Predicting COVID-19 Resistance Using JH-CROWN Dataset
Kai-Wen K. Yang1, Chloé Paris1, Kevin Gorman1, Ilia Rattsev1, Rebecca Yoo1, Yijia Chen1, Tony Wei1, Jacob Desman1, Joseph Greenstein1, Casey Taylor2,3, Stuart Ray4,5

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

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.

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

INCLUSION/EXCLUSION CRITERIA

RESULTS
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].

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.

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

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.