Are We Testing Those That Need it Most?

Analyzing COVID-19 testing and outcomes across income and race in New York City.

James Buzaid
Cornell University
B.A. Economics and Computer Science
Class of 2022
jjb422@cornell.edu

July 7, 2020

Abstract

New York City has seen a great deal of COVID-19 cases and deaths since the first positive test in March. This paper analyzes ZIP-code level COVID-19 outcomes with 2018 American Community Survey neighborhood data. I find residents of poor ZIP-codes and ZIP-codes with a greater fraction of the population identifying as black are more likely to test positive, more likely to die, and less likely to get tested for the virus. Such raises questions of how the virus spreads, how testing is allocated, and whether the cost of testing is different across different income groups

Contents

1 Introduction

In March and April of 2020, New York City had a large number of positive COVID-19 cases when compared to other cities. After beginning testing in March, by April 4th, 2020, New York City had 60,435 people test positive for the virus with a total of 104,096, meaning over 50% of tests were returned positive. As the city reopens in June and July the percent of cases returned positive has fallen and far more people have been tested. By June 22nd, 2020, 202,235 people tested positive out of 985,862 tests, meaning roughly 20% of cases returned positive. The true incidence of COVID-19 in the city is hard to determine with policies that encourage testing but ultimately keep it voluntary. More random sample tests would be conducted, like the random sample antibody test in April that showed roughly 20% of people leaving grocery stores testing positive for antibodies. Selection bias may have been a factor in this high positive rate, but it shows that the number of COVID-19 positives per capita may have been higher than expected. It is important to gauge how testing should be allocated for the safety of others. COVID-19 has been shown to disproportionality affect poorer communities [1]. A factor here is occupation: poorer individuals are more likely to work in occupations that do not allow for adequate social distancing [3]. Keeping these correlations in mind, if a second wave were to strike New York, how should testing be allocated going forward? This study addresses this question using COVID-19 ZIP-code level data and 2018 Five Year American Community Survey results. This study also analyzes how different socioeconomic groups fare with the virus and looks at the economic costs of testing.

The NYC Department of Health and Mental Hygiene (DOHMH) has been publishing COVID-19 data for each Modified Zip Code Tabulation Area (177 in New York City) since March 30th. The data includes COVID-19 positive tests, deaths, and total tests. I merge this data with 2018 Five Year American Community Survey findings to find indicators of COVID-19 testing disparities and disparities in COVID-19 outcomes across different

ZIP-codes. [1]

On June 14th, NYC DOHMH first published death counts on the ZIP-code level. This study uses these counts to come to conclusions about which socioeconomic groups are more likely to pass away from the virus. Additionally, this study attempts to reach some idea of how the true infection rate varies for different socioeconomic groups. While positive tests per capita does not give the best indication for a true infection rate, as the amount of testing has increased it becomes a better estimate for the true rate of infection, assuming people are not being tested multiple times. As of June 22nd, 985,862 tests have been conducted for COVID-19. Assuming each person has been tested once, this is a large portion of New York's population of eight million.

New York City's COVID-19 testing has not been distributed uniformly across ZIP-codes. While New York City COVID-19 testing is free, more testing is received in ZIP-codes that are more white and wealthier. There is likely a role being played by the number of testing centers in a neighborhood as well, even though some private testing centers like CityMD have made their testing free. There are other paid COVID-19 testing centers however and not all COVID-19 testing is free. Additionally, antibody tests, albeit less accurate, can be expensive. Many antibody test visits without insurance cost upwards of $300, although price transparency is hard to come by. Still, testing per capita for each ZIP-code is positively correlated with deaths per capita and with positive tests per capita. Additionally, socioeconomic variables including a ZIP-code's median income are good determinants of the positive cases per capita. This study looks at, accounting for other covariates, how testing per capita is allocated.

2 Data

The NYC Department of Health and Mental Hygiene (DOMNH) publishes data for COVID-19 by Modified ZCTA on the NYC Health GitHub page. The file data-by-modzcta.csv from 6/22/2020 was used in this study. It contains information for the total number of tests, the total number of positive tests/cases, and the total number of confirmed deaths (including comorbidity), all by modified ZCTA (of where the person resides). The file

[1] A Modified Zip Code Tabulation Area (Modified ZCTA) is NYC's way of combining ZCTA's as to allow a reliable estimate of the population of an area and removing ZIP codes with zero population (Midtown, Wall Street, etc.). ZIP-codes are a collection of points and ZCTA's triangulate these points. ZCTA's and ZIP-codes are interchangeable. Most Modified ZCTA's are equivalent to their respective ZCTA. For instance, ZCTA 10003 is the same as Modified ZCTA 10003. A crosswalk from ZCTAs / ZIP-codes to Modified ZCTAs is provided here https://github.com/buzaidj/coronavirus-data/blob/master/Geography-resources/ZCTA-to-MODZCTA.csv, on the NYC Health github page. The 177 "ZIP-codes" I refer to in my study are actually modified ZCTA codes.

tests-by-zcta from 4/4/2020 was also used for comparison. It only contains information for the total number of tests and the total number of positive tests/cases but does not provide death data.

