**Exercise: Planning a Hypothesis Generating Study**

A paper by Agier et al is posted on Canvas. It describes a study conducted in Europe interested in generating hypotheses regarding the early-life exposome and lung function in children.

Study Description:

The study population is the HELIX cohort, a compilation of data from six existing population-based birth cohorts in Europe. The study reports complete outcome data for 1033 children. The primary outcome of interest is FEV1, the forced expiratory volume in 1 second and a common measure of lung function. It is measured using spirometry, a common clinical lung function assessment. Using both questionnaires and geographic residence to link to external data, the authors have compiled a large number of environmental factors for the children in the study, 85 prenatal measures and 125 postnatal measures. They are listed in Table 1. More details about the study are presented in the paper. The authors describe two approaches they used to generate hypotheses within their high-dimensional exposure data: an EWAS which examines each feature of interest individually and the DSA algorithm which is able to examine two-way interactions in an iterative, data-driven manner. Briefly, DSA is a linear regression model search algorithm that typically starts as an intercept only model. It then iterates to find the best model, at each iteration, there is either a removal of a feature (Deletion), substitution of one feature for another (S) or the addition of a feature (A) to the model. The iterative process is trying to minimize the room mean squared error within a cross-validation framework. This was extended to include 2-way interactions between features.

Class Exercise:

Forget what the authors did. You and your group have been tasked with using these data to generate hypotheses regarding combinations of environmental features and lung function in children.

1. What are some limitations of these data that may impact your ability to draw inferences from your analysis?
2. Sketch out an analytic pipeline to address your task. Similar to Assignment 7, your pipeline should include the following components:
   1. Data Preparation
   2. Partitioning/Resampling
   3. Tuning of Hyperparameters
   4. Evaluation
3. Read the “Research in Context” section of the manuscript. Do you agree with the authors’ interpretations of their findings and the implications of their work? What if anything would you have changed or rephrased in this section?