Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Crowding Effects Dominate Demographic Attributes in COVID-19 Cases

Awi Federgruen\textsuperscript{a,}*, Sherin Naha\textsuperscript{b}

\textsuperscript{a} Graduate School of Business at Columbia University, 419 Uris Hall, 3022 Broadway, New York 10027, NY, United States
\textsuperscript{b} Management Science & Engineering Masters Program at Columbia University, United States

\textbf{A R T I C L E   I N F O}

Article history:
Received 8 July 2020
Received in revised form 19 October 2020
Accepted 22 October 2020

Keywords:
COVID-19
Socio-economic factors
Demographic factors
Average household size
Population density
Racial/ethnic factors

\textbf{A B S T R A C T}

\textbf{Objective:} With an eye toward possible public policy implications, our objective is to identify the socio-economic and demographic factors that drive the large variation in COVID-19 incidence rates observed within relatively compact geographic regions, and to quantify the relative impact of each of these factors. We use international comparisons as a starting point.

\textbf{Methods:} New York City, consisting of some 175 zip codes, is an ideal arena to pursue the above study given the large variation in case incidence rates across zip codes. We conducted systematic regression studies employing data with zip code granularity. Our model specifications are based on a well-established epidemiologic model that explains the effects of household sizes on $R_0$.

\textbf{Results:} Average household size emerges as the single most important driver behind the large variation in COVID-19 incidence rates. It independently explains 62% of the variation. The percentage of the population above the age of 65 and the percentage below the poverty line are also strongly positively associated with zip code incidence rates. As to ethnic/racial characteristics, the percentages of African Americans, Hispanics and Asians within the population are significantly associated, but the magnitude of the impact is smaller. (The proportion of Asians within a zip code has a negative association.) Contrary to common belief, population density, by itself, does not have a significantly positive impact (other than when a high population is driven by large household sizes).

\textbf{Conclusion:} Our findings support implemented and proposed policies to quarantine patients and separate infected individuals from families or dormitories; they also support newly revised nursing home admission policies.

\textcopyright 2020 The Author(s). Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

\textbf{Introduction}

The world continues to search for a fundamental understanding of the dynamics of the current pandemic. For example, we try to understand why, in the United States (US), as of May 17, 2020, 192,000 COVID-19 cases have been reported in New York City (NYC) alone, while the country-wide total has been mercifully restricted to 1.5 million. Thus, a city representing 2.5% of the US population accounted for 12.8% of the reported cases. Identifying the main drivers of the disease spread has important implications for public policies to contain the current epidemic and mitigate the widely expected “second wave”.

\textit{Population density} is widely believed to be a main driving force. This theory has intuitive appeal. After all, the number of infections in a given region depends on the basic reproductive number $R_0$, the average number of cases directly generated by one individual, in a population where all individuals are susceptible. $R_0$, in turn, depends, \textit{in part}, on the number of individuals with whom a single case has physical contact during the time interval in which she is contagious, presumably positively correlated with the population density.

However, the theory is challenged, first of all, on an international basis. Many cities with population density as great as or greater than NYC’s 10,198 residents per square kilometre (sq. km) have reported much lower case rates. For example, Manila, Baghdad, Mumbai, Seoul, Mexico City and Singapore have, respectively, 46128, 32874, 32103, 16000, 9800 and 8358 residents/sq. km. Their case incidence rates per 100,000 residents vary, however, between 9.4 (Seoul) and 635.4 (Singapore), as compared to NYC’s rate of 2286 cases per 100,000 residents.

These cities’ lower rates may be explained by ex-ante differences in international traffic patterns in and out of the country,

\textsuperscript{a} Corresponding Author
E-mail address: af7@columbia.edu (A. Federgruen).
affecting the cluster of “imported” cases, or specific containment and testing policies adopted by the respective governments. However, among states within the US, California had one of the highest population densities (2513 residents per square mile) but one of the lowest COVID-19 mortality rates (8 per 100,000 residents), while Louisiana, with a population density 2.5 times lower than that of California, has reported 49 COVID-19 deaths per 100,000 residents.

And stark differences are apparent within NYC itself. (Wadhera et al., 2020) reported recently that among the city’s five boroughs, Manhattan had by far the fewest hospitalizations per 100,000 residents, but the greatest population density, 2.5 times the citywide average (25,106 vs 10,198 residents/sq.km). (Because the percentage of confirmed cases that require hospitalization is remarkably robust throughout the country, hospitalization rates can be viewed as proxies for incidence rates.) In fact, at zip code granularity, the rates of reported cases and the population densities are negatively correlated, with a correlation coefficient of $r = -0.33$ (Table 2).