Death estimates by ZIP-code were recently added to NYC's COVID-19 data. Most COVID-19 deaths occur in people with co-morbid conditions. Death data may be underreported for those that die soon after entering an emergency room or die at home from COVID-19, since autopsies for it are rather rare. Deaths from COVID-19 are ruled probable or confirmed, however ZIP-code level data only takes into account confirmed deaths.

2018 5-year American Community Survey ZIP-code level data was used for all demographic variables in the study. [2] The API variables used are:

1. B02001_001E — Population Estimate
2. B02009_001E — Estimate of population that is black or black and other races
3. B02001_002E — Estimate of population that is white alone
4. B01002_001E — Estimate of median age
5. B19326_001E — Estimate of median income (in the past 12 months)
6. B08201_001E — Estimate of number of housing units

Nursing home and assisted care facility death data was retrieved from here with June 14th, 2020 data. Tabula was used to extract the tables from PDF's. Google's geocoding API was used to retrieve ZIP-code information for each nursing home to retrieve data from nursing home / ACF deaths by ZIP-code in NYC.

A quick look at testing in figure 1 shows testing is not evenly distributed across different ZIP-codes. Summary statistics for variables are given by Table 1.

3 Results

3.1 Deaths per Capita and Tests per Capita

ZIP codes with more deaths per capita have more tests per capita in the interest of public safety. There is a positive relationship between number of deaths and number of tests conducted. However, by the nature of the exponential growth of epidemics, the expectation might be that tests should be exponentially higher in neighborhoods with more deaths. However, this does not appear to be the case, as shown above. Regardless, areas with higher death rates consume more tests. Additionally, number of deaths is very positively correlated with positive deaths per capita. See Figure 2.

[2] https://api.census.gov/data/2018/acs/acs5.html

	Min	5th	25th	Median	75th	95th	Max
Population (1000s)	3.028	8.4694	27.726	43.703	67.094	94.503	112.425
Percent black or black and other races	0.004	0.015	0.042	0.110	0.368	0.751	0.936
Average household size	1.595	1.848	2.338	2.730	3.026	3.523	4.068
Median Income	15,778	17,829	25,773	32,027	44,599	87,642	147,538
Median Age	27.8	30.4	34.2	37.1	40.6	45.5	49.4
Tests per 100,000 (6/22)	3668	8360	10233	11249	12910	15517	33616
Positives per 100,000 (6/22)	597	911	1626	2328	2913	3640	4792
Tests per 100,000 (4/4)	111	490	931	1153	1479	2342	6387
Positives per 100,000 (4/4)	38	247	473	621	795	1452	3919
Deaths per 100,000 (6/22)	0	42	115	180	239	384	612
Percent Positive (6/22)	7.4%	8.9%	15.7%	20.6%	24.5%	27.4%	29.6%
NH/ACF Deaths (6/14)	0%	0%	0%	0%	23%	88%	185%

