HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Demograpic Determinants of Testing Incidence and COVID-19 Infections in New York City Neighborhoods

Faculty Research Working Paper Series

George J. Borjas
Harvard Kennedy School

April 2020

RWP20-008

Visit the **HKS Faculty Research Working Paper Series** at:
https://www.hks.harvard.edu/research-insights/publications?f%5B0%5D=publication_types%3A121

The views expressed in the **HKS Faculty Research Working Paper Series** are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. Faculty Research Working Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s). Papers may be downloaded for personal use only.

Demographic Determinants of Testing Incidence and COVID-19 Infections in New York City Neighborhoods

George J. Borjas
Harvard University

April 7, 2020

Abstract: New York City is the hot spot of the COVID-19 pandemic in the United States. This paper merges information on the number of tests and the number of infections at the New York City zip code level with demographic and socioeconomic information from the decennial census and the American Community Surveys. People residing in poor or immigrant neighborhoods were less likely to be tested; but the likelihood that a test was positive was larger in those neighborhoods, as well as in neighborhoods with larger households or predominantly black populations. The rate of infection in the population depends on both the frequency of tests and on the fraction of positive tests among those tested. The non-randomness in testing across New York City neighborhoods indicates that the observed correlation between the rate of infection and the socioeconomic characteristics of a community tells an incomplete story of how the pandemic evolved in a congested urban setting.

Demographic Determinants of Testing Incidence and COVID-19 Infections in New York City Neighborhoods

George J. Borjas*

1. Introduction

The apparent difficulty in managing the COVID-19 pandemic raises a multitude of research questions that need to be answered in order to minimize both the human suffering and the economic costs. Why did hot spots arise in some cities or countries, but not in others? Is the pandemic disproportionately affecting particular demographic or socioeconomic groups? Can the destructive path of the pandemic be effectively controlled by a more careful targeting of the scarce testing resources? What is the net cost of the restrictions on geographic mobility, work arrangements, and social gatherings that are now common in countries affected by the pandemic? Addressing all of these questions, however, requires the availability of detailed data that would allow a researcher to search for systematic empirical patterns that might suggest an effective path towards a reduction in future costs.[1]

The New York metropolitan area is currently a hot spot of the COVID-19 pandemic in the United States. The first COVID-19 test in the city was conducted on January 29, 2020, with the first positive result not confirmed until February 23, 2020. Nevertheless, by April 6, 2020, New York City had already identified 68,776 persons infected with the virus, and 2,738 persons had succumbed to the infection.

---

* Robert W. Scrivner Professor of Economics and Social Policy, Harvard Kennedy School; Research Associate, National Bureau of Economic Research; and Program Coordinator, IZA Program on Labor Mobility, Institute for the Study of Labor. I am grateful to Hugh Cassidy, Daniel Hamermesh, and Gordon Hanson for helpful comments and suggestions.

[1] Economists have already begun to analyze many of the questions raised by the pandemic; see, for example, Baker et al (2020), Bergen, Herkenhoff, and Mongey (2020), Harris (2020, and Lang, Wang, and Yang (2020).

Beginning on March 30, 2020, the NYC Department of Health and Mental Hygiene began to release detailed information on the number of tests administered and the number of positive results for persons residing in each of 177 zip codes, which I will roughly interpret as urban "neighborhoods."[2] Note, however, that the average population in a zip code in New York City is around 46,000 persons (so that the typical New York neighborhood actually has the population of a small city). This paper exploits the data to identify the demographic characteristics that correlate with the very different patterns in COVID-19 testing and infections across neighborhoods.

Specifically, I merged the neighborhood-level counts of tests administered and positive results with data from both the 2010 decennial census and the 2010-2014 American Community Surveys (ACS). The merging allows me to paint a detailed picture of the neighborhoods where COVID-19 testing was more prevalent and where the number of positive cases was relatively high.

The evidence suggests that a small number of neighborhood characteristics help explain a relatively large fraction of the variance (over 70 percent) in positive test results (conditional on being tested) across New York City neighborhoods. In particular, the conditional probability of a positive test result is far greater for persons living in poor neighborhoods, in neighborhoods where large numbers of people reside together, and in neighborhoods with a large black or immigrant population. At the same time, however, persons residing in poor or immigrant neighborhoods were less likely to be tested.

Much of the discussion over the spread and impact of the COVID-19 epidemic focuses on the evolution of a single statistic: the number of infections per 100,000 persons in the

[2] The data file, however, does not report the number of fatalities resulting from the pandemic.

population. The analysis of the New York City data starkly illustrates the problem with focusing on this single number. The rate of infection in the population depends on two distinct factors: the frequency of tests in a particular area and the fraction of positive tests among those tested.

Because the incidence of COVID-19 testing was not random across New York City neighborhoods, it is crucial to examine how socioeconomic characteristics correlate with each of the two determinants of the rate of infection. It turns out that some characteristics—for instance, household income—are correlated in opposite directions with the incidence of testing and with the likelihood that a test leads to a positive result. In the end, we might find little correlation between household income and infection rates, but this "zero correlation" masks the fact that household income is related to *both* the rate of testing and the likelihood that a test yields a positive result. Persons residing in poorer neighborhoods were less likely to be tested, but once the test was administered, were more likely to be afflicted with the virus. In short, the finding that the rate of infection in the population is uncorrelated with a particular socioeconomic characteristic may not provide the information that is required to understand the evolution of the pandemic.

2. Data

The key data file analyzed in this paper reports the cumulative number of COVID-19 tests administered as well as the cumulative number of positive results for each zip code in the City of New York.[3] These data were first released on March 30, 2020 by the NYC Department of Health and Mental Hygiene (and have been updated since). Unless otherwise noted, the analysis in this paper uses the cumulative counts as of April 5, 2020. The data consist of 177

[3] The New York City zip codes are numbered from 10001 to 11697.

identifiable zip codes, with some persons being allocated to a "non-identifiable" category that is excluded from the empirical exercise that follows.

It is important to emphasize that the zip code in the publicly available data refers to the *zip code of residence* for the person who received the test. For expositional convenience, I will use the terms "zip code" and "neighborhood" interchangeably.

Figure 1, prepared by the New York City Department of Health, shows the geographic spread in infection across neighborhoods. Even a superficial look at the map suggests that many neighborhoods in Manhattan have had relatively few COVID-19 cases, while some neighborhoods in the Bronx, Brooklyn, or Queens have seen a very high case load. Some of the initial media reports of these data provide a detailed discussion of the pandemic in specific neighborhoods (Honan, 2020; Buchanan et al, 2020; and Marsh 2020).

The first two rows of Table 1 summarize key characteristics of the data. The cumulative number of persons who had been tested for COVID-19 infection in the average zip code prior to April 5, 2020 was 598.2, but there was a very large variation across neighborhoods in the incidence of testing. Around 219 persons had been tested in the neighborhood at the 10^{th} percentile, and almost 5 times as many (1013) had been tested in the neighborhood at the 90^{th} percentile. The table also shows a similarly large geographic dispersion in the number of persons who tested positive for the virus, from 101 persons in the neighborhood at the 10^{th} percentile to 624 persons in the neighborhood at the 90^{th} percentile.