Several authors ([Barr & Tassier, 2020] and [Bassett, 2020]) have cautioned against relying on population density as a prime explanatory variable.

First and foremost, populations consist of “households” and the inter-household and intra-household dynamics of an epidemic differ fundamentally. In particular, the contact rate between a pair of individuals living in the same household tends to be much greater than that between two individuals from different households. Starting with the seminal papers (Bartoszynski, 1972) and (Becker, 1977) this observation has been at the core of many epidemiological models. See Section II for a seminal model by (Becker & Dietz, 1995). (Rader et al., 2020) report that infection patterns across China can only be explained with “hierarchically-structured populations where individuals had high rates of contact within their households”.

The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team (2020) reported that 80% of transmission clusters of the coronavirus occurred within households. Nearly three weeks into the initial outbreak in Wuhan, the government instituted social distancing and travel bans, reducing $R_0$ from 3.88 to 1.26. However, it was not until Wuhan instituted a centralized quarantine where anyone with COVID-19-like symptoms would be centrally quarantined at hotels or dormitories, that $R_0$ was reduced to 0.32, a 75% improvement over social distancing alone! Indeed, this intervention was the final step in achieving full containment in China.

In NYC, average household size varies by zip code, between 1.57 and 3.97 (Table 1). It is also highly positively correlated with the number of confirmed cases per 100,000 residents (Table 2). And although Manhattan is the most densely populated of all boroughs, it is also the one with the smallest average household size. Thus, any statistical model intended to explain the variation in case intensities across the City should include the average household size as an explanatory variable.

More broadly, we need to identify demographic and socio-economic indicators that have significant predictive power. NYC is an ideal arena in which to pursue this investigation; as mentioned, the city has, by far, the highest confirmed incidence rate in the country, but it also exhibits remarkable diversity with respect to many potentially relevant demographic and socio-economic factors. Combining various data sources, described below, we have been able to assess these factors for 174 of the 177 NYC zip codes; see Table 1.

A basic household-based epidemiology model

(Becker and Dietz, 1995) consider a population with households of varying sizes. Let $\mu_H$, $\sigma_H$ and $c_VH$ denote the mean, standard deviation and coefficient of variation of the distribution of household sizes, respectively. Let $b$ denote the mean number of infectious contacts that an individual makes with individuals outside her own household during her entire infectious period. An infectious contact is one that is close and intensive enough to result in disease transmission.

The following closed form expression for $R_0$, the basic reproductive number of the epidemic, was derived from [6, p.211]:

$$R_0 = \frac{b}{2} \left[ 1 + \left(1 + 4\left(\frac{\mu_H(1 + CV_H^2) - 1}{b}\right)\right)^{\frac{1}{2}} \right] \quad (1)$$

In a population where all individuals live alone, $\mu_H = 1$ and $c_VH = 0$, so that $R_0 = b$, as in elementary models that do not account for household clusters. In contrast, where people live in households of size 4 (e.g., married couples with 2 children), $\mu_H = 4$ and, if $b >= 1$ (i.e., an average infected individual infects at least one individual outside her own household):

$$R_0 = 0.5 \left( b + \sqrt{b^2 + 12b} \right) = b + 0.5 \left( \sqrt{b^2 + 12b} - b \right)$$

adding at least a full 1.3 points to the $R_0$ value! And when household sizes vary, as they typically do, the household effect is even greater, see (1). For example, if the household distribution approaches a geometric distribution,

$$CV_H = \sqrt{\frac{1 - \mu_H^{-1}}{\mu_H}} = \sqrt{0.75} = 0.87, R_0 = 0.5 \left( b + \sqrt{b^2 + 24b} \right)$$

$$= b + 0.5 \left( \frac{12}{\sqrt{1 + 24b^{-1}}} \right) \geq b + 2$$

| Table 1 | Descriptive statistics by NYC zip code. |
|---------|---------------------------------------|
| # confirmed cases per 100,000 residents | Minimum | Maximum | Average |
| Population density= # residents per sq. mile | 1389.08 | 126,067.69 | 38,480.75 |
| Average household size | 1.57 | 3.97 | 2.65 |
| Percentage below poverty line | 2.20 | 45.90 | 16.83 |
| Percentage above the age of 65 | 0.46 | 28.98 | 14.26 |
| Percentage of total Hispanic population | 1.12 | 75.77 | 26.35 |
| Percentage total Black non-Hispanic population | 0.37 | 90.51 | 20.06 |
| Percentage total Black population | 0.37 | 91.81 | 23.69 |
| Percentage total Asian population | 0.07 | 72.62 | 14.78 |
Population density, by itself, may affect R0 via the parameter b, depending on the interactions among individuals in different households. But if it does, (1) shows that it interacts strongly with average household size (and its coefficient of variation) as well.

