## Problem Statement

The client has a product in development for a rare disease in which physicians recognize differences in severity, but no guidelines exist for our client to map individual patients directly to severity levels. The client's product will be targeted in patients with moderate to severe disease. The client has a database that contains clinically-relevant information about the disease (i.e., flags indicating the presence of symptoms and a variable counting the total number of symptoms) and each patient has been rated as mild, moderate, or severe by an actual physician. The client would like to extract the mental heuristic physicians are using when they label a patient with a severity

## Approach

- Initially, I started out with loading the data set and finding any null values - Then, I explored the feature space:

## Data Cleaning & Exploration

- Found out that the dataset is balanced ie., it has equal number of observations for each level of the target variable
- Further, plotted relationships between every feature and the target variable and analysed the relationship
  - Note: Did not perform correlation, feature scaling and feature engineering because I planned to use decision trees (that do not require any of these) so as to obtain rules classifying observations into one of the target classes

## Data Modeling & Fit Evaluation

- Started by fitting a general decision tree model with default parameter values. Computed training and test accuracies to confirm that this model is overfitting
- Found the best parameters using grid search, which will actually lead to overfitting so as to use general conscience in choosing the parameter value wisely
- Finally, fitted the decision tree model by improving on the parameter value so as to avoid overfilling. Training and test accuracies come out to be much similar than the last time, it appears to be a good model
- Confirm the efficacy (goodness of fit) of the model by fitting a Random Forest Classifier to the data, which should fit better than decision tree given the nature of the model of bootstrap aggregation that avoids overfitting. It is observed that test accuracy obtained for this model is similar to that obtained for decision tree model and hence, we go ahead with our decision tree model to formulate the rules

## Confusion Matrix: Evaluating the fit of the model

Confusion Matrix: [[25 7 1]
[ 0 27 6]
[ 0 3 30]]
Train Accuracy: 87.06467661691542
Test Accuracy: 82.82828282828282

We look at the accuracy values because the dataset is balanced such that we have equal number of observations for each type of target variable namely, mild, moderate and severe.

## Rules

#For Mild,
Rule: At most 2 symptoms and no depression or cramps
#For Moderate
Rule1: At most 2 symptoms, has depression but no spasms #Rule2: Number of symptoms between 2 and 3, but no cramps
#For Severe
Rule1: At most 2 symptoms, has cramps but not depression
Rule2: At most 2 symptoms, has cramps and depression
Rule3: Number of symptoms between 2 and 3 but not cramps
Rule4: Number of symptoms more than 3 and may or may not have spasms