

P8105: Final Project Proposal

11/13/2021

Participants

- Zachary Katz (zak2132)
- Jimmy Kelliher (jmk2303)
- Tanvir Khan (tk2886)
- Tucker Morgan (tln2152)
- Hun Lee (sl4836)

Tentative Title

Exploring the Correlates of COVID-19 Transmission in New York City

Motivation

It's difficult to overstate the extent to which the COVID-19 pandemic has tested the world's public health infrastructure over the past two years. At the same time, the COVID-19 "experience" has manifested unequally – not just country to country, but even city to city. As one of the most heterogeneous urban areas in the world, New York City provides a fascinating case study into the ways in which socioeconomic status may be associated with, or even mediate, disparities in health outcomes. (For instance, it's already well-documented that [income and race](#), along with [socioeconomic privilege and political ideology](#), drive inequities in COVID vaccination rate across US cities.)

Knowing that socioeconomic factors have historically been associated with health outcomes, we aim to start teasing out relationships between a range of potential predictors (e.g. race/ethnicity, high school graduation rate, bachelor degree completion rate, broadband internet access, household income / occupational income score, public vs. private health insurance) and COVID-19 health outcomes – namely, hospitalization, death, and vaccination.

We're curious to understand what kinds of structural barriers may be at play and in what ways. A few questions relevant here may include:

- To what extent does prior use of government services and welfare predict vaccination rate (controlling for socioeconomic status)?
- How much do immigrants' vaccination status correspond to where they moved to NYC, and how long ago, compared to where they emigrated from?
- How does vaccination intent associate vaccination status, and what can this tell us about structural barriers?

Anticipated Final Products

Broadly, we expect our efforts to lead to:

- A set of visualizations that help encapsulate the ways in which covariates affect the COVID-19 health outcomes stated above
- Logistic regression models to ascertain which, if any, of the above covariates (e.g. race/ethnicity) mostly contribute to the correct prediction of COVID-19 health outcomes and assess our model through ROC curves

Anticipated Data Sources

We expect to merge a complex set of data tables, pulled from:

- Demographic and macroeconomic data from the American Community Survey (ACS) 2019 five-year estimate via [IPUMS](#).
- Monthly health outcomes from [NYC Department of Health and Mental Hygiene \(DOHMH\)](#).
- Geographic data of a crosswalk between NYC ZCTAs and PUMAs from [Baruch College](#).
- NYC vaccination rate by race (ages) from [NYC government site](#).
- Broadband Adoption and Infrastructure by Zip Code (Internet access) from [NYC Open DATA](#).
- High school graduation rate and bachelor degree completion rate from [data2go.nyc](#).

Planned Analyses/Visualizations/Coding Challenges

An early challenge we've identified is the need to develop geographical mapping between "PUMA"-level data in the census, zip code level data from DOHMH, and community-district level data from other sources.

In addition, our work will entail creating plots and visualizations of merged data, as well as developing linear models that predict hospitalizations, vaccinations, and death due to COVID-19 from a variety of potential covariates, including socioeconomic status and race/ethnicity. Without doing causal inference, we'll need to be careful not to overstate causal relationships in our findings!

Planned Timeline

- 11/16-11/19: Project Review Meeting
- 11/19-11/24: Merge datasets
- 11/24-11/29: Exploratory analysis
- 11/29-12/4: Regression analysis (modeling)
- 12/4-12/8: Construct report and generate webpage/screencast
- 12/8-12/11: Finishing touches
- 12/11: Report due
- 12/16: "In class" discussion of projects