Machine Learning Optimization: Defining Exposome-Metabolome Associated Aerodigestive Disease

G. Crowley[1], S. Kwon[2], B. Rushing[3], G. Grunig[4], S. Podury[5], S. McRitchie[3], S. Sumner[3], M. Liu[1], D. J. Prezant[6], A. Nolan[5]; [1]NYU School of Medicine, New York, NY, United States, [2]New York University School of Medicine, New York, NY, United States, [3]University of North Carolina at Chapel Hill, Kannapolis, NC, United States, [4]NYU Grossman School of Medicine, New York, NY, United States, [5]New York University Grossman School of Medicine, New York, NY, United States, [6]Fire Department of the City of New York, Brooklyn, NY, United States

RATIONALE Particulate matter(PM) exposure contributes to 7-million deaths annually. Aerodigestive complications include airway hyperreactivity(AHR), gastroesophageal reflux disease(GERD) and premalignant Barrett's Esophagus(BE). While studies are limited in clarifying the pathophysiological link between AHR, GERD and BE, discovery of biomarkers may yield biologically plausible pathways and identify populations needing targeted intervention. Detecting complex associations in high-dimensional datasets is a unique challenge. Machine learning(ML) is important in the analysis of untargeted metabolomics. Our work focused on improving the speed, precision, and accuracy of data-mining through use of random forests(RF) with gradient-boosted tree modeling to identify metabolite classifiers of AHR, GERD and BE. METHODS World Trade Center(WTC)-PM exposed firefighters with normal pre-exposure lung function were defined as cases of AHR by electronic medical record(EMR), PC_{20}<16mg/ml, and/or positive bronchodilator-response; GERD and BE status determined by EMR or endoscopy biopsy diagnoses. Controls were defined as not meeting any case criteria and were members of the 10%-randomly selected cohort controls. Metabolite serum profiles were analyzed by liquid chromatography-mass spectrometry and peaks were matched using in-house and public libraries. Machine Learning. RF of the untargeted, filtered, normalized metabolome(n=14,969 peaks) yielded a refined metabolite profile of the top 5% by mean decreased accuracy(MDA), and were included in a gradient-boosted tree model(xgboost,R-Project) to build a case classifier. The final model was determined by random hyperparameter space search, maximizing cross-validated AUC_{ROC}. RESULTS The refined profiles for all cases of AHR, GERD, and BE are shown with values of log_2(median-fold-change), Figure-1. AHR cases had decreased mesobilirubinogen compared to controls(5-fold cross-validated AUC_{ROC}=0.95;15% classification error(CE)),Fig-1A. GERD identified metabolites related to estrogen/testosterone metabolism: 11-oxo-androsterone glucuronide is a metabolite of testosterone; 2,3-bis(4-hydroxyphenyl) propionitrile is an agonist for estrogen receptor beta, 4-hydroxyandrostenedione glucuronide is a metabolite of the aromatase inhibitor 4-hydroxyandrostenedione,Fig-1B. No sex hormone metabolites were identified as key metabolites of BE. Acetaminophen and moxifloxacin were key exposures in GERD and BE, respectively. No overlapping metabolites identified cases of GERD and BE,Fig-1B-C. Final models AUC_{ROC}=0.88;19.5% CE for GERD and an AUC_{ROC}=0.93; CE=19.3%; for BE. CONCLUSIONS Our findings included differential expression of mesobilirubinogen, a metabolite of biliverdin. Heme oxygenase (HO)-1 metabolizes heme into biliverdin. Sex hormones and exogenous therapeutics were linked to GERD and BE. Transient receptor potential(TRP)V1 gene expression is altered in esophageal mucosa of reflux patients and TRPV1 is an acetaminophen target. Metabolic profiling provides a contemporaneous snapshot of an organism's physiology and may identify plausible pathways of disease that could be attenuated pharmacologically.