I merged the data released by the NYC Department of Health with information from other sources, specifically the 2010 decennial census and the 2010-2014 American Community Surveys (ACS). This merging allows me to obtain a relatively simple description of the demographic and socioeconomic characteristics for each of the 177 neighborhoods.

The demographic data from the 2010 decennial census at the zip code level is publicly available in the data archives maintained by the Census Bureau.[4] The characteristics that I downloaded from the decennial census are: the population of the neighborhood, the fraction of the population that is male, the fraction of the population that is over age 60, the fraction of the population that is minority (i.e. Hispanic, non-Hispanic black, or non-Hispanic Asian), and the number of persons in the average household. Table 1 shows that all of these demographic characteristics vary substantially across NYC neighborhoods. For instance, the mean household size is 2.6 persons, but the range between the 10th and 90th percentile neighborhood goes from 1.7 to 3.2 persons. Similarly, the range for the fraction of the neighborhood's population that is black goes from 1.3 to 52.1 percent, with a mean of 16.1 percent.

I use the neighborhood's population to normalize the data on COVID-19 testing and infections. Specifically, Table 1 also reports both the number of tests and the number of infections per 100,000 persons. In the 10th percentile neighborhood, the incidence of testing was 902.9 persons, while the incidence was twice as large, or 1840.6, in the 90th percentile neighborhood. These large geographic differences in the incidence of testing will play a role in the analysis that follows.

The table also reports the rate of infection per 100,000 persons, calculated simply as the number of persons infected with COVID-19 divided by population (times 100,000). This number also differs substantially across neighborhoods, from 442.2 in the 10th percentile neighborhood to 1107.4 in the 90th percentile neighborhood.

[4] The data are available at census.data.gov.

Finally, the table reports the percent of tests that confirmed an infection. The fraction of tests that had positive results was 54.0 percent in the average neighborhood, but it ranged from 39.4 in the 10th percentile neighborhood to 64.3 percent in the 90th percentile neighborhood.

Unfortunately, the zip code-level data available from the decennial census does not contain information on the income level of the neighborhood or on other socioeconomic variables that might be relevant in understanding the spread of the pandemic. I used data from the merged 2010-2014 ACS to obtain this type of additional information. In particular, I used the ACS to calculate the mean household income of the neighborhood and the fraction of the neighborhood's population that is foreign-born.[5]

Table 1 also reports the summary statistics for these two additional variables. As with the data drawn from the decennial census, there is a great deal of dispersion across neighborhoods in these characteristics. For example, the range for mean household income goes from $46,800 in the 10th percentile to $164,000 in the 90th percentile, and the range for the fraction of the neighborhood's population that is foreign-born goes from 19.3 to 52.5 percent.

Let V_i be the number of persons in neighborhood i who tested positive for COVID-19, and let P_i be the population of the neighborhood. The fraction of persons in a neighborhood who tested positive (which can be easily translated into the number of persons infected with the virus per 100,000 people) can be written as:

$$\frac{V_i}{P_i} = \left(\frac{T_i}{P_i}\right) \times \left(\frac{V_i}{T_i}\right), \tag{1}$$

[5] These data are downloadable from the usa.com website; http://www.usa.com/rank/.

where T_i gives the number of residents in the neighborhood who have been tested.

Equation (1) shows that the infection rate in the population (V_i/P_i) depends on two distinct ratios: (1) the rate at which persons were tested; and (2) the rate at which those who were tested, in fact, tested positive for the virus. Testing in the United States was very sporadic in the early stages of the pandemic. Even in a city as large as New York City, fewer than 1,000 persons had been tested by March 10, 2020, and only 110,606 persons (out of a population of 8.6 million) had been tested by April 5, 2020.[6]

Equation (1) also shows that the infection rate in the population depends on the fraction of tests that turn out to be positive. This variable, too, is likely to be correlated with specific socioeconomic characteristics (for example, there may be more infections among people living in neighborhoods where large families cluster together).

It is unclear at the outset whether a given socioeconomic variable x should be similarly correlated (even in sign) with each of the two ratios in the right-hand-side of equation (1). As a result, the correlation between the population infection rate (V_i/P_i) and x may confound the two very different roles that x might be playing in the pandemic.

Let me conclude with an important caveat. The demographic and socioeconomic characteristics chosen to describe the neighborhood were selected not only because they tend to be the usual suspects used in many studies in the social science literature, but also because it is unlikely that they affected the evolution of test frequency and testing results during the month of March 2020.

---

[6] The cumulative number of tests for March 10 was reported in the testing.csv file prepared by the NYC Department of Health and available at https://github.com/nychealth/coronavirus-data. The file was downloaded on April 2, 2020, but had disappeared from the website by April 5.

It is not difficult to find neighborhood characteristics that likely affected the allocation of testing resources as the pandemic took hold in the city. By the middle of March 2020, for instance, it became well known that the large Hassidic community in Brooklyn was particularly hard hit by the virus (Stack and Schweber, 2020; Sales, 2020). The Hassidic population in New York City is concentrated in a small number of neighborhoods (Gallagher, 2009).[7] After learning of this specific problem, NYC health officials might have responded to the outbreak by reallocating resources to increase the incidence of testing in the affected neighborhoods. Such a behavioral response would lead to an increased number of tests in Hassidic neighborhoods, with a likelihood that many of those tests were positive.

A positive correlation between, say, the fraction of a neighborhood's population that is Hassidic and the fraction of positive COVID-19 tests would then measure a response to an actual outbreak rather than a pre-existing "predisposition." The socioeconomic variables chosen for the analysis and summarized in Table 1 were, in part, chosen because they tend to avoid this endogeneity problem.[8]

3. Descriptive evidence

Before proceeding to the regression analysis, it is instructive to illustrate graphically the simple correlation between some of the neighborhood characteristics and both the incidence of testing and the likelihood of a positive test result. Figure 2, for example, presents scatter diagrams that show how differences in the mean household income of the neighborhood's

[7] The main zip codes are 11211, 11218, 11219, 11204, and 11230, which include the Williamsburg and Borough Park sections of Brooklyn.

[8] One potential exception is the variable that gives the fraction of the population that is over age 60. As the pandemic progressed, it became apparent that the elderly were severely affected by the illness and testing resources may have been reallocated to neighborhoods that had a disproportionately large number of older persons.

population relate to: (a) the incidence of testing per 100,000 persons ($V_i/P_i \times 100{,}000$); (b) the percent of tests that confirm an infection ($V_i/T_i \times 100$); and (c) the number of infections per 100,000 persons ($V_i/P_i \times 100{,}000$).