### Methods

The numbers of reported COVID-19 cases per zip code were obtained from the NYC Department of Health’s Github repository, and population sizes and population densities from http://zipatlas.com/us/ny/zip-code-comparison/population-density.htm.

The demographic and socio-economic characteristics below were extracted from https://github.com/BuzzFeedNews/2020-05-covid-city-zipcodes/blob/master/data/processed/census/zip_census_data.csv, a database created from 2018 5-year American Community Service (ACS) estimates, obtained from the US Census Bureau.

### Average household size

- **Age**
  - age_60_and_over
  - age_65_and_over
  - age_75_and_over
  - median_age

- **Economic**
  - pct_more_than_one_occupant_per_room
  - pct_below_poverty_level
  - household_median_income
  - household_mean_income

- **Racial/Ethnic**
  - total_non_hispanic
  - total_white_nonhispanic
  - total_black_nonhispanic
  - total_asian_alone
  - total_black
  - total_hispanic

We applied standard linear regression to estimate best fit regression equations, assuming independent noise terms, all with an identical normal distribution. We have conducted our study in two stages: in the first stage, we focused on crowding, age and income related factors; in the second stage, we investigated to what extent any of the racial/ethnic characteristics have additional explanatory power.

Based on the empirical and theoretical observations above, we specified a model with (a) population density and (b) average household size as potential explanatory variables, along with (c) an interaction term between them, (d) the percentage of the population above the age of 65, and (e) the percentage below the poverty line. The relevance of the latter two demographics is both intuitive and is explained in our Discussion section. In the remainder of this paper, we employ the following variable names:

- **population_density**: Population density=# residents per sq. mile
- **avg_household_size**: Average household size
- **pct_below_poverty_line**: Percent below poverty line
- **pct_age>65**: Percent above the age of 65
- **cases_per_100k**: # confirmed cases/100,000 residents

As mentioned, the correlation between population density and cases_per_100k is -33%. The correlation between the standard interaction term population density*avg_household_size and cases_per_100k is also negative – albeit weaker (-12%). We therefore specified the interaction term as (population density*avg_household_size)^2 which is positively correlated with cases_per_100k.

We have based our regression studies on the reported cases on May 12, 2020, and report the results for that day. However, to assess the robustness of our results, we have repeated the studies for a full 15 day window around this target date, starting with May 5, 2020 and ending with May 19, 2020.

### Results

As the correlation figures in Table 2 suggest, average household size is, by far, the single most important explanatory variable in the model. Indeed, it independently explains 62% of the variation in incidence rates. In a model that includes only household size as an explanatory variable, the following equation provides the best fit.

\[
\text{# confirmed cases/100,000} = 450.5 + 986.9 \times (\text{avg household size} - 1)
\]

The coefficient in front of the average household size variable is highly statistically significant. The equation indicates that in zip code A, where the average household has one more member than that in otherwise comparable zip code B, the case rate would be 986.9 cases/100,000 higher than in zip code B (Table 3).

**Table 2**

| Population density= # residents per sq mile | Average household size | Percent below poverty line | Percent above age 65 | # confirmed cases/100,000 residents |
|---------------------------------------------|------------------------|----------------------------|----------------------|------------------------------------|
| Population density= # residents per sq mile | 1                      | -0.324                     | 0.285                | -0.064                             |
| Average household size                      | -0.324                 | 1                          | 0.189                | -0.207                             |
| Percent below poverty line                  | 0.285                  | 0.189                      | 1                    | 0.622                              |
| Percent above age 65                        | -0.064                 | -0.207                     | -0.367               | 0.952                              |
| # confirmed cases/100,000 residents         | -0.325                 | 0.622                      | 0.229                | 0.092                              |

**Table 3**

Regression results with average household size as the single explanatory variable; avg_household_size_adj = average household size - 1.

| Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|--------------|----------------|--------|---------|-----------|-----------|
| Intercept    | 450.51         | 163.46 | 2.76    | 0.006     | 127.86    | 773.54542 |
| avg_household_size_adj | 986.90 | 94.79  | 10.41   | 5.41E-20 | 799.80    | 1174.01   |
However, a significantly superior fit can be obtained by adding the remaining explanatory variables (Table 4). This model accounts for 72% of the variation in case rates, while R² and the adjusted R² are 51% and 50%, respectively, resulting in the following equation:

\[ \text{cases}/100,000 = -284.4 + 896.8 \times (\text{average household size})^{-1} + 49.9 \times \text{age 65} + 24.4 \times \text{percentage below poverty level} - 0.006 \times \text{population density} - 3.52E-05 \times (\text{population density} \times \text{average household size})^{-1} \]  

(2)

The average household size continues to be the main determinant of the number of reported COVID-19 infections. In contrast, neither population density, nor its interaction term with the average household size, has a significant impact on the case intensity. (The intercept value is non-significant as well.)

