Predicting COVID-19 Resistance Using JH-CROWN Dataset

Kai-Wen K. Yang[1], Chloé Paris[1], Kevin Gorman[1], Ilia Rattsev[1], Rebecca Yoo[1], Yijia Chen[1], Tony Wei[1], Jacob Desman[1], Joseph Greenstein[1], Casey Taylor[1,2,3], Stuart Ray[2,4]

[1]Department of Biomedical Engineering, Johns Hopkins University; [2]Department of Medicine, Johns Hopkins School of Medicine;
[3]Division of General Internal Medicine, Johns Hopkins School of Medicine; [4]Division of Infectious Diseases, Johns Hopkins School of Medicine

Team Mountain Goats

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Background

Little is known about the human genetic and immunological basis of resistance to SARS-CoV-2. It has been observed that mean secondary attack rates for SARS-CoV-2 infections can reach up to 70% in some households, and several families reported that all their members except one of the spouses were infected. This suggests that some highly exposed individuals may be resistant to infection. In addition, little is known about whether the occurrence of COVID-19 resistance differs between people by health characteristics as noted in the electronic health record. This study aims to provide such insights among individuals indicating previous exposures.

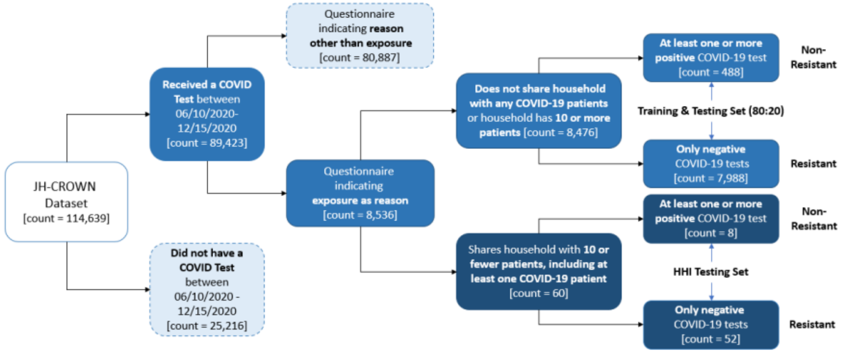
Objective

To build a data-driven computational model to predict COVID-19 resistance in individuals with a COVID-19 exposure.

Methodology

Patient data from the JH-CROWN dataset was filtered using the inclusion/exclusion criteria shown in Figure 1 into resistant and non-resistant cohorts. The cohort was further split into high and low confidence exposure groups based on the household index (HHI). Maximal-frequent All-confident pattern Selection and Pattern-based Clustering (MASPC) was used to identify clusters of patients in the resistant cohort and significant patterns between cohorts. Classification models of patient resistance status using XGBoost (XGB), Random Forest (RF), and Logistic Regression (LR) algorithms were constructed and evaluated for performance.

INCLUSION/EXCLUSION CRITERIA



JH-CROWN Dataset [count = 114,639]

Received a COVID Test between 06/10/2020-12/15/2020 [count = 89,423]

Did not have a COVID Test between 06/10/2020 - 12/15/2020 [count = 25,216]

Questionnaire indicating **reason other than exposure** [count = 80,887]

Questionnaire indicating **exposure as reason** [count = 8,536]

Does not share household with any COVID-19 patients or household has **10 or more patients** [count = 8,476]

Shares household with **10 or fewer patients, including at least one COVID-19 patient** [count = 60]

At least one or more positive COVID-19 test [count = 488] Non-Resistant

Training & Testing Set (80:20)