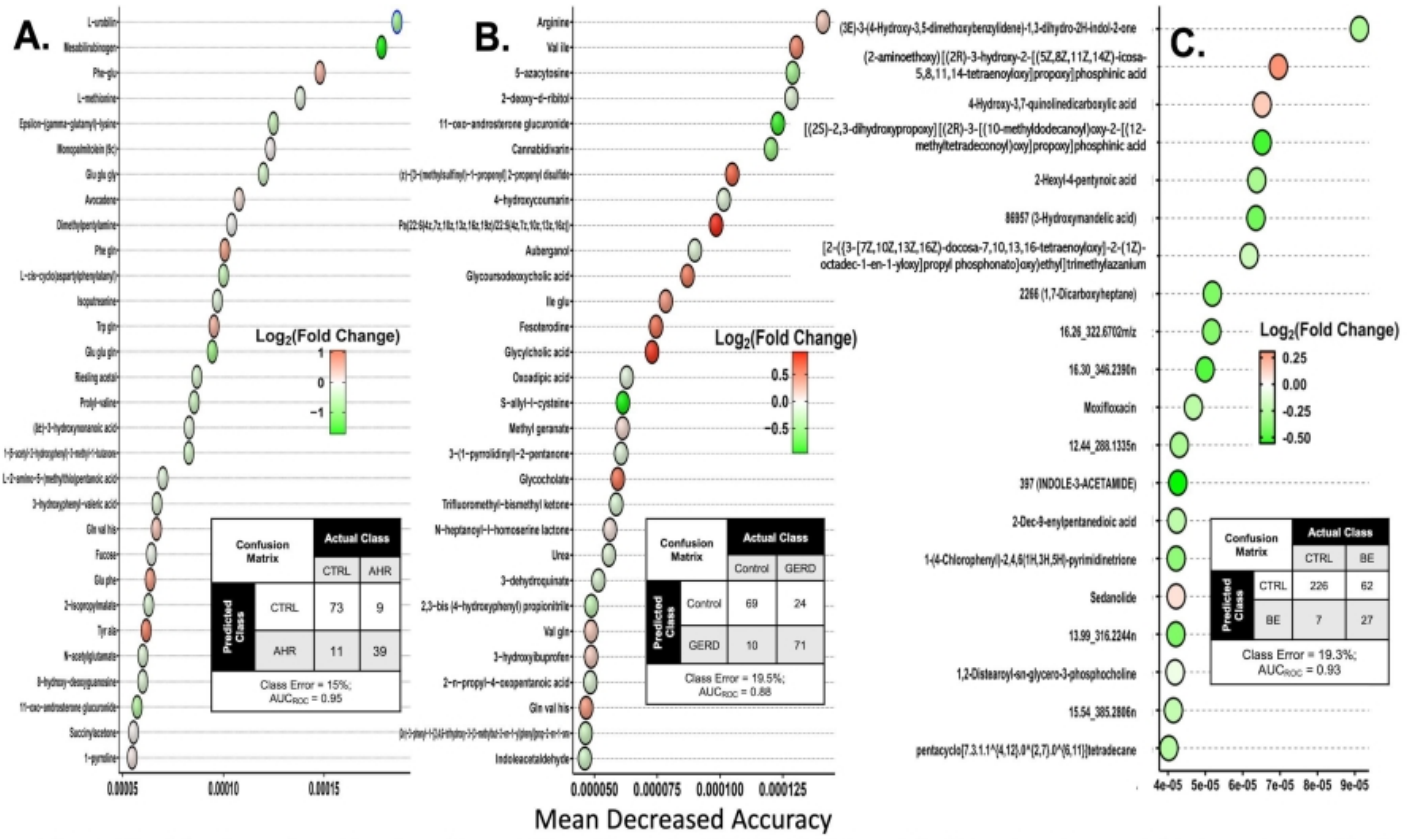

Fig. 1: Machine Learning: Random Forest Classifiers. A. AHR **B.** GERD and **C.** BE ordered on variable importance. Median fold change represented as Case/controls. In each assessment controls were individuals that didn't meet case status criteria.

This abstract is funded by: CDC/NIOSH U01-OH11855, -OH11300, and -OH01269; NCATS: UL1TR001445; KL2TR001446