Average household size, the percentage of the population above the age of 65 and the percentage below the poverty line, all have positive coefficients that are statistically significant at a very high confidence level. Including the other demographic variables in the model reduces the coefficient of the average household size variable to 896.8 but allows us to conclude at a 95% confidence level that an increase of but one individual to the average household size in the zip code leads, ceteris paribus, to an increase of at least 540.4 cases per 100,000 residents.

An increase of the percentage of senior residents by one percentage point augments, ceteris paribus, the case rate by approximately 50 cases, more than double the effect of an increase of the poverty rate by a single percentage point. The association of both demographic factors with zip code incidence rates is plausible as well; senior residents may be equally susceptible to getting infected as other age brackets. However, they are at increased risk to develop significant symptoms and complications, thus disproportionally contributing to the counts of confirmed cases, the only counts we have. (Bendavid et al., 2020) estimated that the number of confirmed cases in Santa Clara county may have been a mere 2% of the total number of infections.

Likewise, the a priori health status of individuals in the lowest income bracket is typically inferior to that in higher income brackets. Infected individuals with incomes below the poverty line therefore contribute disproportionally to the counts of confirmed cases as well. Additionally, a much larger percentage of this population segment is employed on location, as opposed to being able to work from home, and is thus more susceptible to infections.

We have checked for the presence of other interaction effects, omitting the population density-related variables that were found to have an insignificant impact. The only significant interaction effect is that between the percentage of the population below the poverty line and the percentage of senior residents. See Table 5 for the revised regression results. (The addition of the interaction term has a negligible impact on the R² or adjusted R² value).

Thus, after omitting the population density related variables whose impact was found to be statistically insignificant, our estimate for the coefficient of avg_household_size_adj has a much smaller standard error, resulting in a much narrower 95% confidence interval. The point estimate of this coefficient is now 1051.7, and above 876, at a 95% confidence level. The revised model specification has minimal impact on the estimated coefficients of the remaining explanatory variables.

As mentioned, we have repeated the regression studies for each of 15 days around the target date. The estimated coefficients are all remarkably robust, all well within the 95% confidence interval for the May 12 estimates; see Table 6.

We have created an interactive heat map of NYC. The color of each area is determined by the zip code’s residual in the above regression equation, which denotes the number of cases per 100,000 residents which the zip code experienced in excess of or below what could be expected based on its demographic characteristics. When clicking on a specific zip code, a table with all of the above measures emerges. See Fig. 1 for a snapshot.

### The impact of racial and ethnic factors

It has been widely conjectured that certain ethnic or racial groups, in particular the black and Hispanic communities, are at increased risk for COVID-19 morbidity. For example, (Departments of Public Health of Illinois; Mississippi and Virginia, 2020) reported that black Americans experienced a disproportionately greater number of reported COVID-19 cases compared to other Americans. (Yanci, 2020) concludes: “the pattern is irrefutable; underrepresented minorities are developing COVID-19 infection more frequently”

However, it remains difficult to assess whether the disproportionately greater incidence rate among the black population, as reported in the above three states, is pervasive throughout the country. The CDC publishes cumulative COVID-19 data reported by State Health Departments; however, 78% of those data were missing race/ethnicity disaggregation as of April 15, 2020, and 50%

---

1. [https://colab.research.google.com/drive/1CEY6UadCa3Nzwl-WoQHPVx_e7dVSIIKAlU?usp=sharing](https://colab.research.google.com/drive/1CEY6UadCa3Nzwl-WoQHPVx_e7dVSIIKAlU?usp=sharing)
as of May 30, see (Millett et al., 2020) and (Godoy & Wood, 2020), respectively.

It is easier to relate the case incidence rates in the country’s 3142 counties to the proportions of black residents in these counties, since complete data are available here. Figure 2 in Millett et al. (2020) shows that the case incidence, mostly, although not universally, increases with the county’s proportion of black residents: on the estimated curve, the case incidence rate doubles when comparing counties with a 1% black population with those where 75% of the population is black.

Moreover, even if the higher incidence rate is pervasive, we need to understand whether the difference with the general population is entirely due to other socio-economic or demographic factors (such as average household size or the poverty rate) or not.