Panel A of the figure shows the relationship between the incidence of testing and mean household income (in logs). It is obvious that there is one outlying zip code where residents were far more likely to be tested. The neighborhood with the largest (normalized) number of tests, with about 4,400 tests per 100,000 persons, has zip code 10018. This zip code roughly encompasses the west side of midtown Manhattan between 34th and 42nd streets, from 5th Avenue to the Hudson River.[9]

The graph suggests a positive relation between the incidence of testing and household income in the neighborhood. In other words, persons residing in wealthier neighborhoods were more likely to be tested than persons residing in poorer neighborhoods.

The middle panel of Figure 2 shows the relation between the probability that a test yields a positive result and mean household income. It is obvious that this probability is far higher in poorer neighborhoods. The figure suggests that the fraction of tests that reveal an existing infection is around 60 percent in neighborhoods at the 10th percentile of the income distribution (where the log of household income is 10.8), but the statistic falls to about 45 percent in neighborhoods that are in the 90th percentile (where the log of household income is 12.0).

[9] It is difficult to understand why the incidence of testing was so high in the west midtown area. There are no media reports suggesting the existence of a notable COVID-19 outbreak in this neighborhood. The second highest incidence of testing, with 2,537 tests per 100,000 persons, has zip code 11370, and it encompasses the Elmhurst area of Queens. That area, however, became notorious in mid-March because of the large number of virus-related admissions and fatalities at the Elmhurst Hospital Center (Rothfeld et al, 2020; Russell, 2020). It seems plausible that health officials might have responded to the crisis by increasing testing in that neighborhood in the last half of March, "artificially" inflating the testing incidence.

The bottom panel of the figure shows the net impact of these two conflicting correlations by linking the incidence of COVID-19 infection in the population and household income. Although the resulting correlation is negative, it is visually evident that it is relatively weak simply because persons residing in poorer neighborhoods were less likely to be tested and hence have a smaller chance of showing up in the infection counts.

Figure 3 provides an analogous summary of the correlation between the average household size in the neighborhood and the various outcomes. Panel A reveals that the incidence of testing was essentially independent of average family size. This allocation of testing resources might seem questionable given the very strong positive correlation between household size and the likelihood that a test yields a positive result illustrated in Panel B of the figure. In neighborhoods where small households are prevalent, fewer than 50 percent of the tests came back positive. In neighborhoods with larger households, over 60 percent of the tests came back positive.

The combination of these two effects attenuates the relationship illustrated in Panel C of the figure. There is a weaker positive correlation between the incidence of infection in the population and household size simply because testing was not particularly common in the neighborhoods where large families cluster.

Panel A of Figure 4 shows the relationship between the incidence of testing and the minority composition of the neighborhood, defined as the fraction of the neighborhood's population that is either Hispanic, non-Hispanic black, or non-Hispanic Asian. The graph suggests a negative relation between the incidence of testing and the percent of the neighborhood's population that is minority. In other words, persons residing in predominantly

minority neighborhoods were less likely to be tested than persons residing in mainly white neighborhoods.[10]

The middle panel of Figure 4 shows the relation between the probability that a test yields a positive result and the minority composition of the neighborhood. It is obvious that this probability is far higher in minority neighborhoods. The figure suggests that the fraction of tests that reveal an existing infection is around 45 percent in neighborhoods where 20 percent of the population is minority and increases to nearly 60 percent in neighborhoods that are over 60 percent minority.

The bottom panel of the figure shows the net impact of these two conflicting correlations by linking the incidence of COVID-19 infection in the population (per 100,000 persons) and the percent minority variable. Although the resulting correlation is positive, it is relatively weak simply because persons residing in minority neighborhoods were less likely to be tested and hence have a smaller chance of showing up in the infection counts.

Finally, Figure 5 illustrates the relationship between the outcome variables and the percent of the neighborhood's population that is foreign-born. The top panel of the figure reveals that persons residing in immigrant neighborhoods were less likely to be tested. The middle panel, however, shows a strong positive relation between the probability that a test is positive and percent immigrant. The combination of these two correlations summarized in Panel C leads to a weaker positive relation between the "immigrant-ness" of a neighborhood and the incidence of COVID-19.

[10] One important caveat: The regression analysis presented below conducts a more detailed examination of this correlation and shows that the negative correlation between the incidence of testing and the racial composition of the neighborhood is driven by Asian neighborhoods.

As this graphical description of the data suggest, it is crucial to examine the distinct channels through which any socioeconomic or demographic variable can affect the incidence of testing and the likelihood that any given test yields a positive result. At least in the early stage of the pandemic in New York City, testing resources were not randomly allocated across neighborhoods, contaminating the correlation between any socioeconomic characteristic and the (observed) population incidence of COVID-19 infection. The next section will estimate regression models that confirm many of the implications of these scatter diagrams even after controlling for the role played by other neighborhood characteristics.

4. Regression Results

Table 2 reports the main regressions that relate each of the three key outcome variables to the set of socioeconomic characteristics introduced earlier. The regressions are estimated using a grouped logit estimator. For expositional convenience, Table 2 reports the marginal effect of each of the regressors, where the marginal effect is defined as the percentage change in the number of "successes" in the dependent variable.[11]

The regressions confirm the insight emphasized in the previous section—that focusing solely on the incidence of COVID-19 in the population masks a lot of what is actually going on in New York City. Consider, for example, the impact of mean household income (introduced in logs in the regressions). The last column of the table shows that although the relative number of infections in the population is positively related to household income, after holding constant other neighborhood characteristics, the coefficient is not very significant. This finding would

[11] Specifically, the regressions are estimated using the *glogit* command in STATA, and the marginal effects are estimated using the *eydx* option.

lead to the inference that COVID-19 infections affected poor and wealthy NYC neighborhoods in roughly the same way.

However, the regression reported in the first column of the table documents that persons residing in wealthy neighborhoods were tested at much *higher* rates than persons residing in poor neighborhoods. The 90-10 gap in log household income is about 1.2 log points. This implies that the number of tests administered to persons residing in the wealthier neighborhoods was about 21.6 percent greater than the number of tests administered to persons residing in the poorer neighborhoods (for a given population size).

At the same time, however, the persons who were tested and who resided in the wealthy neighborhoods were *less likely* to test positive for the virus.[12] The number of positive test results declines by about 5.5 percent when comparing the 90th percentile neighborhood with the 10th percentile neighborhood (for a given number of tests administered). In short, the analysis of the distinct effects of household wealth on the incidence of testing and on the conditional probability of testing positive (holding constant other neighborhood characteristics) provides a different picture than the one implied by the weaker positive correlation between household income and the incidence of COVID-19 infections in the population.

It is important to note that the correlation between household income and the various outcome variables reported in Table 2 can be interpreted in different ways, and that the currently available data do not allow us to distinguish among alternative explanations. For instance, the regressions indicate that residents from wealthier NYC neighborhoods were tested relatively more often than their counterparts from poorer neighborhoods. But *why* did this difference arise?