Only negative COVID-19 tests [count = 7,988] Resistant

At least one or more positive COVID-19 test [count = 8] Non-Resistant

HHI Testing Set

Only negative COVID-19 tests [count = 52] Resistant

Figure 1. Inclusion and Exclusion Criteria used to filter JH-CROWN dataset

Results

DESCRIPTIVE STATISTICS OF RESISTANT STUDY SUBJECTS

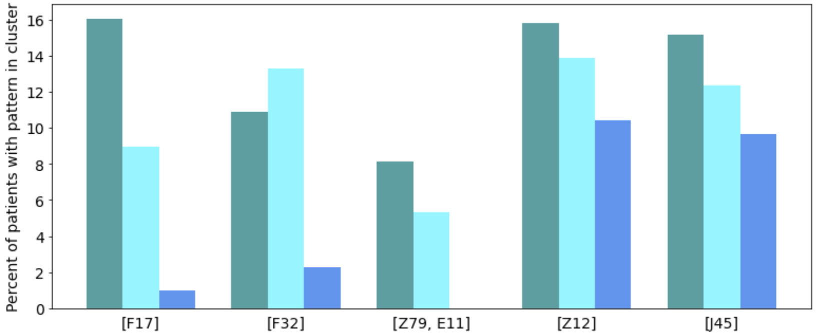


Figure 2. Clustering results from MASPC. Distribution of patients with patterns of ICD10 codes: nicotine dependence [F17], depressive episode [F32], long term drug therapy & Type 2 diabetes [Z79,E11], screening for malignant neoplasms [Z12], and asthma [J45].

DISCREPANT PATTERNS FOUND IN PATIENTS

Pattern	Odds Ratio	P-Value
Long-term (current) drug therapy; Screening for infectious and parasitic diseases	0.59	0.0055
Screening for infectious and parasitic diseases; Encounter for immunization	0.67	0.01
Disorders of fluid, electrolyte and acid-base balance	0.70	0.021
Dorsalgia (back pain)	0.73	0.043
Personal history of malignant neoplasm	1.54	0.036

Figure 3. Prevalence of patterns found using MASPC method in both resistant and non-resistant patients. Five diagnostic code patterns were found with a p-value less than 0.05. Odds Ratio less than 1 indicates prevalence in non-resistant cohort, whereas odds ratio greater than 1 indicates prevalence in resistant cohort.

PREDICT PATIENT RESISTANCE STATUS

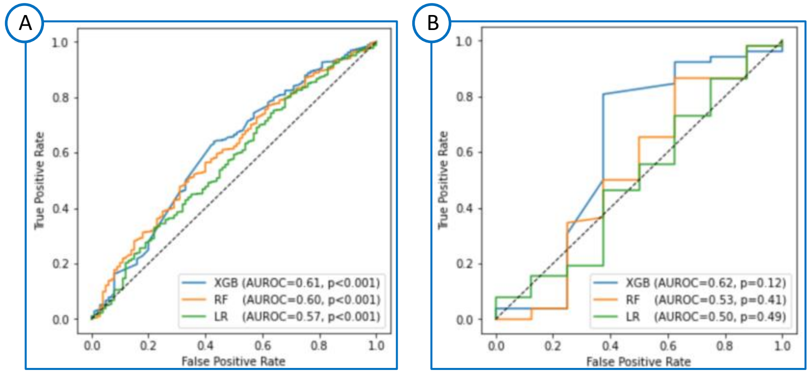


Figure 4. Receiver Operating Curves of XGBoost (XGB), Random Forest (RF), and Logistic Regression (LR) models.

4A. Testing Set: XGB is the best performing model and all three models have statistically significant AUROCs (p<0.001)

4B. Household Index Testing Set: XGB is again the best performing model, yet the p-values are less statistically significant due to the small sample size

Conclusion

The clustering results from MASPC indicate that one pattern is associated with SARS-CoV-2 resistance and four patterns are associated with SARS-CoV-2 non-resistance. We also observe that patients in the non-resistant cohort have more comorbidities and medications than the resistant cohort. Our model can predict the patient resistance status with AUROC of 0.61 and p-value of <0.001 on the testing set, indicating statistical significance.

Future Directions

We implement the above methods in a restricted dataset. In the future we wish to evaluate how the model and feature associations identified using our cohorts can be generalized to the general population. In addition, it would be important to validate features associated with resistance/non-resistance through more advanced association studies.