(Millett et al., 2020) provided initial insights. They partitioned the country’s 3142 counties into two sets: the first set of 677

| Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|--------------|----------------|-------|---------|-----------|-----------|
| Intercept    | −549.53        | 246.39| −2.23   | 0.027     | −1035.93  | −63.14    |
| avg_household_size_adj | 1051.69 | 89.00 | 11.82   | 7.17E-24  | 875.99    | 1227.38   |
| pct_below_poverty_level | 25.48  | 5.43  | 4.692   | 5.57E-06  | 14.76     | 36.20     |
| pct_age65    | 48.62          | 9.73  | 4.957   | 1.44E-06  | 29.41     | 67.82     |
| pct < 2 * pct_age65 | −0.004 | 0.001 | −3.03   | 0.003     | −0.006    | −0.001    |

Table 6
Full Regression with interaction effect, but population density omitted.

Fig. 1. Snapshot of the heat map of NYC showing the zip code’s residuals calculated from regression Eq. (2).
Table 7
Regression results when including total_hispanic population as an additional independent variable.

| Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|--------------|----------------|-------|---------|-----------|-----------|
| Intercept    | 607.02         | 239.93| -2.53   | 0.0123    | -1080.66  | -133.37   |
| avg_household_size_adj | 918.81         | 90.26 | 10.18   | 0.0000    | 740.64    | 1096.99   |
| pct_below_poverty_level | 3.82           | 5.95  | 0.64    | 0.5220    | -7.93     | 15.56     |
| pct_age=65   | 54.71          | 9.57  | 5.72    | 0.0000    | 35.83     | 73.60     |
| total_hispanic | 12.35          | 3.03  | 4.08    | 0.0001    | 6.37      | 18.32     |

| Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|--------------|----------------|-------|---------|-----------|-----------|
| Intercept    | 607.02         | 239.93| -2.53   | 0.0123    | -1080.66  | -133.37   |
| avg_household_size_adj | 918.81         | 90.26 | 10.18   | 0.0000    | 740.64    | 1096.99   |
| pct_below_poverty_level | 3.82          | 5.95  | 0.64    | 0.5220    | -7.93     | 15.56     |
| pct_age=65   | 54.71          | 9.57  | 5.72    | 0.0000    | 35.83     | 73.60     |
| total_hispanic | 12.35          | 3.03  | 4.08    | 0.0001    | 6.37      | 18.32     |

counties with a disproportionately large (> 13%) black population and the second set with the remaining 2465 counties. The estimated case incidence rate in the first set is more than 4 times that in the second, but the associated 95% confidence intervals are overlapping.

Similar concerns apply to the Hispanic/Latino population. With the above proviso that in 48% of confirmed cases the race or ethnicity of the patient is unknown, the COVID-19 Racial Tracker reports that in 42 states plus Washington D.C., Hispanics/Latinos make up a greater share of confirmed cases than their share of the population. In eight states, it is more than four times greater.

To assess the impact of various ethnicity/racial factors, we have evaluated four additional regression models, each time adding to the model in Eq.(2), one of the following four ethnic /racial population percentages as an explanatory variable: (a) total_hispanic, (b) total_black_nonhispanic, (c) total_black and (d) total_asian_alone. (Once again, we omitted the population density related variables that were shown to have an insignificant impact, see above.). See Tables 7–10.

The addition of the ethnic/racial descriptors has little impact on the estimates of the coefficients and standard errors of our two main explanatory variables: average household size and the percentage of residents age 65 or older.

The coefficient of the percentage below the poverty line is halved, from approximately 25 to a number in the range [12.27,14.79] in Tables 8–10. In the remaining model (Table 7) where the impact of the proportion of Hispanics is assessed, this coefficient loses its significance. In this case, it is difficult to separate its impacts of the ethnic and the economic variable, as the two are highly positively correlated (65%, see Table 15).

The coefficients of each of the ethnic/racial descriptors are significant at a very high confidence level, with p-values below 0.0028, confirming the above described conjectures. However, the magnitude of these coefficients is modest, in particular when compared with the percentage of seniors and the percentage below the poverty line. An increase of the percentage of Hispanic residents by one percentage point augments, ceteris paribus, the case rate by approximately 12 cases. Thus, if in zip code A, the proportion of Hispanic residents is 20 percentage points higher than that in otherwise comparable zip code B, the case rate would be 247 higher than in zip code B.

The impact of the percentage of black residents, and that of black non-Hispanic residents, while strongly significant, is even smaller. An increase of this proportion by one percentage point augments, ceteris paribus, the case rate by approximately 6 cases. Thus, if in zip code A, the proportion of black residents is 20 percentage points higher than that in otherwise comparable zip code B, the case rate would be 115 higher than in zip code B.

In contrast, the proportion of Asian residents in a zip code has a (highly significant) negative association with the case rate, each additional percentage point decreasing, ceteris paribus, the case rate by some 14 cases.²

As with the basic model, we have repeated the regression analyses for each of the 15 days from May 5–May 19; once again the estimates are remarkably robust, see Tables 11–14.