[12] It is worth noting that the R-squared of the grouped logit regression on the conditional probability of a positive test result is quite high, exceeding 0.7. A small number of neighborhood characteristics do a remarkably good job of explaining the geographic dispersion in this probability.

It is possible that the scarce testing resources were disproportionately allocated to wealthier neighborhoods. But it is also possible that persons residing in the wealthier neighborhoods might have had better information networks or could more easily afford to obtain tests elsewhere in the city. It would be important to resolve this puzzle, but the publicly available data released by the NYC Department of Health reports only the zip code of residence for the persons tested and does not report the zip code where the test was administered.

The regression in Table 2 shows that household size is an important determinant of the various outcome variables, and in this case the effects on the two components of the infection rate work in the same direction, so that the correlation between the incidence of infection in the population and household size is very strong. In particular, persons residing in neighborhoods where large families were cluster were far more likely to be tested. Similarly, given that a test was administered, infections were much more likely to be detected in those neighborhoods. This is not surprising, of course, as the grouping together of a larger number of people raises the risk of exposure to COVID-19. The fact that the two effects work in the same direction leads to a very strong correlation between household size and the incidence of infections in the population: One additional person in the household increases the number of infections by about 46.4 percent.[13]

Table 2 also shows that predominantly male neighborhoods were more likely to be affected by COVID-19 infections. This finding, however, is solely by the fact that those neighborhoods were tested more intensively. The relative number of positive cases among persons who were tested is independent of the gender composition of the neighborhood's

[13] Although this may seem implausibly large, it is consistent with the raw data illustrated in Panel C of Figure 3. The figure suggests that an increase in household size from 1.5 to 2.5 increases the number of infections per 100,000 persons from about 500 to about 750, a 50 percent increase.

population. There is a similar pattern in the correlation between infections and the percent of the neighborhood's population that is over age 60. Persons residing in these neighborhoods were also far more likely to be tested, but the outcome of the test was independent of the age composition of the neighborhood.

The regressions examine the link between minority neighborhoods and COVID-19 infections in more detail than the graphical analysis presented in the previous section. In particular, the regression introduces a vector of variables measuring the percent of the neighborhood's population that belongs to each of the three main ethnic/racial groups. The analysis reveals striking differences across the groups.

Persons residing in neighborhoods with a predominantly black or, to a lesser extent, Hispanic population were more likely to be tested. In the case of black neighborhoods, these tests were also more likely to result in the detection of an infection. As a result, there is a strong positive correlation between the incidence of COVID-19 infections and the percent black (and, to a lesser extent, percent Hispanic) variables.

The regression also shows, however, that the sign of these correlations is reversed for Asian neighborhoods. Persons residing in those neighborhoods were less likely to be tested. Once tested, the tests were less likely to come back positive. As a result, the incidence of COVID-19 infections among persons residing in predominantly Asian neighborhoods was far lower than for persons residing in other types of neighborhoods.

Given the geographic origin of the COVID-19 pandemic in Wuhan, China, the results regarding the incidence of testing and frequency of positive test results in Asian neighborhoods are unexpected. Note, however, that 60 percent of Asian immigrants in the city were not born in China. Nevertheless, it would be interesting to uncover why the spread of the virus in

neighborhoods with a large Asian population behaved in different ways than the spread in black or Hispanic (or white) neighborhoods.

Finally, New York City has long been an urban area that attracts very large numbers of immigrants. The regression reported in the last column of Table 2 indicates that the "immigrant-ness" of the neighborhood's population is uncorrelated with the incidence of an infection in the population (holding the other neighborhood characteristics constant). In other words, the number of infections per 100,000 persons was essentially the same for persons living in immigrant or in "native" neighborhoods. However, this result arises because persons in immigrant neighborhoods were less likely to be tested, and far more likely to test positive once the test was administered. The evidence regarding the correlation between immigrant neighborhoods and CUVID-19 infections in New York City, therefore, again reveals how the nonrandom allocation of testing resources in the city affects the perception of which groups are disproportionately affected by the virus.

I should note that the regression analysis summarized in Table 2 only examines the role of a small number of variables that hope to capture key aspects of how the COVID-19 pandemic affected the various neighborhoods of New York City. By focusing on a small number of regressors, the analysis avoids the problem of multicollinearity that a kitchen-sink approach would introduce, particularly given the limited nature of the sample (a single cross-section of outcomes summarizing the cumulative impact of the pandemic on NYC as of April 5, 2020) and the relatively small number of observations.

Nevertheless, the regressions ignored the obvious fact that the various neighborhoods are located in one of the five distinct boroughs that make up the City of New York (Manhattan, the Bronx, Brooklyn, Queens, and Staten Island). Figure 1 suggests that there may be noticeable

differences across the five boroughs in the incidence of COVID-19 infections per 100,000 persons (which, of course, partly reflect the different kinds of neighborhoods that compose the various boroughs).

Table 3 re-estimated the regressions after adding a set of borough fixed effects. The results tend to resemble those reported in Table 2, although many of the coefficients are not statistically significant. For example, the correlation between household income and the various outcome variables is weaker once the regression controls for borough of residence. This is not surprising, as it is well known that Manhattan has a far higher level of household income than the other boroughs, so that the Manhattan fixed effect attenuates the impact of household income on the incidence of testing and on the likelihood of positive test results.[14] Nevertheless, the regressions in Table 3 suggest that—even *within* boroughs—the population in the neighborhoods that were predominantly black or Hispanic, or had larger households, or had a large immigrant population, are more likely to test positive once tested. Similarly, the population in the neighborhoods where large households cluster or are predominantly male—even within a borough—are still more likely to be tested.[15]

6. Summary

This paper documents the characteristics of New York City neighborhoods that were most affected by the COVID-19 pandemic. It uses data compiled by the New York City

[14] Manhattan's average household income is $125,900, as compared to between $48,700 and $90,300 for the other four boroughs.

[15] The coefficients of the borough fixed effects in Table 3 suggest the presence of borough-wide impacts that would be interesting to pursue in further research, as they may provide insight into the evolution of the pandemic. For example, persons residing in Brooklyn had a relatively lower incidence of testing than persons from other boroughs, while persons residing in the Bronx had a relatively higher incidence. In addition, the likelihood a test yielded a positive result was far higher in Brooklyn and Queens than in the other boroughs.

Department of Health and Mental Hygiene that reports the number of COVID-19 tests administered and the number of persons infected at the level of a zip code. I merged these administrative counts with data from the 2010 decennial census and the 2010-2014 American Community Surveys to describe the neighborhoods where COVID-19 testing was relatively more common and where the number of positive test results was relatively high.

The probability of a positive test result (conditional on testing) is larger in poorer neighborhoods, in neighborhoods where large numbers of people reside together, and in neighborhoods with a large black or immigrant population. At the same time, however, persons residing in poorer or immigrant neighborhoods were less likely to be tested.

