Genetic Correlates and Causal Inference in Covid19 Severity of Response

Dominik Damjakob
damjakob@stanford.edu

Jacob Edelson
jedel@stanford.edu

Xinyu Hu
xhu17@stanford.edu

Mentor: Dr. Manuel A. Rivas

Introduction

Covid-19, also known as the Coronavirus, is a viral disease that was first discovered in Wuhan, China, in December 2019. Subsequently, it spread around the world causing a worldwide pandemic.

During the past year, a unique scientific collaboration occurred to fight and treat the virus. On the genetic front, a general data base for the genetic properties of people affected by different outcomes of Covid-19 has been created. However, no greater study has so far evaluated the genetic properties of the virus itself. We will try to model this problem by comparing the genetic architecture of Covid-19 to other, already known traits, such as the metabolite exposure. As these traits are already well-researched, this allows for a thorough mapping of the virus. To conduct this research we analyze two problems:

Dataset and Approach

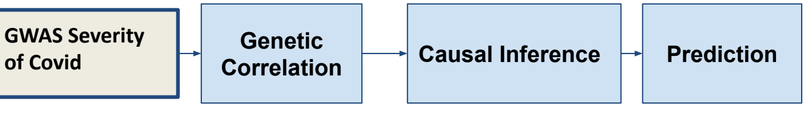
Dataset:

The COVID-19 Host Genetics Initiative (HGI) was developed in March 2020, at the height of the COVID-19 global pandemic from the Institute of Molecular Medicine in Finland (FIMM) and the Broad Institute of MIT and Harvard. For this Analysis we used the Freeze 3 Data, which included participants from both the UK Biobank and 23andMe. We als used publicly available summary statistics from the Ben Neale Lab and Metabolite and Cytokine Summary statistics from Univerisyt of Bristol.

Approach and Experimental Setup:

We will seek to understand and define:

1. The genetic correlation between Covid-19, Metabolites and Cytokines
2. The causal relationship between Metabolites, Cytokines and Covid-19
3. Prediction of Covid19 Response Severity based on clinical features..



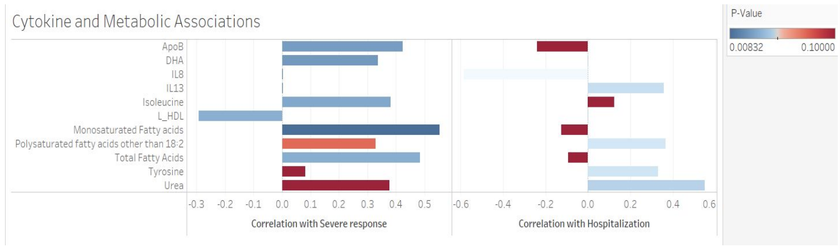
References:

1. Jason W.Wei and Kai Zhour. "EDA: Easy Data Augmentation Techniques for boosting Performance on Text Classification Tasks," In: CoRR abs/1901.11196 (2019). arXiv: 109.11196
2. Rico Senrich, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data", In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol I), Berlin, Germany, Aug 2016
3. Toxic Comment Classification Challenge, https:www.Kaggle.com/c/jigsaw-toxic-comment-classification-challenge-discussion

Correlation Analysis

Linkage Disequilibrium Score Correlation (LDSC):

- examines pairwise shared genetic architecture between any two traits
- computes sum over all SNPs of R-squared for regression with all other SNPs
- Severe COVID and risk of hospitalization most strongly correlated with high BMI data
- also Cytokines and Metabolites significantly correlated



Causal Inference

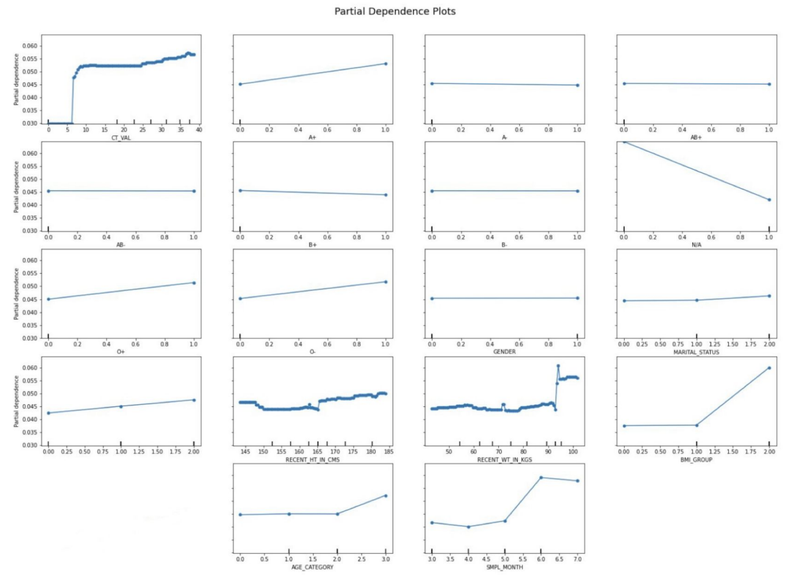
Mendelian Randomization

- replicate random study to avoid confounding error or reverse causal effects
- checks for intermediary effects via known biological effects
- 5 Metabolites were inferred to be causal in COVID hospitalization: CTACK, MCP1, MIG, MIP1b and RANTES

Individual Severity Prediction

The predictive models being used include logistic regression, support vector machine, random forest, XGBoost and neural network classifier. Based on accuracies obtained by the above models, we select the weight for the ensemble model. Overall, the ensemble model achieves an F1 score of 0.71 and AUROC of 0.75.

In order to better understand the mechanism of these models, their dependence plot is shown below.



Conclusion / Future Work

We have successfully found a strong effect of BMI related data, a relationship that matches with existing findings, for both causal and predictive models. More factors like smoking habits could be used, which we did not include due to a lack of data.