Discussion
Average household size emerges as the single most important demographic variable associated with the large variation in infection rates. We estimate that an additional single household member increases the number of cases by 892 or more (Table 15).

This result is quite intuitive. It confirms both the above empirical findings and its theoretical underpinnings; see Section II. Of course, the same contagion potential exists in other settings where a significant number of individuals reside together in the same dwelling, for example dormitories and nursing homes. (Givvan, 2020) reported that in the US, 40% of COVID-19-induced deaths occurred in nursing homes and assisted living facilities. Similarly, across the globe, half of COVID-19 fatalities occurred in nursing homes and related facilities (Comas-Herrera & Zalakain, 2020). Their residents face the hazard of living in close proximity to many others, while belonging to an age group with significantly increased potential for serious symptoms and hence confirmation of their infections.

The proportion of the population above the age of 65 emerges as a second significant explanatory variable, with each additional percentage point contributing 50 cases per 100,000 residents. The association of the proportion under the poverty line with the outcome, such that each percentage point contributes an estimated 24 cases per 100,000, is also plausible: the health status of this segment is generally poorer, enhancing infections with significant symptoms. Moreover, far more individuals in this income bracket are employed as essential workers and are dependent on public transportation, with limited opportunities for social distancing; for both reasons, this population segment contributes more than others to the case count. The second factor pertains almost exclusively to individuals below retirement age, possibly explaining the significantly negative coefficient of the interaction term in Table 6.

Our findings about the great importance of household size support such policy initiatives as isolation policies for individuals, either immediately upon being identified as infected or post-hospitalization (see, for example, (Chan et al., 2020)). Based on their model (Chan et al., 2020), the authors show: "that the time to reopening [of the economy] can be shortened by 11%. In addition, assuming symptomatic people are infectious, if 50% of them are quarantined before getting sick enough to go to the hospital, we can reduce the risk of developing severe COVID-19 illness and the time till reopening can be shortened by 86%." Our findings also support the May 11 decision by New York State to cancel its mandate requiring nursing homes to accept COVID-19 patients.

Racial/ethnic factors have a significant, albeit more modest, association with case incidence rates above and beyond differences in average household size and the percentage below the poverty line with which they are correlated. (Richardson et al., 2020).

---

² The results thus partially confirm and partially challenges conjectures offered by various scholars, e.g., “Does this mean Latinos and Asians are being infected with, and dying from, COVID-19 at higher rates than other New Yorkers? We don’t know for certain, but it sure seems that way.” (Kendi, 2020)
### Table 8
Regression results when including %total_black_nonhispanic population as an additional independent variable.

| Coefficients          | Standard Error | t Stat | P-value | Lower 95%  | Upper 95%  |
|-----------------------|----------------|--------|---------|------------|------------|
| Intercept             | −593.85        | 245.16 | −2.42   | 0.0165     | −1077.82   | −109.87    |
| avg_household_size_adj| 961.76         | 90.63  | 10.61   | 0.0000     | 782.84     | 1140.67    |
| pct_below_poverty_level| 14.79        | 5.01   | 2.95    | 0.0036     | 4.89       | 24.69      |
| pct_age≥65            | 50.42          | 9.71   | 5.19    | 0.0000     | 31.26      | 69.59      |
| %total_black_nonhispanic| 5.87         | 1.94   | 3.03    | 0.0028     | 2.05       | 9.70       |

### Table 9
Regression results when including %total_black population as an additional independent variable.

| Coefficients          | Standard Error | t Stat | P-value | Lower 95%  | Upper 95%  |
|-----------------------|----------------|--------|---------|------------|------------|
| Intercept             | −600.22        | 244.74 | −2.45   | 0.0152     | −1083.37   | −117.08    |
| avg_household_size_adj| 961.57         | 90.44  | 10.63   | 0.0000     | 783.04     | 1140.11    |
| pct_below_poverty_level| 13.79        | 5.08   | 2.71    | 0.0073     | 3.76       | 23.83      |
| pct_age≥65            | 50.75          | 9.70   | 5.23    | 0.0000     | 31.61      | 69.89      |
| %total_black          | 5.77           | 1.86   | 3.10    | 0.0022     | 2.10       | 9.43       |

### Table 10
Regression results when including %total_asian population as an additional independent variable.

| Coefficients          | Standard Error | t Stat | P-value | Lower 95%  | Upper 95%  |
|-----------------------|----------------|--------|---------|------------|------------|
| Intercept             | −484.66        | 238.91 | −2.03   | 0.0441     | −956.29    | −13.02     |
| avg_household_size_adj| 1081.07        | 86.49  | 12.50   | 0.0000     | 910.34     | 1251.80    |
| pct_below_poverty_level| 12.27        | 4.89   | 2.51    | 0.0131     | 2.61       | 21.93      |
| pct_age≥65            | 55.48          | 9.45   | 5.87    | 0.0000     | 36.82      | 74.13      |
| %total_asian_alone    | −14.73         | 3.17   | −4.64   | 0.0000     | −20.99     | −8.47      |