The rate of infection in a given population (i.e., the number of infected persons divided by the size of the population) depends on two separate factors: the frequency of tests and the fraction of positive tests among those tested. The New York City experience suggests that the net correlation between socioeconomic characteristics and the rate of infection sometimes captures the net impact of perhaps two conflicting forces. As a result, an understanding of which types of neighborhoods are disproportionately affected by the pandemic requires an examination of how socioeconomic characteristics correlate with each of the two determinants of the rate of infection in the population.

References

Baker, Scott R. R.A. Farrokhnia, Steffen Meyer, Michaela Pagel, and Constantine Yannelis, "How Does Household Spending Respond to an Epidemic Consumption During the 2020 COVID-19 Pandemic," NBER Working Paper No. 26949, April 2020.

Berger. David W., Kyle F. Herkenhoff, and Simon Mongey, "An SEIR Infectious Disease Model with Testing and Conditional Quarantine," NBER Working Paper No. 26901, March 2020.

Buchanan, Larry, Jugal K. Patel, Brian M. Rosenthal, and Anjali Singhvi, "A Month of Coronavirus in New York City: See the Hardest-Hit Areas," *New York Times*, April 1, 2020.

Gallagher, Patrick. "Identification and Analysis of Orthodox Jewish Enclaves in Brooklyn, New York: A GIS Based Approach," *Middle States Geographer*, 2009, 42: 83-89
Russell, David. "Elmhurst deals with coronavirus crisis," Queens Chronicle, April 2, 2020. https://www.qchron.com/editions/central/elmhurst-deals-with-coronavirus-crisis/article_c4e30040-cc13-5762-bfa1-4521d1daafc7.html.

Harris, Jeffrey E. "The Coronavirus Epidemic Curve is Already Flattening in New York City," NBER Working Paper No. 26917, April 2020.

Honan, Katie. "The New York Neighborhoods with the Most Coronavirus Cases," *Wall Street Journal*, April 1, 2020.

Lang. Hanming, Long Wang, Yang Yang, "Human Mobility Restrictions and the Spread of the Novel Coronavirus (2019-nCoV) in China," NBER Working Paper No. 26906, April 2020.

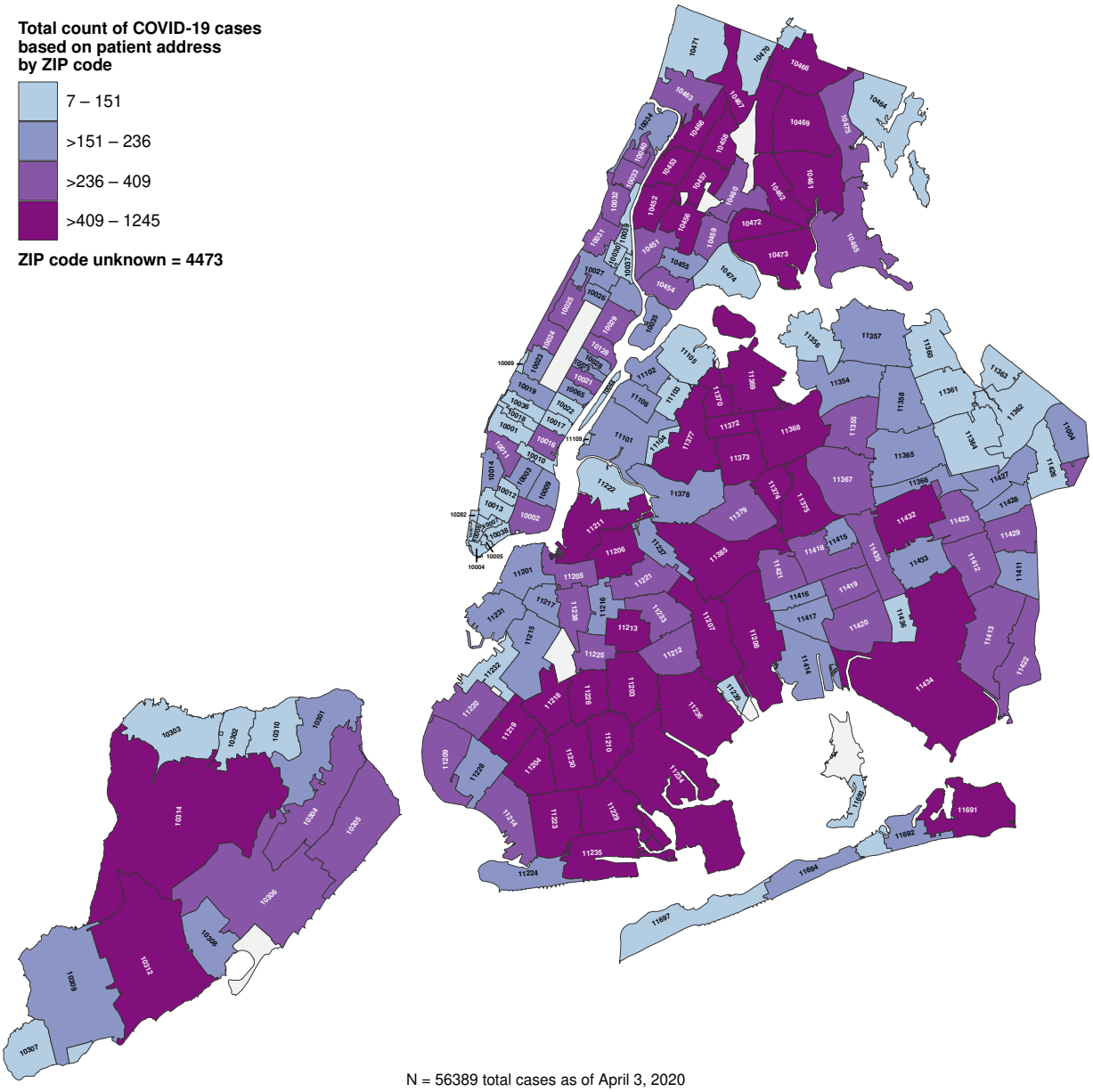
Marsh, Julia. "NYC Map Shows Total Cases Testing Positive for Coronavirus by ZIP Code," *New York Post*, April 1, 2020.

Rothfeld, Michael, Somini Sengupta, Joseph Goldstein, and Brian M. Rosenthal. **"**13 Deaths in a Day: An 'Apocalyptic' Coronavirus Surge at an N.Y.C. Hospital," *New York Times*, March 25, 2020.

Sales, Ben. "Brooklyn's Orthodox neighborhoods have especially high rates of the coronavirus," Jewish Telegraphic Agency, April 2, 2020. https://www.jta.org/2020/04/02/united-states/brooklyns-orthodox-neighborhoods-have-especially-high-rates-of-coronavirus.