Table 1: Summary Statistics for Variables

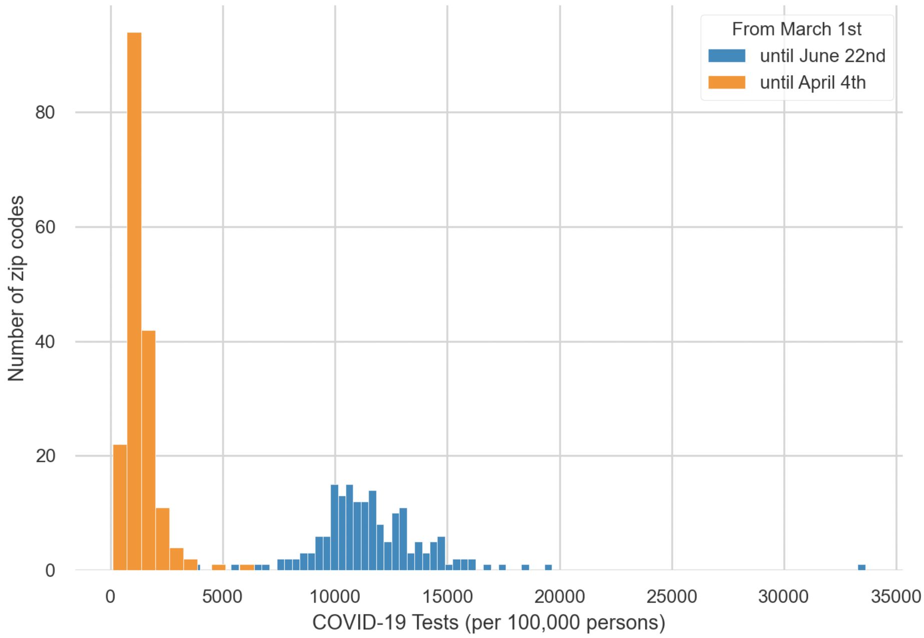

Figure 1: COVID-19 Tests per 100,000 persons

3.2 Outcome Regressions with American Community Survey data

It has held since COVID-19 data was first released on the ZIP code level that those tested in poorer areas were more likely to test positive for COVID-19, even though there is a great deal of variation between ZIP codes. Figure 3 shows the relationship between lower income and disparate COVID-19 outcomes. All outcome regressions are weighted least squares regressions, with data weighted by population.

There are multiple explanations for this phenomenon. One is that the rate of infection is different across income groups. Another considers the economic costs of testing. Poorer individuals might choose to get tested only when sick, especially with the high demand for testing creating wait times for it, explaining the higher positive rate in poorer ZIP codes. Explaining how this holds with a roughly uniform distribution of testing across ZIP codes by income level is harder but omitted variable bias may be at play. One explanation however is that those in poorer zip codes may get tested if they are not sick, out of fear of contracting the disease. Regardless, even if the rate of infection was the same across income groups in April, it may not be the case as of June 22nd considering the increase in positive tests per capita for a decrease in median income as well as a greater number of deaths in poorer communities seen in Figure 6.