### Table 11
Maximum, minimum and average values of coefficients for regression including %total_hispanic population as an additional independent variable for 15 day period.

| Coefficients          | Minimum       | Average      | Maximum       |
|-----------------------|---------------|--------------|---------------|
| avg_household_size_adj| 875.61725     | 911.73546    | 931.62228     |
| pct_below_poverty_level| 2.6686427    | 4.0941206   | 5.226449      |
| pct_age≥65            | 51.994143     | 55.349999    | 58.726034     |
| %total_hispanic       | 1165.5236     | 1238.7307    | 1310.935      |

### Table 12
Maximum, minimum and average values of coefficients for regression including %total_black_nonhispanic population as an additional independent variable for 15 day period.

| Coefficients          | Minimum       | Average      | Maximum       |
|-----------------------|---------------|--------------|---------------|
| avg_household_size_adj| 916.27133     | 955.9674     | 981.29695     |
| pct_below_poverty_level| 13.036153    | 15.172045    | 17.115713     |
| pct_age≥65            | 47.943724     | 51.047048    | 54.177515     |
| %total_black_non_hispanic| 553.63002   | 578.35984    | 600.34054     |

### Table 13
Maximum, minimum and average values of coefficients for regression including %total_black population as an additional independent variable for 15 day period.

| Coefficients          | Minimum       | Average      | Maximum       |
|-----------------------|---------------|--------------|---------------|
| avg_household_size_adj| 916.17228     | 955.76292    | 980.97525     |
| pct_below_poverty_level| 12.10048     | 14.187076    | 16.112607     |
| pct_age≥65            | 48.251701     | 51.369147    | 54.503164     |
| %total_black          | 542.78349     | 567.9981     | 588.8834      |

### Table 14
Maximum, minimum and average values of coefficients for regression including %total_asian population as an additional independent variable for 15 day period.

| Coefficients          | Minimum       | Average      | Maximum       |
|-----------------------|---------------|--------------|---------------|
| avg_household_size_adj| 1030.0448     | 1074.2913    | 1101.9448     |
| pct_below_poverty_level| 10.529976    | 12.603313    | 14.40379      |
| pct_age≥65            | 7.94E-09      | 56.092721    | 59.369881     |
| %total_asian_alone    | −1514.9971    | −1471.4693   | −1421.1716    |
studying 5700 patients in the NYC area, identified hypertension, obesity, and diabetes as the most common comorbidities. These comorbidities are known to be more prevalent among the black and Hispanic population segments, thus providing an explanation for the observed ethnic/racial associations. In turn, increased prevalence of these comorbidities is, in large part, the result of poor dietary habits and more limited access to healthcare, themselves driven by lower incomes.

Future research should investigate other crowding factors, e.g., numbers of individuals residing in nursing homes or dormitories.

**Financial Support**

The authors received no financial support for this study.

**Conflicts of Interest**

The authors have no conflict of interest with respect to this study.

To our knowledge, no outside individual or institution has a material interest in the results of this study.

**Declaration of interests**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Signed by
Awi Federgruen and Sherin Naha

**Acknowledgement**

We express our gratitude to Dr. Daniel Berman, Infectious Disease Specialist at Montefiore Hospital, Professor Jing Dong of the Columbia Business School, Judith Jacobson, Professor of Clinical Epidemiology at the Columbia School of Public Health, as well as Professor Assaf Zeevi of the Columbia Business School for their insightful suggestions and comments with respect to earlier drafts of this paper.

**Table 15**

Correlation coefficients among different variables.

| Average household size | Percent below poverty line | Percent above age 65 | Percentage of total Hispanic population | Percentage total Black non-Hispanic population | Percentage total Black population | Percentage total Asian population | # confirmed cases/100,000 residents |
|------------------------|---------------------------|---------------------|----------------------------------------|---------------------------------------------|---------------------------------|---------------------------------|---------------------------------|
| 1.00                   | 0.19                      | –0.21               | 0.26                                   | –0.56                                       | 0.26                            | 0.34                            | 0.62                            |
| 0.19                   | 1.00                      | 0.37                | 0.65                                   | 0.28                                        | 0.34                            | 0.29                            | 0.23                            |
| –0.21                  | –0.37                     | 1.00                | –0.34                                  | –0.13                                       | –0.16                           | 0.19                            | 0.09                            |
| 0.34                   | 0.65                      | –0.34               | 1.00                                   | –0.02                                       | 0.05                            | –0.26                           | 0.42                            |
| 0.26                   | 0.28                      | –0.13               | –0.02                                  | 1.00                                        | 0.99                            | –0.48                           | 0.34                            |
| 0.07                   | –0.29                     | 0.19                | –0.26                                  | –0.48                                       | –0.51                           | 1.00                            | –0.19                           |
| 0.62                   | 0.23                      | 0.09                | 0.42                                   | 0.34                                        | 0.35                            | –0.19                           | 1.00                            |