Stack, Liam and Nate Schweber, "Coronavirus: 'Huge Spike' in Brooklyn Hasidic Community," New York Times, March 18, 2020. https://www.nytimes.com/2020/03/18/nyregion/Coronavirus-brooklyn-hasidic-jews.html.

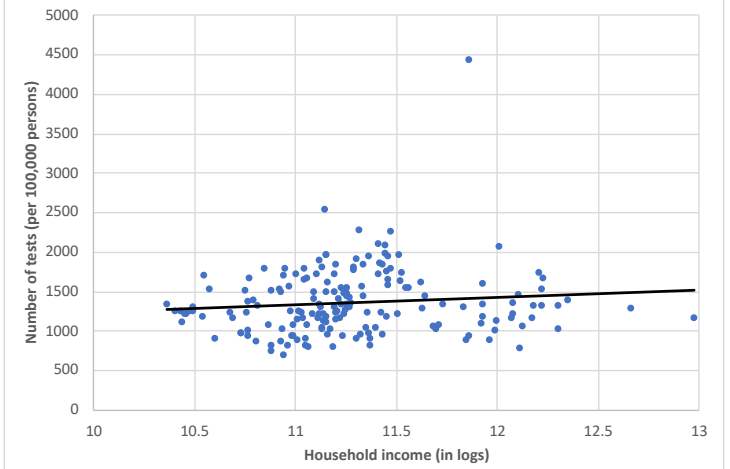
Figure 1. Covid-19 cases in New York City, by zip code (as of April 3, 2020)



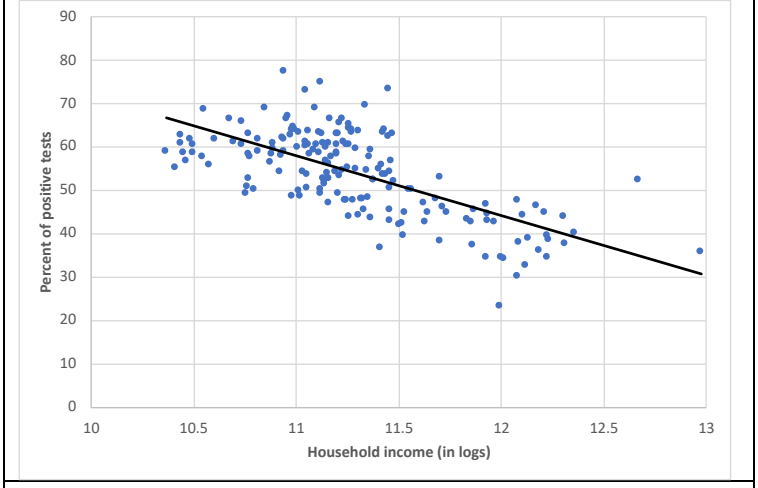
Source: Map prepared by the NYC Department of Health; https://www1.nyc.gov/assets/doh/downloads/pdf/imm/covid-19-cases-by-zip-04032020-1.pdf.

Figure 2. Household income and COVID-19

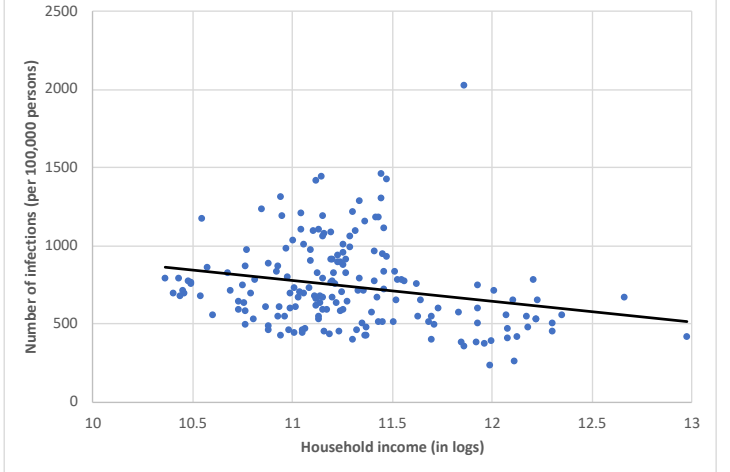
A. Incidence of testing (per 100,000 persons)