Additionally, even with strong social distancing measures in place, in-

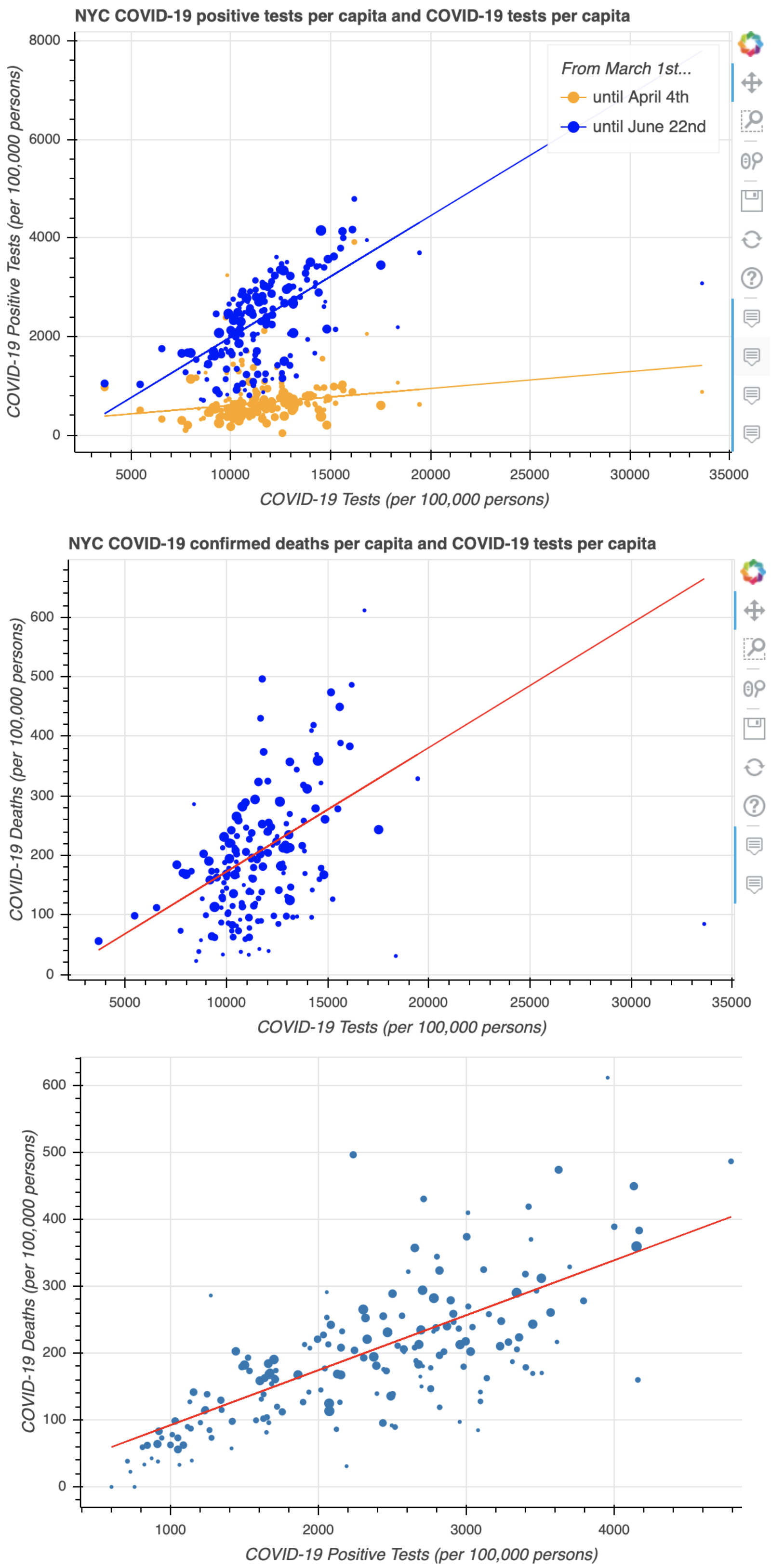

Figure 2: Distribution of testing and deaths by outcomes

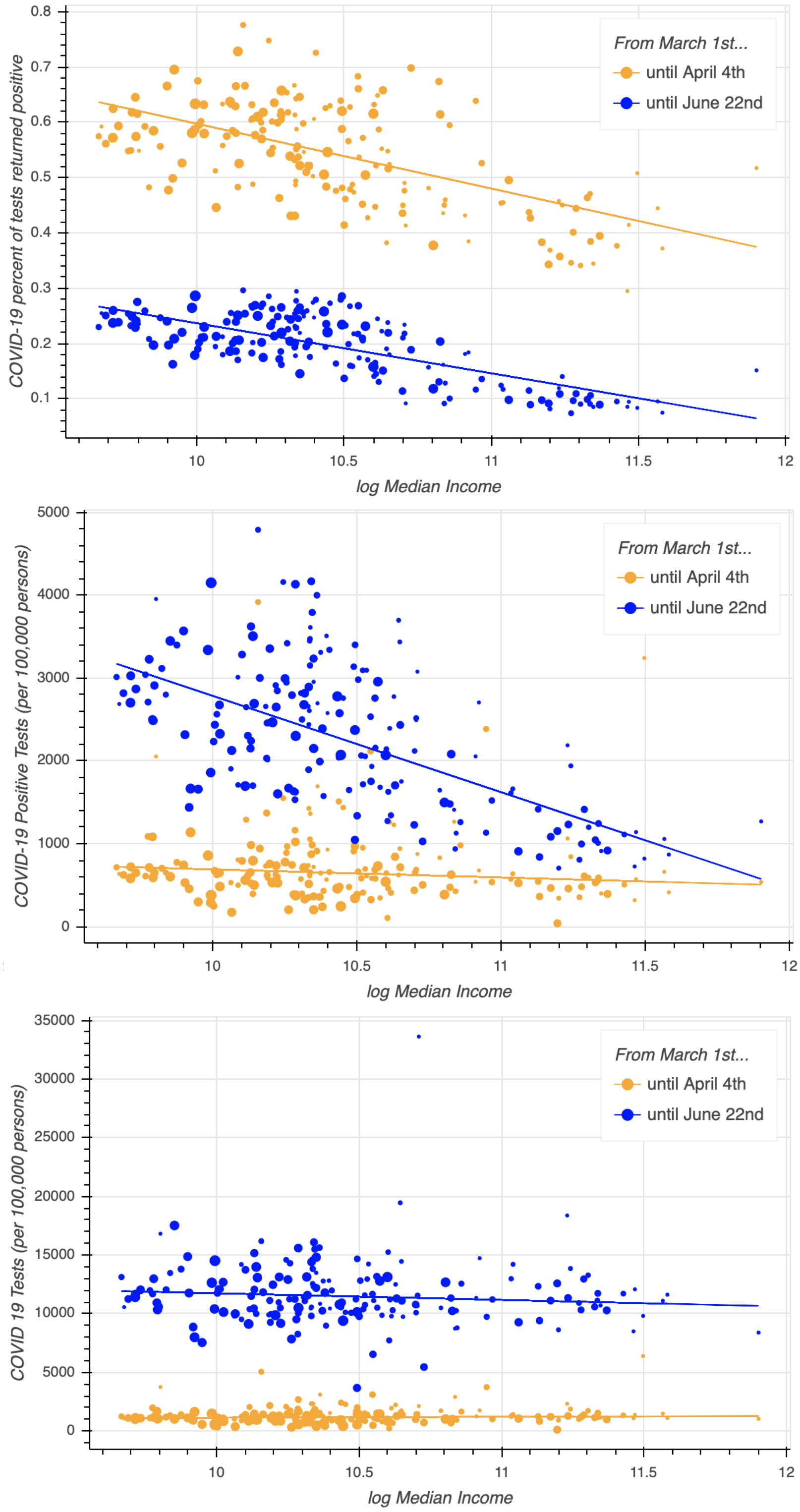

Figure 3: Outcomes by income

trahousehold spread is hard to prevent. If one person in a household has COVID-19, everyone in the household could get it. Such might explain some positive relationship between average household size and the percent of COVID-19 tests returned positive. With more tests coming in as of June, average household size, a measure of density, has become a good predictor of positive tests per capita as well, even though there was little correlation in April. Somewhat surprisingly, testing is distributed fairly uniformly across ZIP codes with differing average household size, even though denser ZIP-codes are more likely to test positive. Figure 4 shows COVID-19 outcomes by average household size in a ZIP-code.

Average household size is unsurprisingly somewhat correlated to median income. Larger households economically make sense as a means of subsidizing rent for poorer tenants and making household labor cheaper. Figure 5 shows average household size by income.

Even assuming COVID-19 has some fixed death rate for a positive case across socioeconomic factors, the expectation is that ZIP-codes with lower incomes and higher average household sizes would have higher death rates with more infected. Deaths are correlated with income and average household size, albeit slightly loosely. However, controlling for omitted variables such as how bad the pandemic was in certain areas very early, a ZIP-codes median age, and how many nursing home and assisted care facility deaths is needed. Figure 6 shows death data by income and by average household size.

3.3 Multivariate Regressions

Regressions with many American Community Survey ZIP code level variables that may impact COVID-19 deaths are shown below. Tables 2, 3, and 4 give weighted least squares regressions, predicting deaths per 100,000 persons, positives per 100,000 persons, and tests per 100,000 persons.

In predicting deaths per 100,000 persons, average household size is significantly negative. In the single variable case, it appeared that average household size was somewhat positively correlated with deaths and positive cases. Average household size remains a positive predictor of positive cases per capita however. Perhaps when accounting for covariates like income, larger average household sizes tend to indicate more families with young people who are very unlikely to die from the disease if they contract it.

Additionally, when accounting for income and other covariates, the percent of a ZIP-code that is black is positively correlated with deaths and positive tests, and negatively correlated with the number of tests conducted. Income also is positively correlated with deaths and positive tests, and negatively correlated with deaths.

