Project Title: Prediction of coronavirus infections and complications in individuals from UK BioBank Data

Company Background:

Catalyst Health Tech Integration (HTI) located on 3515 Brighton Blvd in downtown Denver's RiNo district includes an off-site location for Mines site focusing on an initiative for AI in bio and health. The Catalyst HTI is a networking hub for stakeholders in the health domain in Denver featuring over 70 companies, academic and non-profit institutions. Mines has an off-campus location at the Catalyst HTI. The vision for Mines@Catalyst is to foster collaboration between entrepreneurs, industry leaders, engineers & researchers as the fastest way to accelerate innovation, advance care, and improve lives. The Colorado School of Mines has joined the Catalyst community to showcase its presence in the health-tech sector and the city of Denver. The Catalyst provides opportunities for seamless integration of Mines with the health tech community in Denver. In the current crisis, our projects focus on COVID-19.

Background and Motivation: COVID-19 Pandemic:

April 2, 2020 marked a grim milestone in the COVID-19 epidemic caused by the novel coronavirus: more than a million people are confirmed to be infected and more than 50,000 people have died. The largest number of confirmed cases of any country is in the US: 250,000 people diagnosed, with an unknown number of undiagnosed cases. Life as we know it is on hold, affecting everyone in the US and the world. Stay home orders are in place since several weeks but the number of cases still rises exponentially [1]. We therefore urgently need to investigate the role diverse factors play at the individual and the population levels in determining disease spread. We here propose to predict at the individual level, if a person is infected or not, and if that person will develop severe symptoms that require hospitalization or not.

Team Size: 3-4 Students

Location: Remote, client/team meetings with zoom, frequent communication via slack/email

Project Summary: Large amounts of clinical, epidemiological, sequence and molecular data has been collected as evidenced by more than 1000 scientific articles relating to the novel coronavirus and a number of dedicated databases and websites disseminating these large data. We have previously worked on predicting host-pathogen interactions with the goal of target discovery from host-pathogen interaction networks, including HIV, Salmonella and several other human pathogens. In all of these projects, we collected from large numbers of publications and databases gold standard datasets used for training binary classifiers and diverse biological data for use as features. We here propose the hypothesis that we can develop two binary classifiers, one for infection and one for complications, by integrating diverse data relating to COVID-19. There is published clinical data for >800 patients for which the truth is known (infection and its severity), so it is suitable as a gold standard to learn from. From clinical data we can use correlations with disease history, co-

morbidities, blood type and molecular profiling data where available from patients, combined with analysis of available omics datasets (more than 4000 novel coronavirus sequences are available and two high quality proteomics datasets).
From the proteomics data, we can infer

- Biomarkers of Coronavirus Infection (relevant to diagnostic yes/no infected or not)
- Biomarkers of Coronavirus complications (relevant to severe - mild symptoms)

We can link proteomic (from the DisGeNet database) and clinical data (from the UK biobank, see Table 1) through mapping of the disease classification id's. By integrating the clinical with the proteomics data, we propose to predict two binary classifiers:

(a) Patient is infected or not
(b) Patient will develop severe or mild symptoms

The key to this project is to use the UK Biobank Data to create machine learning models.

Table 1. UK Biobank data (Basket ID: 2008223). [This basket contains: 41 tabular, 0 bulk, 1 HES record, 0 genetic SNP and 0 dataset fields.]

Alcohol
Asthma outcomes
Baseline characteristics
Blood pressure
Breathing
COPD outcomes
Circulatory system disorders
Coronavirus COVID-19
Death register
Ethnicity
General health
MET Scores
Medical conditions
Medical information
Medications
Ongoing characteristics
Reception
Smoking
Summary Diagnoses

Key Skills/Technologies: Database handling, conversion into training/test set, extraction and encoding of features, classification.

Student Benefits:

- Learn about collaborating with non-experts (here biologists)
- Work on a high impact project

Contact Information:

- Prof. Judith Klein, judithklein@mines.edu