# Molecular Epidemiology of the COVID-19 pandemic in Chicago

Ted Ling Hu[1,2] , Lacy M Simons[1,2], Egon A. Ozer[1,2], Ramon Lorenzo-Redondo[1,2], & Judd F. Hultquist[1,2]

[1]Division of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL, L 60611, USA. [2] Center for Pathogen Genomics and Microbial Evolution, Northwestern University Institute for Global Health

## Background

The COVID-19 pandemic has severely ramped up genomic surveillance on a global level that allows researchers to monitor shifts in viral trends that could affect pathogenesis, virulence, host range and immune escape. Although most mutations in SARS-CoV-2, the causative agent of COVID-19, are often deleterious, there is a small proportion that will survive and develop new functionalities that allow the virus survive longer or infect quicker. In Chicago, we see four main peaks of COVID-19 cases, each associated with a different variant of SARS-CoV-2. The beginning of the pandemic was marked by Nexstrain clades 19A and 19B. The second peak, which began November of 2020, marked the highest case counts of Chicago and consisted primarily of clade 20G. Then subsequent waves of COVID-19 in April of 2021 and August of 2021 were marked by the Alpha variant and the Delta variant. Although 20G was the prevalent variant during the height of the pandemic, it was marked by the lowest hospitalization to cases ratios as well as death to cases ratios. **Therefore, we hypothesize that 20G clade become predominant in the United States in the later half of 2020 because it caused less severe disease and caused asymptomatic infections.**

## Methods

We collected nasopharyngeal swabs and blood samples from 7080 patients(6448 outpatient, 632 inpatient). Of these 7080 patients, we have sequencing data for 1373 of them. Clinical and demographic data were extracted from the Northwestern Medicine Enterprise Data Warehouse (NMEDW). We were also able to utilize publicly available data from the city of Chicago City Data Portal on case counts, hospitalizations, deaths and the breakdown of these by demographics. Multiple hypothesis testing was done using the Kruskal-Wallis Test and posthoc analysis was performed using Dunn's test to perform pairwise comparison. Significance was determined where $p<0.05$. Statistical analysis and modeling were run on Python (v3.8.8) using packages Scipy (v1.6.2), sklearn (v0.24.1) and pandas (v1.2.4).