**References**

Barr J, Tassier T. Are Crowded Cities the Reason for the COVID-19 Pandemic?. Scientific American 2020; https://blogs.scientificamerican.com/observations/are-crowded-cities-the-reason-for-the-covid-19-pandemic/.

Bassett MT. Opinion | The Coronavirus Isn’t Worse in Cities Because of Density - The New York Times. 2020. https://www.nytimes.com/2020/05/15/opinion/sunday/coronavirus-cities-density.html.

Becker NG. On a general stochastic epidemic model. Theoretical Population Biology 1977;11(1):23–36, doi:http://dx.doi.org/10.1016/0025-5564(77)90004-1.

Becker NG, Dietz K. The effect of household distribution on transmission and control of highly infectious diseases. Mathematical Biosciences 1995;127(2):207–19, doi:http://dx.doi.org/10.1016/0265-9266(94)00055-5.

Bendavid E, Mulaney B, Sood N, Shah S, Leng E, Bromley-Duffino R, Lai C, Weissberg Z, Saavedra R, Tedrow J, Tversky D, Bogan A, Kupiec T, Etchner D, Gupta R, Ioannidis J, Bhattacharya J. COVID-19 Antibody Serorelevance in Santa Clara County, California. MedRxiv 2020; doi: http://dx.doi.org/10.1101/2020.04.14.20064980. 2020.04.14.20064980.

Corns-Herrera A, Zalakain J. Mortality associated with COVID-19 outbreaks in care homes : early international evidence. In Resources to support community and institutional Long-Term Care responses to COVID-19 (Issue April), LTCovid.org; 2020.

Godoy M, Wood D. What Do Coronavirus Racial Disparities Look Like State By State?. National Public Radio (NPR); 2020. https://www.npr.org/sections/health-shots/2020/05/30/865413079/what-do-coronavirus-racial-disparities-look-like-state-by-state.

Illinois Department of Public Health. COVID-19 Statistics. Retrieved from Illinois Department of Public Health; 2020. https://www.dph.illinois.gov/covid19-statistics.

Millet GA, Jones AT, Benesder D, Baral S, Mercer L, Beyer C, Honennery B, Lanjkiewicz E, Mena L, Crowley JS, Sherwood J, Sullivan P. Assessing Differential Impacts of COVID-19 on Black Communities. Annals of Epidemiology 2020; doi: http://dx.doi.org/10.1016/j.amepi.2020.05.003.

Mississippi State Department of Health. Coronavirus Disease 2019 (COVID-19). Retrieved from Mississippi State Department of Health; 2020. https://msdh.ms.gov/msdhsite/_static/140,420.html.

Rader B, Scarpino S, Nande A, Hill A, Dalziel B, Reiner R, Pigott D, Gutierrez B, Shrestha M, Brownstein J, Castro M, Tian H, Greshell B, Pybus O, Metcalf J, Kraemer MUG. Crowding and the epidemic intensity of COVID-19 transmission. MedRxiv 2020; doi: http://dx.doi.org/10.1101/2020.04.15.20064980. 2020.04.15.20064980.

The Novel Coronavirus Pneumonia Emergency Response Epidemic Team. Vital Surveillances: The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) — China, 2020. February Retrieved from Chine CDC Weekly; 2020. http://www.chinacdc.cn/en/article/id/e5359462-c654-41e9-a0db-a3a8bf51.

Virginia Department of Health. COVID-19 in Virginia. Retrieved from Virginia Department of Health; 2020. http://www.vdh.virginia.gov/coronavirus/.

Wadhera RK, Wadhera P, Gaba P, Figueroa JF, Joyn Maddox KE, Yeh RW, Shen C. Variation in COVID-19 Hospitalizations and Deaths Across New York City Boroughs. JAMA - Journal of the American Medical Association 2020;323(21):2192–5, doi:http://dx.doi.org/10.1001/jama.2020.7197.

Yancy CW. COVID-19 and African Americans. JAMA - Journal of the American Medical Association 2020;323(19):1891–2, doi:http://dx.doi.org/10.1001/jama.2020.6548. American Medical Association.