Disparities in COVID-19 Risk Exposure: Evidence from Geolocation Data[*]

Milena Almagro[†] Joshua Coven[‡] Arpit Gupta[§]
Angelo Orane-Hutchinson[¶]

February 23, 2021

Abstract

We examine the determinants of COVID-19 risk exposure in the context of the initial wave in New York City. During the beginning of the first wave of the pandemic, out-of-home activity related to commuting was strongly associated with COVID-19 cases at the ZIP code level and hospitalization at an individual level. After layoffs of workers decreased commuting, case growth continued through household crowding. A larger share of individuals in crowded housing, or commuting to essential and frontline work, are Black, Hispanic, and lower-income—which contributes to disparities in disease risk. As a result, our paper shows that structural socio-economic inequalities help determine the cross-section of COVID-19 risk exposure in urban areas.

[*] The views expressed in articles and other content on this website are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. This paper previously circulated under the title: "Racial Disparities in Frontline Workers and Housing Crowding during COVID-19: Evidence from Geolocation Data."

[†] Federal Reserve Bank of Minneapolis and Booth School of Business, University of Chicago. Email: milena.almagro@chicagobooth.edu

[‡] Department of Finance, Stern School of Business, New York University. Email: joshua.coven@stern.nyu.edu

[§] Department of Finance, Stern School of Business, New York University. Email: arpit.gupta@stern.nyu.edu

[¶] Department of Economics, New York University. Email: angelo.orane@nyu.edu

I Introduction

The novel coronavirus disease 2019 (COVID-19) has disproportionately and negatively impacted disadvantaged populations. The hardest-hit regions of New York City include parts of the Bronx, Brooklyn, and Queens with high fractions of Black, Hispanic, and low-income populations as has been noted by Borjas (2020) and Schmitt-Grohé et al. (2020).[1] Nationwide, infections are three times as likely among Latinos and African-Americans compared to infections among whites.[2] While the disparities in COVID-19 disease burdens across dimensions of income, race, and ethnicity have been widely recognized, the ultimate drivers of these inequities remain unclear.

This paper identifies two fundamental drivers of COVID-19 risk exposure: out-of-home behavior and housing crowding, which help account for the cross-sectional disease burden of different population groups. We focus on mobility measures which serve as proxies for the social interactions and direct contacts which place individuals at risk for infection. We also analyze household crowding because of the importance of physical proximity as another risk factor in coronavirus exposure. In contrast to population density, which measures the concentration of individuals within a geographic area, housing crowding measures the number of close physical contacts. It provides a more granular and heterogeneous measure of individual risk, and follows prior medical literature which has highlighted the role of within-household spread of coronavirus, particularly in China.[3]

We use anonymized mobile phone Global Positioning System (GPS) data to create and analyze our risk factors at the individual, building, and neighborhood levels, which allows us to control for important ZIP code level demographic variables. To do so, we link individual mobile phone data with ZIP code-level data on daily COVID-19 infections, as well as census data on occupation and household occupancy. We then use within-tract variation in commuting and housing characteristics to identify the impact of mobility on COVID-19 risk.

[1] We document the demographic associations of COVID-19 in Section III.A. New York City official data suggest that African-Americans were 59% more likely to be diagnosed with COVID-19 relative to whites, while Hispanics were 64% more likely. See: https://www1.nyc.gov/site/doh/covid/covid-19-data.page.

[2] See data from the C.D.C. https://www.nytimes.com/interactive/2020/07/05/us/coronavirus-latinos-african-americans-cdc-data.html.

[3] See Jing et al. (2020).

We focus on New York City, the global epicenter for the pandemic in Spring 2020. Our work splits the first wave of the COVID-19 pandemic in New York City into two periods. In the initial stage of the crisis—lasting from March until early April—we document that the commuting behavior of essential and frontline workers placed them at greater risk of infection. Namely, during our first week of analysis (March 25th to March 31st), a 10% increase in the number of hours residents of a building spend outside the home is associated with a 15.5% increase in the number of hospitalizations per housing unit for that building.

We find that this association decreased after early April, the start of the second stage of the crisis, when many workers were laid off. At this point, disease spread continued through a household crowding channel—by the week of April 8th to April 14th, the previous coefficient for hours outside the home declines to zero, while the effect of a 10% increase in crowding correlates with a 12.4% increase in hospitalizations per unit. Both measures remain significantly associated with infection when included together. Importantly, we find that racial minorities and low-income individuals are over-represented in both risk measures, pointing to important disparities in disease exposure.

Our results suggest sizeable effects of mobility on disease exposure: increasing time outside of an individual's home census tract from the 10th to the 90th percentile is associated with a 4.2 times higher hazard rate of hospitalization. Similarly, individuals at the 90th percentile of housing crowding have a 2.2 times higher hazard rate than those at the 10th percentile. The importance of both the commuting and housing crowding channels highlights concerns for policies that focus on one specific transmission route while neglecting the other. For example, shutting down workplaces or outdoor public spaces through lockdowns may lower infectious spread through a commuting or workplace channel, but may instead result in individuals interacting more in crowded home settings.

