PREDICTING LONG COVID RISK USING EHR, GENETIC, AND SURVEY DATA

Christopher Guardo, Srushti Gangireddy, QiPing Feng, Henry H. Ong, V. Eric Kerchberger, Wei-Qi Wei
Wei Lab, Center for Precision Medicine, Department of Biomedical Informatics

PREDICTING LONG COVID CASES

As the number of Covid19 variants rise, the focus of the healthcare industry has shifted to predicting the most extreme cases of Covid19 with long term health risks.

EHR DATA BASED MODEL

The original model made by the N3C project team was a Random Forrest Classifier, which assigned patients a 0-1 risk score based on their given EHR medical history.

INCLUSION OF GENETIC DATA

The genetic portion of data included six different genetic components which were found as significant indicators in a Genome-wide association study.

INCLUSION OF SURVEY DATA

Survey data gathered included 42 questions on:

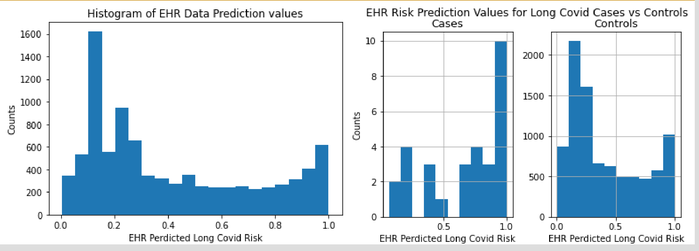
- Overall Health and Drug Use
- Lifestyle and Living Situation
- Insurance type and Availability

METHODS

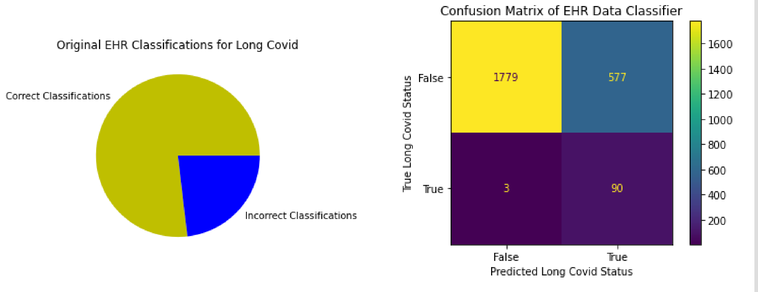
- Construct two complementary machine learning models for genetic and survey data to give greater insight to a patient's risk for long COVID
- Evaluate a patient's overall risk using a weighted average, between the EHR model's prediction and the prediction from the new model.
- Using tuning parameters for the weighted average, increase the accuracy of our prediction method by balancing the influence of each data type.

TRAITS FROM THE ORIGINAL EHR BASED MODEL

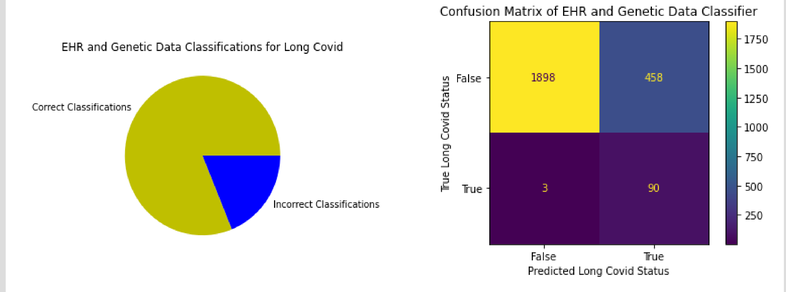
- Fairly accurate prediction of long COVID cases
- Misidentified many controls, leading to high type 2 error
- Skewed distribution of long COVID risk scores



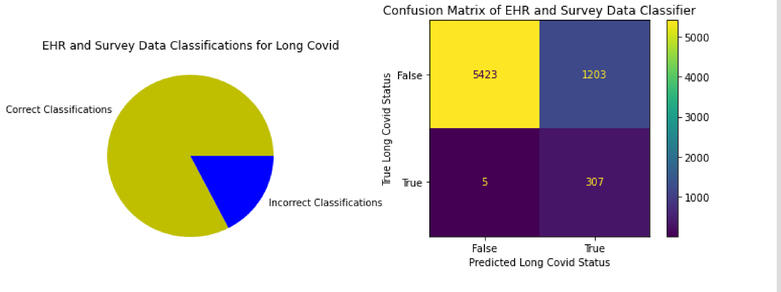
EHR DATA ALONE PREDICTING LONG COVID



GENETIC DATA AND EHR DATA MODEL



SURVEY DATA AND EHR DATA MODEL



CONCLUSIONS

The inclusion of genetic or survey data to our original EHR based prediction increases accuracy of our overall prediction by decreasing our type 2 error. This means the genetic and survey data gathered give us power to better differentiate cases that may appear from their medical history as long covid risks, when their risk for long covid is relatively low. Further research will be directed in combining all three medical data sources and updating the models with new data.

Please contact the corresponding author at chris.guardo@vumc.org

VANDERBILT UNIVERSITY MEDICAL CENTER

All of Us RESEARCH PROGRAM