). In short, an ML algorithm was developed that can predict how experts would rank, or label, chemicals in terms of their risk, or "level of concern". The ML algorithm was developed using descriptive information collected from widely available data sources for 166 chemicals coupled with multiple rounds of model refinement based on expert labeling of a subset of these chemicals. Thus, the ML algorithm was underpinned by a large dataset and required only limited label inputs from the experts to "train" it. The algorithm predicts how experts would label any chemical, and could therefore stand-in for the experts when the risk of a new chemical needs assessed, given the availability of similar descriptive data for that chemical. Additionally, a ranked list of features determined to have particularly predictive ability in the model was generated; several of the top ten of these features reflect chemical toxicity. Using this approach, chemical agents of most concern from among the enormous body of non-traditional chemical agents can be rapidly identified without reliance on additional expert elicitations.

5. *Leveraging Natural Language Processing to Review Health Impacts of Air Pollution*

Natassja Lewinski, Virginia Commonwealth University

With the rapid expansion of our scientific knowledge and associated publications, it is becoming more challenging and laborious to curate and structure data for analysis to review current evidence on a research question. Over the past seven years, we have been collaborating to develop natural language processing tools to automatically extract information from scientific literature. We have created annotated corpora on nanomedicines and utilized publicly available corpora to evaluate different supervised learning approaches on their identification of experimental design factors and product characterization. Here, we demonstrate the application of our developed approaches to extracting information from journal articles describing experiments related to health effects of air pollution. We then present an analysis of the performance from two perspectives, the ability of the model to handle challenging language (e.g. lexically diverse entities, non-contiguous mentions, and dependencies between entities within categories) and the ability of the model to accelerate the literature review process for synthesizing knowledge in a mature research area. We conclude with reflections and recommendations.

6. *Interactive Reference Flow (I-REFF) Diagrams: A New Approach to Increase Efficiency and Transparency in Systematic Review Reporting*
Kristen Magnuson and Courtney Lemeris, ICF International, Inc.

Literature flow diagrams are a critical reporting element used to support transparency in systematic reviews. They are used to display the extent of the body of evidence that addresses a specific research question by efficiently conveying the literature search and screening process and documenting the number of included studies as well as how many studies were excluded and why. Literature flow diagrams are particularly important for hazard and risk assessments due to the high level of scrutiny these reviews face from scientific, political, and public communities. These diagrams have historically been manually generated, and while they add transparency to the reporting, this traditionally static format presents a number of limitations. Developing, maintaining, and ensuring the accuracy of literature flow diagrams can be time-intensive, particularly for large reviews or ones that require literature search updates. Furthermore, static literature flow diagrams provide limited, summary-level information. We demonstrate an approach to develop a new study flow diagram that is interactive (called Interactive REFerence Flow [I-REFF] diagrams), leveraging screening tools such as DistillerSR coupled with visualization software such as Tableau to efficiently generate interactive literature flow diagrams that are linked to the literature screening results. Linking screening data to the literature flow diagram allows summary counts to be automatically calculated so minimal effort is required to adjust the I-REFF diagram when a literature screen is updated. By linking to screening data, additional information (e.g., full reference citations and URLs) can be included as interactive elements in an I-REFF diagram without additional effort, allowing readers to quickly and easily identify studies considered in the review, and providing them greater ability to check and confirm, re-create, or build upon the review. The I-REFF approach increases efficiency during diagram development, minimizes potential for errors, enhances transparency and accessibility, and paves the way for further automation in the generation of literature flow diagrams.

7. *Analyzing High-Throughput Assay Data to Advance the Rapid Screening of Environmental Chemicals for Reproductive Toxicity*
Julia Varshavsky, Northeastern University

Rapid screening of environmental chemicals for reproductive toxicity is required to advance predictive toxicology and chemical assessment. High-throughput (HTP) assays have been used as platforms for rapid assessment of reproductive toxicants; however,

additional germline assays and method validation would further advance knowledge regarding fertility and reproductive health. We assessed yeast (S. cerevisiae) and roundworm (C. elegans) HTP assays related to germline function and other reproductive health endpoints by screening toxicity of 134 environmental chemicals and modeling each dataset using a streamlined, semi-automated benchmark dose software (BMDS) approach. We then extracted and modeled mammalian in vivo data from the Toxicological Reference Database (ToxRefDB, Version 2.0). We potency-ranked the data set from each evidence stream by BMD and evaluated rank order correlation between datasets using Kendall's Tau Correlation. Finally, we constructed a prediction model using machine learning to incorporate model descriptors (including data on QSAR, chemical properties, etc.) and tested model performance in predicting time to half-max (yeast assay). We compared agreement of the coefficient of determination (R2) and mean absolute error (MAE) between training and testing sets (80/20 split) to evaluate the model's predictive power. We found good correlation between data sets and great promise in predicting toxicological outcomes with reasonable accuracy (Training R2: 0.97, testing R2: 0.61). The agreement between the training set and the testing set showed that the algorithm has a great potential of making accurate predictions outside its calibration domain and could play a significant role in predicting reproductive toxicity of environmental chemicals.