B. Percent of tests that are positive

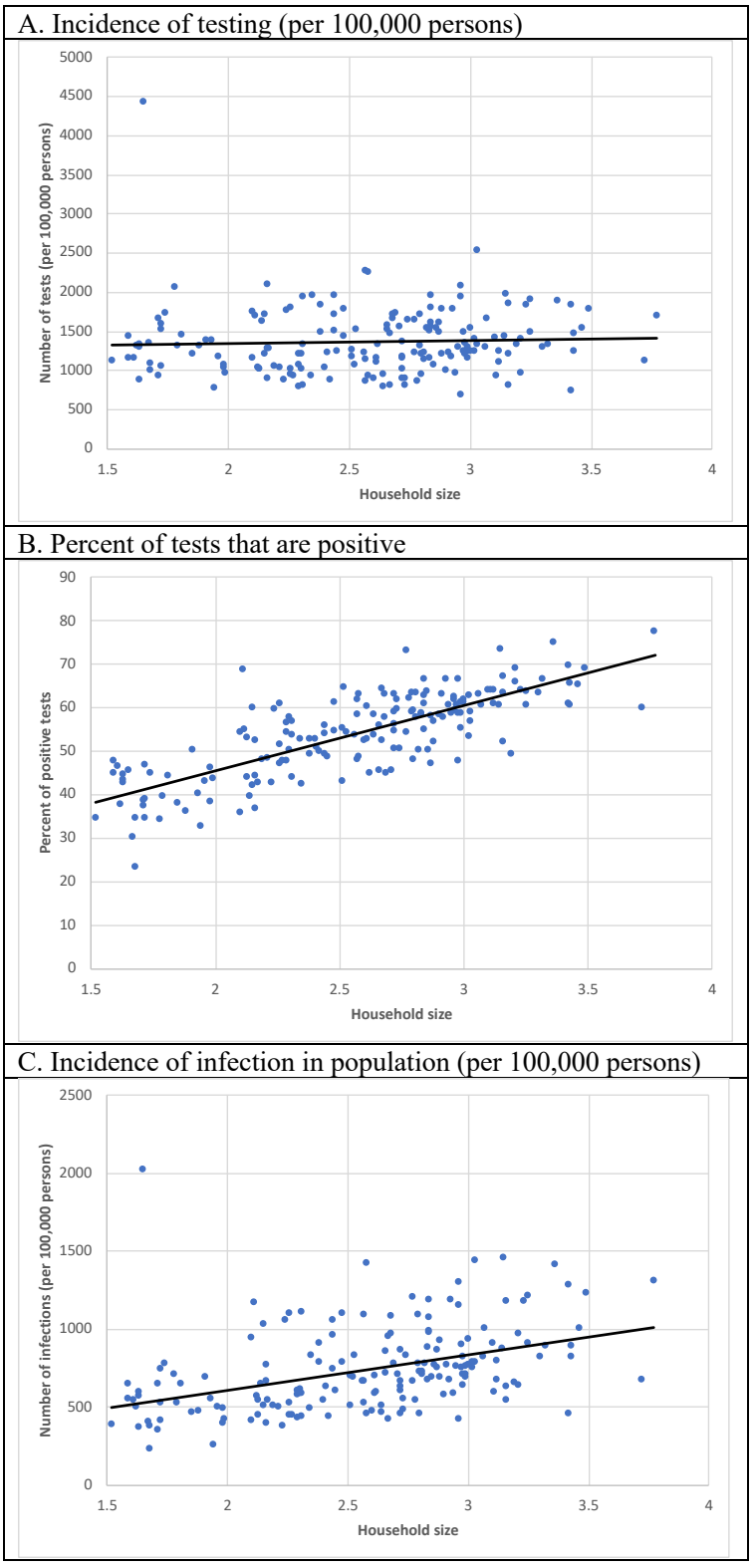


C. Incidence of infection in population (per 100,000 persons)



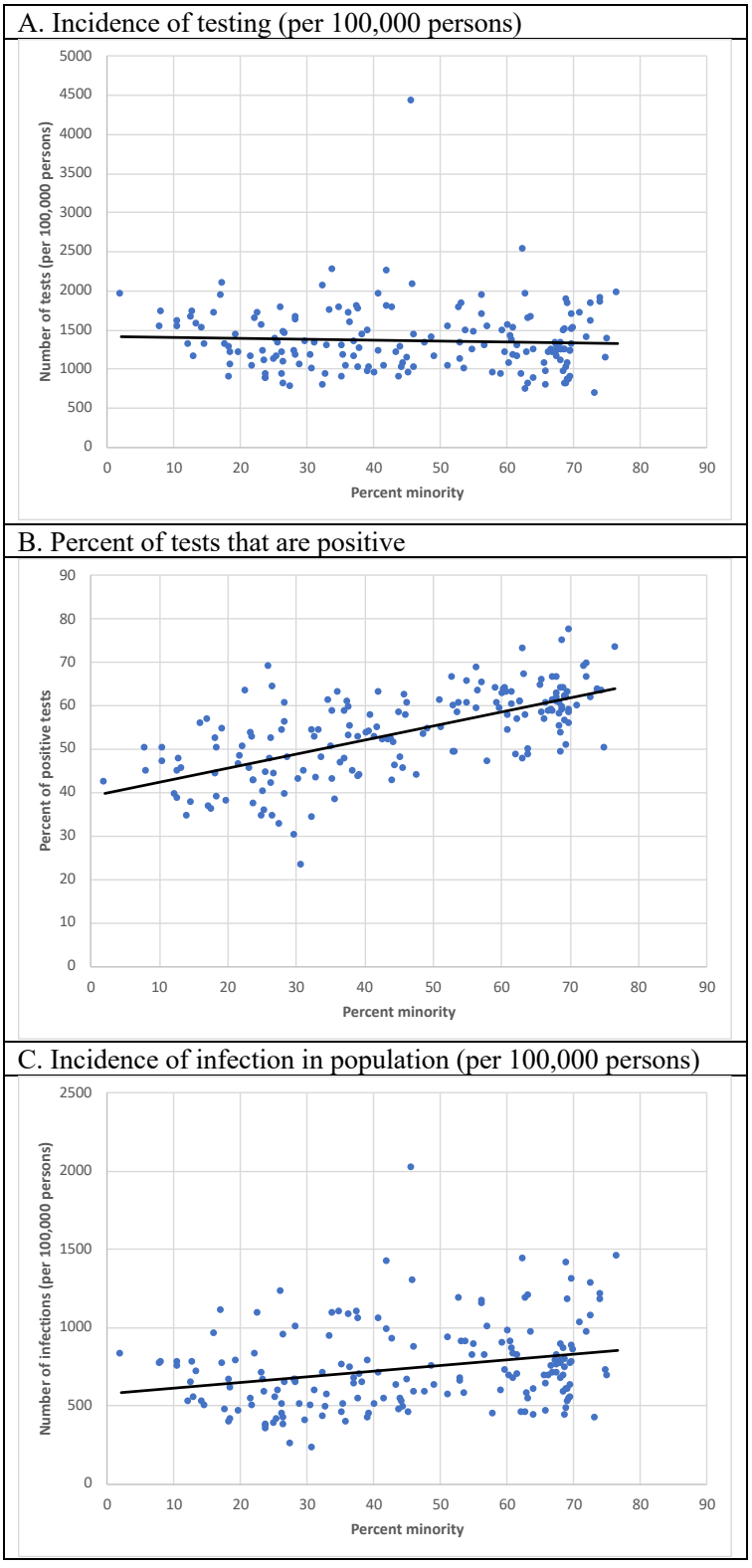
Notes: The income variable gives the household income for the average household in the zip code. See the text for details on the construction of the variables.

