**Predicting Condition Flare Up in Patients with Respiratory Disease History**

**Description**

Respiratory diseases including asthma, cystic fibrosis, lung cancer and smoking related diseases like chronic pulmonary obstructive disease (COPD) are a severe global medical problem which have affected over 50 million people worldwide.  Respiratory disease is a leading cause for death, globally. As their condition worsens, a fraction of patients experience flare ups. Flareups lead to sudden worsening of symptoms such as shortness of breath and increased airway inflammation, often requiring immediate medical treatment and emergency room visits.

Our goal is to design a decision making tool that would assist medical professionals in identifying patients most likely to have a flare up based on a parsimonious subset of predictor variables.  For this specific task, we would like to restrict our patient population to a particular type of respiratory disease and to limit the number of variables to ones that are easily measured in a typical medical office visit.

**Objective**

The objectives are as follows:

* Given baseline data from patients in clinical studies and longitudinal follow-up outcomes, develop an algorithm to predict which patients are more likely to exhibit a flare up.
* Uncover top variables which can help identify and monitor flare up of disease based on the model output. Variable importance is a key metric.

**Data**

In this competition you'll work with observational data from patients with a specific respiratory disease and predict which patients are more likely to exhibit a flare up. Each observation corresponding to one patient belongs to one (and only one) group – Development or Scoring dataset. In order to protect patient confidentiality, all data has been anonymized and the observations are all normalized.

The master dataset includes a total of 1985 patients.

RAW dataset has been transformed and information is as follows:

* Transformed Data: Contains 60 columns of medical derived data that are divided into the following broad categories.  While the exact column names are hidden to protect patient privacy, the following contains examples of potential categories:
  + **Flare\_Up**: is a categorical column where 1= Flare Up and 0 = No Flare Up.  This is the variable to be predicted.  This is provided in column 2 in the dataset.
  + **Demographics**:  This includes variables like Age, Gender, Height, Weight, Race, and Nationality and includes categorical and binary variables.
  + **Disease Stage**: This is two different measure of disease severity.  These are categorical variables and higher values represent more severe disease
  + **Lung Function**: These are 20 continuous variables that were derived from lung function measurements determined by spirometry.  See http://en.wikipedia.org/wiki/Spirometry for an explanation of spirometry.
  + **Disease History**: These include specific history of lung disease symptoms.  These include some continuous variables and some conditional variables.  Conditional variables may include thing like have you ever been hospitalized for respiratory illness and if so how many times.
  + **Other Lung Disease**:  These include history of lung problems like history of cough, asthma or other lung ailments.  There are three types of variables, binary (yes or no) tertiary (yes, no, I don’t know) or continuous. Some of the variables are conditionally related.  Conditional variables may include thing like have you ever been treated for respiratory disease and if so how many times.
  + **Respiratory Questionnaire**: Results from two independent questionnaires to measure overall health well-being in patients with lung disease.  Four of the variables represent different scores derived from the same questionnaire
  + **Smoking History**: 4 measurements that ascertain a subjects smoking habits.
  + Missing values are not transformed and represented as it is in the data.
* Metadata / Data Dictionary: Contain the variable name, the broad category as outlined above, a list of the data types for each column in the data file.  And whether the variables are linked.  There are likely other dependencies in the data and participants are encouraged to identify other relationships. See data dictionary. Columns are annotated as:
  + Binary
  + Continuous
  + Category (The order of the numbers are meaningless)
  + Ordered Category (The order of the numbers is relevant)

This master data set has been split into Development and Scoring data sets for the tasks. Participants will use the development dataset to build their models. The Scoring data set does not have the **Flare\_Up** Variable. Participants need to use the model built to predict the Flare Up in the Scoring Dataset.

**Contest Tasks and Evaluation**

1. Participant will be evaluated based on
   1. **Area Under the Curve (AUC)** criteria in the Scoring Dataset
   2. **Variable Importance** derived from the Development Dataset

Note that the AUC will be computed by the Evaluators

1. Participants are encouraged to engineer new features from the Development Dataset but explain the role of the derived variables and its impact on the dependent variable clearly
2. Codes can be written only in R or Python and should be submitted. Evaluators will re-execute the code and should be written in a way that it works end to end.
3. The following are the artefacts expected from the participants
   1. Predictions – Probability Score and Actual Predicted Value from the Scoring Dataset mapped to the Patient Identifier and explanatory variables
   2. Well commented working R/Python code for evaluation
   3. One/Two slides explaining
      1. Methodology
      2. Model Assumptions
      3. Explanation and Validation of engineered variables
      4. Variable importance of Top Predictors
      5. Any insights and recommendations



1- Data Dictionary