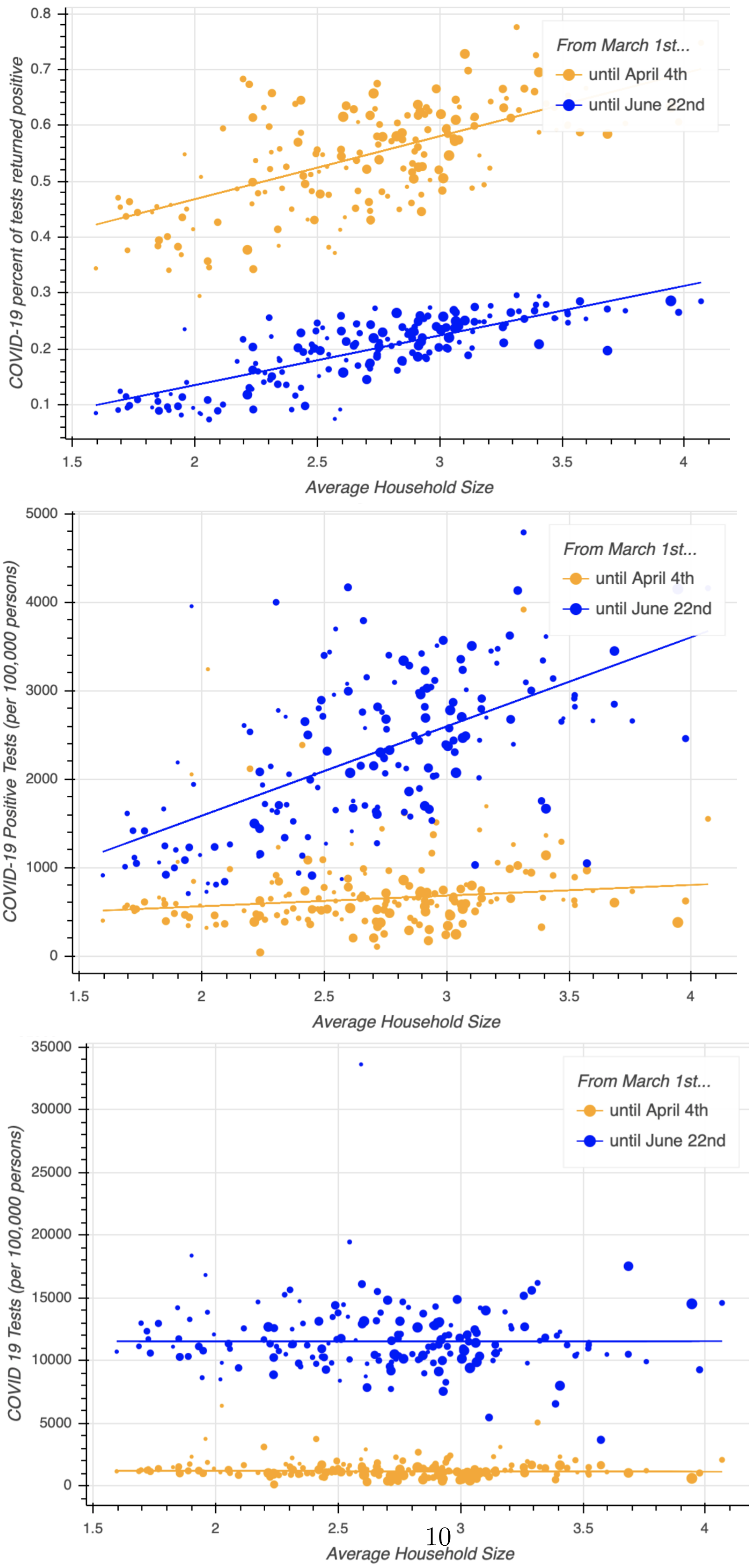

Figure 4: Outcomes by average household size

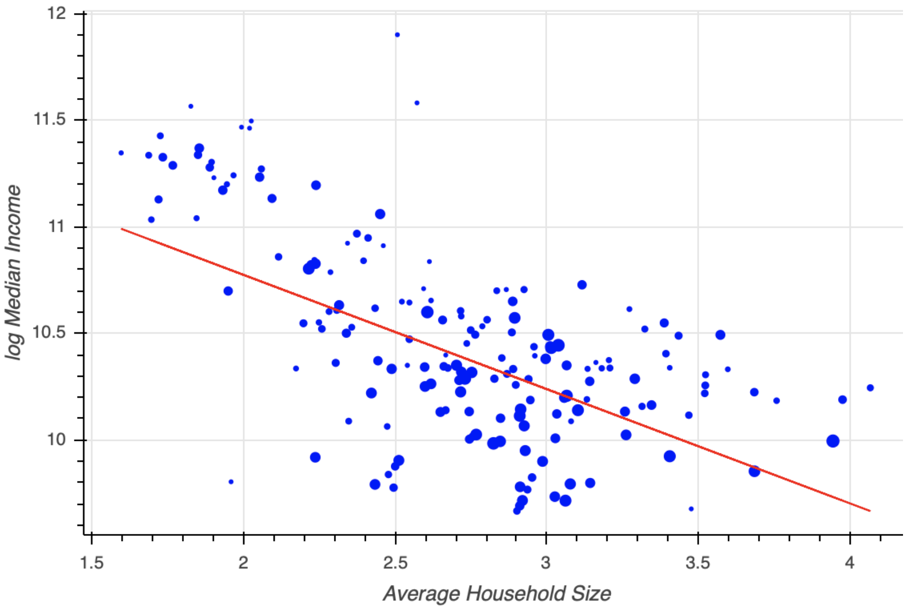

Figure 5: Income by average household size

Model:	WLS (Weighted by Population)	Adj. R-squared:	0.731
Dependent Variable:	COVID-19 Deaths (per 100,000 persons)	AIC:	1923.1444
Date:	2020-06-26 16:28	BIC:	1951.7298
No. Observations:	177	Log-Likelihood:	-952.57
Df Model:	8	F-statistic:	53.40
Df Residuals:	168	Prob (F-statistic):	2.83e-42
R-squared:	0.743	Scale:	1.1432e+08

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
Constant	780.6902	165.5068	4.7170	0.0000	456.3029	1105.0775
Percent of population identifying as black or black and other races	41.7470	19.3329	2.1594	0.0308	3.8553	79.6387
Average Household Size	-40.3288	13.9340	-2.8943	0.0038	-67.6390	-13.0187
log Median Income	-69.7237	15.6699	-4.4495	0.0000	-100.4361	-39.0113
Median Age	1.5267	1.2063	1.2656	0.2056	-0.8376	3.8910
NH/ACF Deaths	0.8494	0.1789	4.7491	0.0000	0.4989	1.2000
COVID-19 Positive Tests until June 22nd (per 100,000 persons)	0.0605	0.0133	4.5498	0.0000	0.0344	0.0866
COVID-19 Tests until June 22nd (per 100,000 persons)	0.0013	0.0033	0.3931	0.6943	-0.0052	0.0078
COVID-19 Positive Tests until April 4th (per 100,000 persons)	0.0186	0.0112	1.6611	0.0967	-0.0034	0.0406