## Overview of COVID-19 in Chicago

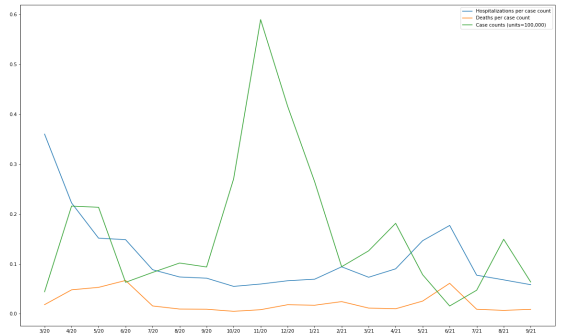

**Figure 1**. An overview of case counts in Chicago.

## Clade Distribution of Sequences from NU

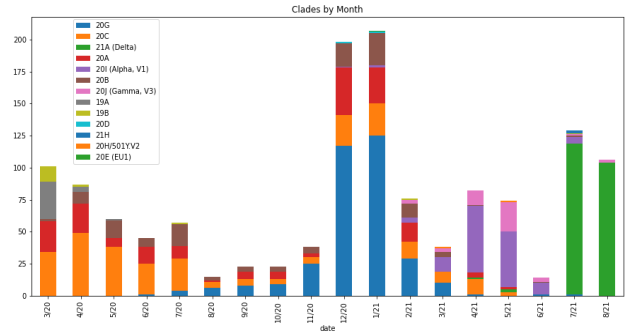


**Figure 2.** Distribution of clades from NU sequencing samples.

## Deterioration Index (DI) between different clades

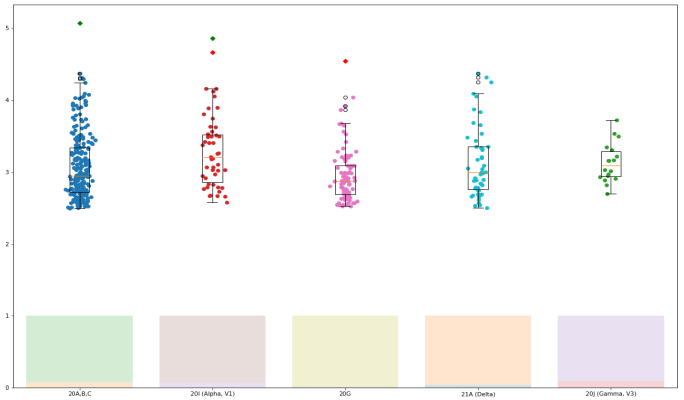


**Figure 3.** *TOP*: The jitter plot of DI scores between each clade. *BOTTOM:* The ratio of ICU vs non-ICU for the associated clade.

The DI score is a proprietary score developed by Epic Systems that utilizes routinely collected data to calculate a risk score that aims to predict which hospitalized patients are more at risk of deterioration and would require higher levels of care. In this graph, statistical significance (Dunn test, p-value < 0.05) is shown between the DI scores of 20I (Alpha variant) compared to 20G and 20A,B,C. The bottom half of the graph shows the proportion of those who entered the ICU within all those infected with the respective

clade. We can see here that 20G has no patients who entered the ICU, whereas every other clade does. Clades 20A,B,C were grouped together due to phylogenetic analysis (not pictured) and the remaining clades were removed due to insignificant sample size.

## Workflow

- Subset data to maximize features and minimize missingness (n=314)
- Multiple Imputation with Chained Equations
- 70/30 stratified shuffle split for training/testing
- GridSearch five-fold cross validation
- Model Fitting

## Results

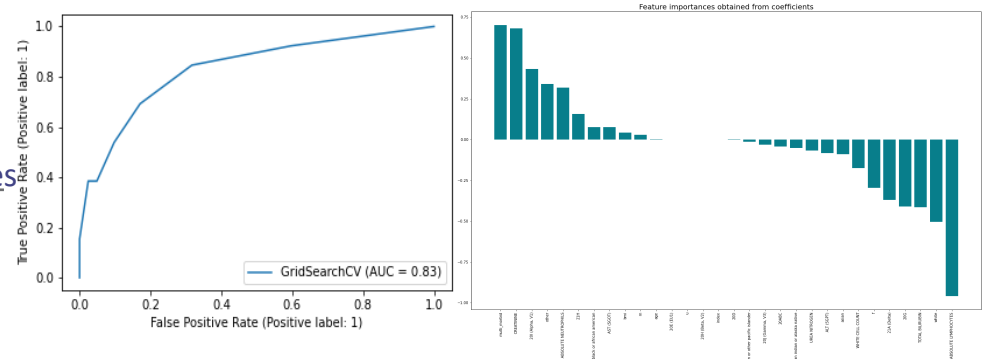


**Figure 4.** *LEFT:* AUC curve of the logistic regression model. *RIGHT:* Coefficients from the logistic regression from highest to lowest.

## Conclusions

- At the height of the pandemic in Chicago, the hospitalization and ICU admittance frequency decreased significantly. This period coincided with the predominance of 20G.
- Further investigation into the differences between 20G and other clades revealed that the DI severity score from Epic Systems revealed statistical significance between 20G and the Alpha variant. 20G also had 0 ICU admittance, and ranked lowest in terms of hospitalization frequency, along with the Delta variant.
- A multivariate logistic regression analysis revealed that using clade, demographic and routinely available lab data were predictive of hospitalization (AUC = 0.83). Within this model, 20G had a significantly negative impact on hospitalization whilst the Alpha variant had a positive impact on hospitalization as measured per their coefficients.
- Further investigation is warranted by using genetic and genomic technologies to understand the evolutionary advantage of 20Gs infectivity yet relatively low pathogenicity.

Northwestern Medicine®
Feinberg School of Medicine