Figure 3. Household size and COVID-19



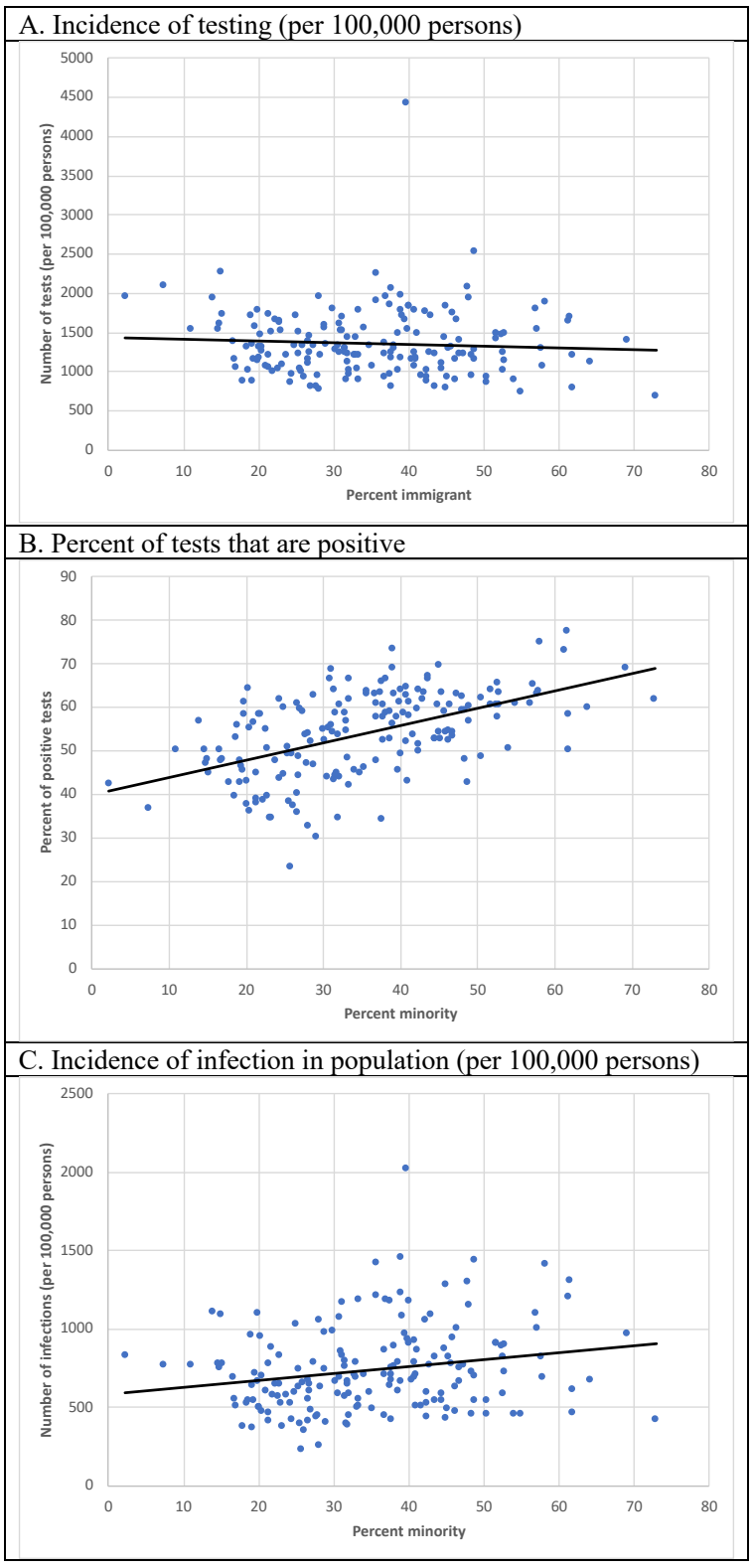
Notes: The household size variable gives the number of persons in the average household. See the text for details on the construction of the variables.

Figure 4. Minority neighborhoods and COVID-19



Notes: The percent minority variable gives the fraction of the zip code's population that is either Hispanic, non-Hispanic black, or non-Hispanic Asian. See the text for details on the construction of the variables.

Figure 5. Immigrant neighborhoods and COVID-19



Notes: The percent immigrant variable gives the fraction of the zip code's population that is foreign-born. See the text for details on the construction of the variables.

Table 1. Summary statistics

Variable:	Mean	10th percentile	Median	90th percentile
Number of tests	598.2	219.0	515.0	1013.0
Number of positive tests	337.9	101.0	274.0	624.0
Population of zip code (in 1000s)	46.1	14.0	41.0	85.9
Number of tests per 100,000 persons	1362.1	902.9	1296.2	1840.6
No. of positive tests per 100,000 persons	740.3	442.2	692.7	1107.4
Percent of tests that are positive	54.0	39.4	54.9	64.3
Average household income (in 1000s)	90.4	46.8	74.0	164.0
Size of household	2.6	1.7	2.7	3.2
Percent male	47.7	44.7	47.5	50.3
Percent age ≥ 60	17.5	11.4	16.8	25.3
Percent black	16.1	1.3	6.4	52.1
Percent Hispanic	18.8	5.6	13.1	44.3
Percent Asian	10.5	1.1	6.9	25.1
Percent immigrant	35.1	19.3	34.0	52.5

Notes: The sample has 177 observations. The data on the number of COVID-19 tests and the number of positive results were compiled by the New York City Department of Health and Mental Hygiene. The counts are cumulative to April 5, 2020. The data for the zip code's population, the average size of the household, the percent male, and the percent in the various ethnic/racial groups are drawn from the 2010 decennial census files. The variables measuring household income and the percent of the population that is immigrant are drawn from the pooled 2010-2014 American Community Surveys.

Table 2. Correlates of COVID-19 testing incidence and infections

	Number of tests relative to the population	Number of infections relative to number of tests conducted	Number of infections relative to the population
Log mean household income	0.182	-0.046	0.107
	(0.066)	(0.028)	(0.077)
Average household size	0.284	0.179	0.464
	(0.055)	(0.023)	(0.064)
Percent male	0.038	0.002	0.043
	(0.010)	(0.004)	(0.012)
Percent age ≥ 60	0.029	-0.000	0.031
	(0.005)	(0.002)	(0.006)
Percent black	0.002	0.002	0.004
	(0.001)	(0.001)	(0.002)
Percent Hispanic	0.002	0.000	0.003
	(0.002)	(0.001)	(0.002)
Percent Asian	-0.008	-0.001	-0.009
	(0.003)	(0.001)	(0.003)
Percent immigrant	-0.003	0.004	-0.000
	(0.002)	(0.001)	(0.002)
R-squared	0.331	0.724	0.448

Notes: Standard errors in parentheses. All regressions have 177 observations and are estimated using the command *glogit* in STATA. This procedure estimates grouped logit regression models using weighted least squares. The coefficients report the marginal effect of the independent variable in percentage terms, specifically giving the percent change in the number of "successful" cases (i.e., the numerator of the dependent variable) resulting from a one-unit change in the independent variable.

Table 3. Correlates of COVID-19 testing incidence and infections, including borough fixed effects

	Number of tests relative to the population	Number of infections relative to number of tests conducted	Number of infections relative to the population
Log mean household income	0.130	-0.040	0.061
	(0.069)	(0.030)	(0.085)
Average household size	0.254	0.154	0.390
	(0.058)	(0.025)	(0.070)
Percent male	0.043	-0.002	0.045
	(0.009)	(0.004)	(0.011)
Percent age ≥ 60	0.021	-0.000	0.023
	(0.005)	(0.002)	(0.006)
Percent black	0.001	0.002	0.003
	(0.001)	(0.001)	(0.002)
Percent Hispanic	-0.005	0.002	-0.003
	(0.002)	(0.001)	(0.003)
Percent Asian	-0.012	-0.001	-0.013
	(0.002)	(0.001)	(0.003)
Percent immigrant	-0.000	0.002	0.001
	(0.002)	(0.001)	(0.002)
Manhattan			---
Bronx	0.228	0.005	0.243
	(0.064)	(0.0302	(0.079)
Queens	0.060	0.121	0.188
	(0.062)	(0.027)	(0.076)
Brooklyn	-0.178	0.097	-0.079
	(0.066)	(0.030)	(0.082)
Staten Island	-0.035	-0.014	-0.024
	(0.088)	(0.041)	(0.111)
R-squared	0.517	0.778	0.567

Notes: Standard errors in parentheses. All regressions have 177 observations and are estimated using the command *glogit* in STATA. This procedure estimates grouped logit regression models using weighted least squares. The coefficients report the marginal effect of the independent variable in percentage terms, specifically giving the percent change in the number of "successful" cases (i.e., the numerator of the dependent variable) resulting from a one-unit change in the independent variable.