Omnibus:	7.393	Durbin-Watson:	1.571
Prob(Omnibus):	0.025	Jarque-Bera (JB):	8.126
Skew:	0.346	Prob(JB):	0.017
Kurtosis:	3.789	Condition No.:	479455

Table 2: COVID-19 deaths per 100,000 persons

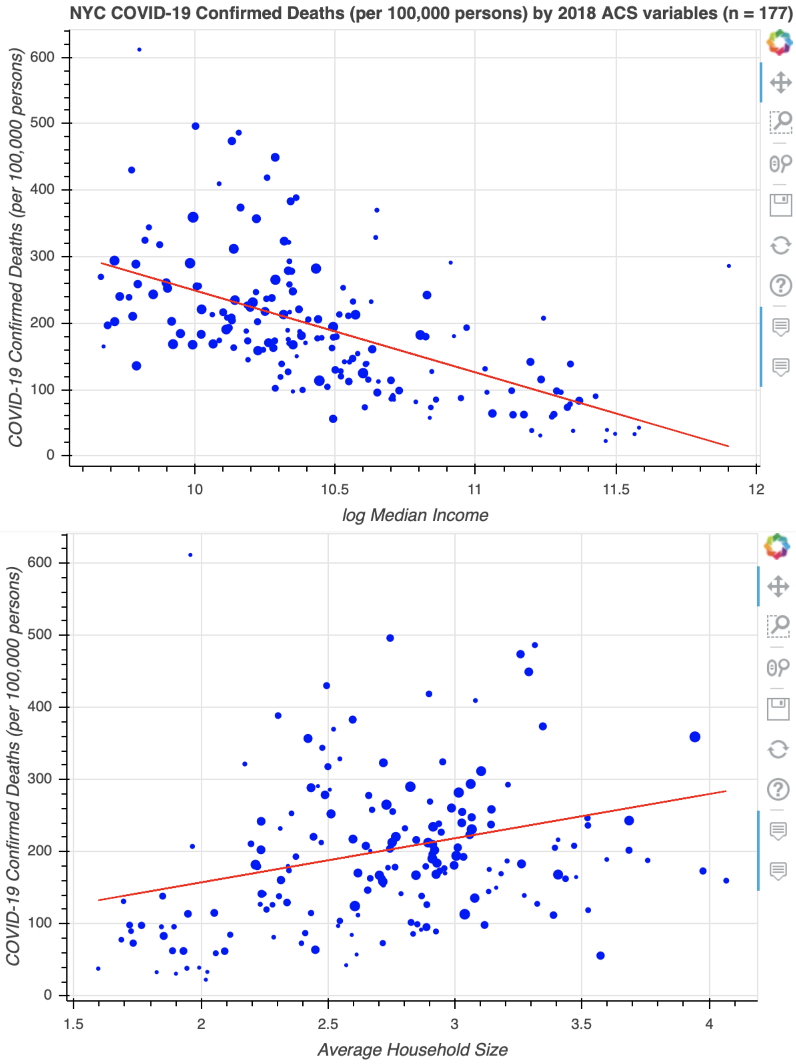

Figure 6: Deaths per capita

Model:	WLS (Weighted by Population)	Adj. R-squared:	0.886
Dependent Variable:	COVID-19 Positive Tests (per 100,000 persons)	AIC:	2547.2367
Date:	2020-06-26 16:28	BIC:	2575.8220
No. Observations:	177	Log-Likelihood:	-1264.6
Df Model:	8	F-statistic:	220.2
Df Residuals:	168	Prob (F-statistic):	7.15e-85
R-squared:	0.891	Scale:	3.8851e+09

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
Constant	-184.0532	855.4955	-0.2151	0.8297	-1860.7936	1492.6872
Percent of population identifying as black or black and other races	470.6636	88.4280	5.3226	0.0000	297.3479	643.9793
Average Household Size	708.2946	56.6421	12.5047	0.0000	597.2782	819.3110
log Median Income	-332.1375	86.5399	-3.8380	0.0001	-501.7526	-162.5225
Median Age	28.1115	7.3382	3.8309	0.0001	13.7289	42.4940
NH/ACF Deaths	-2.5462	0.9197	-2.7687	0.0056	-4.3487	-0.7437
COVID-19 Deaths until June 22nd (per 100,000 persons)	2.0568	0.6032	3.4098	0.0007	0.8746	3.2391
COVID-19 Tests until June 22nd (per 100,000 persons)	0.2093	0.0229	9.1284	0.0000	0.1644	0.2543
COVID-19 Positive Tests until April 4th (per 100,000 persons)	0.1430	0.0635	2.2521	0.0243	0.0185	0.2675

