Risk Identification and Prediction for Long COVID-19

Elliot Kim, Emma Jin, Victoria Song

- **Central hypothesis:**
  - We want to identify the risk factors for long COVID and predict the likelihood that a patient would suffer from Long COVID. Long Covid refers to experiencing persistent symptoms after recovering from COVID-19.
- **Problem Description**
  - The problem we seek to address is: **What risk factors contribute the most to persistent symptoms after recovering from COVID?** More specifically:
    - What are the most prominent risk factors that contribute to persistent symptoms for patients diagnosed with COVID-19 (also known as long COVID)? For each persistent symptom, what are the most prominent risk factors for this particular symptom? Are they different for the risk factors that we identified for persistent symptoms in general?
  - Given the answer to this problem, we also want to further explore:
    - How well can we predict the likelihood of a patient suffering from long COVID in general? What about for each specific symptom that is a part of long COVID?
  - A solution to this problem would look like:
    - We identify a list of risk factors that cause a patient to suffer from persistent symptoms in general; for each possible persistent symptom, we identify a list of risk factors and compare this to the risk factors for persistent symptoms in general. We analyze these factors and identify the most prominent risk factors (and possible ways to address these factors if probable). Based on these risk factors identified, we will also modify our model and have a working model in the end for predicting the likelihood of a patient suffering from long COVID.
  - People who would stand to benefit from a solution to this problem:
    - Patients who are suffering the effects of persistent symptoms from COVID-19, or long COVID. Identifications of the risk factors may provide a primitive insight to how to combat persistent symptoms for COVID-19 in the long run. At the very least,

it can serve as an alert for patients who are susceptible to one or more of the risk factors identified.

- How we would know if we've solved the problem:
  - If we are able to identify the most prominent risk factors for long COVID and build a model that relatively accurately predicts the likelihood of a patient suffering from long COVID, then we know we have solved the problem we are trying to answer.

- **Algorithms:**
  - We want to explore multiple models (e.g. **SVMs, Random Forests, Logistic Regression**...) and look at their weights for each feature in order to identify the most prominent risk factors. Then, we want to test these models on novel data and find the model with the best performance (i.e. that could best predict the likelihood that a patient would suffer from long COVID). We plan to use default models from Scikit-Learn for all these models.
  - **Anomaly detection:** it is possible that we have a highly imbalanced dataset (since there is a relatively small proportion of COVID patients who experience persistent symptoms) that needs to be upsampled or downsampled. We plan to use *sklearn.utils.resample* for downsampling.
  - **Dimensionality Reduction:** Because our dataset has a huge amount of features, we want to conduct dimensional reduction methods to get rid of noise and distraction and focus on the important features. We plan to implement this ourselves.
  - **Rank Prediction:** Instead of having a binary target for predicting whether a person has Long Covid or not, we can implement a ranking system, so that patients are given a score that indicates the possibility of them suffering from Long Covid. We plan to implement this ourselves.
- **Data:**
  - We are using the 2021 Medicare Current Beneficiary Survey COVID-19 Fall and Winter Supplement Public Use Files (link to Public Use Files) provided by the Centers for Medicare & Medicaid Services (cms.gov). These datasets were selected because they contained data on eight different types of "persistent symptoms," which would serve as the labels for our models. Furthermore, the datasets contain highly similar features, so combining them manually would not be particularly difficult to do as a pre-processing step.
- **Experiments:**

- First of all, we plan to train separate models (SVM, random forest, logistic regression) on the same dataset (with the label being whether the patient suffered from long COVID or not) and look at their weights to identify the most prominent risk factors for long COVID in general. Then, we want to look at each persistent symptom separately and train the models on each symptom. In the end, we want to make a comparison between the risk factors identified and analyze whether, how, and why they differ from each other.
- In order to optimize our models, we want to use:
    - Hyperparameter Tuning: Since we are choosing models that have hyperparameters (SVM, random forest), we plan to use hyper-parameter tuning for each model in order to ensure best performance for each model.
    - Stratified K-Fold Validation: We want to perform a stratified k-fold validation on our models to ensure the stability of our models and combat randomness of our data. We also want to look at the validation curves to make sure that we are not underfitting/overfitting our data.
- Secondly, we want to test our models, compare their performance, and pick the one with the best performance. We will evaluate the performance of the models by looking mainly at accuracy, but we will also plot the confusion matrix and look at the precision, recall, f1 score, etc. in turn.

- **Impacts:**
    - This project has the potential to benefit COVID-19 patients and their healthcare providers by revealing the risk factors involved in long COVID. Although many COVID patients who recover from the disease are able to do so fully, there exists a subpopulation of patients who retain COVID symptoms. These symptoms can range from headaches to shortness of breath to anxiety to heart problems. Especially considering the (continued) prevalence of COVID-19, it will be important to mitigate even what may appear to impact a fraction of COVID patients. The better healthcare providers are able to identify patients at risk of long COVID, the more reliably patients will receive the appropriate resources for recovery.
    - Because the dataset's features include patient demographics and accessibility to various resources (information sources, care, etc.), this project also has the potential to characterize the role of social factors in long COVID prognosis. This can help in the preventative healthcare space such that those patients are not underserved in this regard.

- **References:**
  - Our data comes from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/MCBS-Public-Use-File