Towards a Better Understanding of Complex Disease: Identifying Endotypes of Childhood Asthma

ClarLynda Williams-DeVane, David Reif, Stephen Edwards, Elaine Cohen Hubal, Edward Hudgens, Jane Gallagher

National Health and Environmental Effects Research Laboratory, USEPA National Center for Computational Toxicology, USEPA

Complex disease, where the diagnostic criteria cannot distinguish among differing etiologies, is often difficult to diagnose, treat and study due to the inability to classify individuals into suitable subtypes of the disease. Here, we aim to use and compare a combination of methods to identify the probable subtypes or endotypes of a complex disease, childhood asthma, based on three domains of data: gene expression, clinical covariates and disease indicators of allergy and childhood asthma as part of the Mechanistic Indicators of Childhood Asthma (MICA) study. Traditional analysis of complex disease considers one domain of data at a time to define the subtypes of complex disease. The commonly employed methodologies used to do so require a clearly defined phenotype representative of the same underlying disease process for supervised methods or a disease clearly identifiable by genomic or clinical data for unsupervised methods, neither of which is true of complex disease. To better define the endotypes of childhood asthma, we use standard methods such as Student's t-test and single data domain clustering as well as more complex, multi-data domain methods such as multi-step decision tree and modk-prototypes algorithm analysis strategies. We compare the results of each of the analysis methods to determine the best method based on 1) ability to classify known asthmatics and non-asthmatics, 2) ease of interpretation and 3) amount of mechanistic information gained. Secondly, we evaluate the impact of how each domain of data is incorporated in each analysis scheme. Standard methods that incorporate one data domain in the analysis stage, did not classify asthmatics and non-asthmatics well, were difficult to interpret and provided sparse mechanistic insight. Methods that incorporate multiple domains of data in the preprocessing, analysis and interpretation stages performed better in all three criteria. The multi-step decision tree method incorporates clinical covariates and gene expression data in the preprocessing step, indicators of childhood asthma covariates and gene expression in the analysis stage and all domains of data in the interpretation stage. This method classified known asthmatics and non-asthmatics the best, was easier to interpret than other methods and provided the most mechanistic insight. Insights of analysis methods with respect to complex disease and probable endotypes of childhood asthma will be presented. The understanding gained from the use of childhood asthma as a case study will lead to better understanding of complex disease in general through the development of more efficient methodologies.