Omnibus:	13.189	Durbin-Watson:	1.106
Prob(Omnibus):	0.001	Jarque-Bera (JB):	20.881
Skew:	-0.408	Prob(JB):	0.000
Kurtosis:	4.471	Condition No.:	508300

Table 3: COVID-19 positive tests per 100,000 persons

Model:	WLS (Weighted by Population)	Adj. R-squared:	0.771
Dependent Variable:	COVID-19 Tests (per 100,000 persons)	AIC:	3020.5677
Date:	2020-06-26 16:28	BIC:	3049.1531
No. Observations:	177	Log-Likelihood:	-1501.3
Df Model:	8	F-statistic:	89.74
Df Residuals:	168	Prob (F-statistic):	1.26e-56
R-squared:	0.781	Scale:	5.6337e+10

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
Constant	-227.0765	2656.1029	-0.0855	0.9319	-5432.9426	4978.7896
Percent of population identifying as black or black and other races	-2267.7444	373.5868	-6.0702	0.0000	-2999.9611	-1535.5277
Average Household Size	-2304.6881	222.5774	-10.3545	0.0000	-2740.9318	-1868.4443
log Median Income	1583.8174	251.0124	6.3097	0.0000	1091.8422	2075.7926
Median Age	-133.5910	30.8466	-4.3308	0.0000	-194.0493	-73.1327
NH/ACF Deaths	6.3754	3.2750	1.9467	0.0516	-0.0434	12.7942
COVID-19 Deaths until June 22nd (per 100,000 persons)	0.6409	1.5481	0.4140	0.6789	-2.3933	3.6752
COVID-19 Positive Tests until June 22nd (per 100,000 persons)	3.0355	0.1640	18.5046	0.0000	2.7140	3.3570
COVID-19 Positive Tests until April 4th (per 100,000 persons)	-0.3140	0.2053	-1.5299	0.1260	-0.7164	0.0883

Omnibus:	89.628	Durbin-Watson:	1.712
Prob(Omnibus):	0.000	Jarque-Bera (JB):	570.446
Skew:	1.781	Prob(JB):	0.000
Kurtosis:	11.041	Condition No.:	112531

Table 4: COVID-19 tests per 100,000 persons

4 Conclusions

This study's results confirm past research that poorer neighborhoods and neighborhoods with more black people are more likely to test positive and are more likely to have a recorded death from the virus. While the true number of people infected is hard to infer, and the true number of deaths may be higher or lower than recorded, poorer ZIP-codes and ZIP-codes with more black people are more likely to experience disparate outcomes as a result of the virus. However, socioeconomic groups that are more likely to experience disparate outcomes seem to be tested at equal or lesser rates than socioeconomic groups that do not.

More positive cases per capita in poorer areas may have to do with the costs of quarantine for different socioeconomic groups. Jobs that can be done from home, from a computer, also tend to be higher paying. Restaurant work, construction work, and other fields that pay little and were deemed essential during the pandemic are near impossible to be done remotely. Almagro [3] has done more research on this.

There are several explanations that may exist together to explain less testing per capita in poorer and more black areas. These require more research. One explanation is that with more cases in an area, the likelihood that one contracts COVID-19 at a testing center in that area is higher. Since testing is entirely voluntary, it is also likely that more testing is received by sick people with COVID-19 like symptoms. Both factors contribute to those that are not sick and do not show symptoms opting to not get tested. Another explanation that explains disparate outcomes across different income levels is that the economic cost of testing may be higher for those with lower incomes. Poorer people have less access to paid testing services and are less likely to have insurance cover such costs. Also, even though city supplied testing and CityMD testing is free, testing does take time. Additionally, poorer people may be less informed about the risks COVID-19 carries. More research is needed to reach conclusions here.

References

[1] Borjas, G. *Demographic Determinants of Testing Incidence and COVID-19 Infections In New York City Neighborhoods.* Working Paper 26952. April 2020.

[2] Emeruwa, N., et. al. *Associations Between Built Environment, Neighborhood Socioeconomic Status, and SARS-CoV-2 Infection Among Pregnant Women in New York City.* JAMA. June 2020.

[3] Almagro, T., Orane-Hutchinson, A. *The determinants of the differential exposure to COVID-19 in New York City and their evolution over time.* June 2020.