Our analysis also has implications for ongoing debates on the role of density and urban form on disease exposure. In contrast to research which emphasizes the role of static characteristics of urban design such as density (Duranton and Puga, 2020; Carozzi et al., 2020) or subways (Harris, 2020), we highlight the dynamic responses of individuals and groups which depend on access to preexisting resources. Notably, Manhattan—the densest and wealthiest borough—saw many fewer infections than the other boroughs.

Our results suggest that the temporary pockets of density created by the mobility patterns of frontline workers, and the physical proximity of housing crowding, matter more than a static picture of density. We document that these mobility-induced densities are inequitably experienced by vulnerable populations. In turn, structural inequalities lead disadvantaged groups to disproportionately live in crowded housing and specialize in jobs that require physical presence, and therefore to increased COVID-19 exposure.

A limitation of our analysis is that we are unable to observe demographic associations at the individual and building levels, where we are able to identify the effects of COVID-19 risk exposure. As a result, we are not able to fully identify the effects of our estimated disease exposure in explaining the entire racial disparity gap. Instead, our work seeks to identify the drivers of risk exposure at a fine-grained level as resulting from inequities in occupation and housing, and document disparities in the exposure of different populations to these risk factors.

We contribute to a growing literature on COVID-19 by emphasizing both income and racial disparities within urban areas and providing direct evidence of the role of both commuting and housing-related disparities in the spread of COVID-19. Many papers have used geolocation data in the context of COVID-19 (Chen et al., 2020; Chiou and Tucker, 2020; Couture et al., 2020; García-Lopez and Puga, 2020). Our work is most closely related to Glaeser et al. (2020). We differ in four key ways. First, we consider both aggregated data and individual-level mobility data, which allows us to identify individual risk factors for hospitalization. Second, a central focus is examining the role of racial and income disparities in disease burden. Third, we consider an additional housing crowding dimension which was crucial at the stage in the pandemic when many workers stopped commuting due to a combination of job loss and remote work. Finally, we contribute on the identification side by constructing a panel of buildings where our main outcome variable is the hospitalization of a building's resident, which allows us to control for daily unobservables that are common across all individuals who live in buildings in the same census tract.

We build on methods used in prior works such as Athey et al. (2019), Chen et al. (2019), and Chen and Rohla (2018), which used mobile phone geolocation data to examine segregation, racial disparities in voting waiting times, and partisanship.

A growing literature also examines racial disparities specifically in the context of COVID-19 (Borjas, 2020; McLaren, 2020; McCormack et al., 2020; Almagro and Orane-Hutchinson, 2020; Sá, 2020; Karaca-Mandic et al., 2020). Our work adds to this literature by linking important risk components of the racial disparity in case exposure. We complement medical literature, such as Rentsch et al. (2020) and Price-Haywood et al. (2020), which suggests that population differences in risk exposure must account for racial disparities, given that there are no racial differences in mortality among the hospitalized. We also add to a broader medical literature which discusses the role of social factors that lead to disparities in mortality, such as Wong et al. (2002) and Trivedi et al. (2005).

II Data

II.A Geolocation Data

Mobile location data were sourced from VenPath, a holistic global provider of compliant smartphone data. Our data provider aggregates information from approximately 120 million smart phone users across the United States. GPS data were combined across applications for a given user to produce "pings" corresponding to time stamp–location pairs. The provider anonymizes information on individual users. Ping data include both background pings (location data provided while the application is running in the background) and foreground pings (activated while users are actively using the application). Our sample period covers February 1st – July 12th, 2020. Appendix A describes the filters that we apply to our raw ping data in detail.

II.B Constructing individual risk measures: Mobility and household crowding

To create the mobility measures, we first identify residents of NYC, as described in Appendix A. To categorize their activity as at-home or out-of-home-tract, we identify each mobile phone user's modal tract between 6pm and 8am on each date. We then designate the most frequently observed nightly tract as their "home census tract" (HCT). To measure out-of home-behavior, we count the number of hours a user spends entirely outside of the user's HCT during the range of 8am to 10pm, conditional on the user having ping data during those hours.

To define our metric of housing crowding, we identify each user's modal building each night. We count the number of unique users in each building for whom that building is their modal home building that night. We divide this number by the amount of residential housing units in that building to calculate the people per housing unit on each date.

II.C Measuring Hospitalizations

While we cannot see whether an individual has been tested in our data, we can observe whether an individual pings inside a hospital. Our measurement of individual hospitalizations is an important contribution to the literature, which has generally focused on cases measured at more aggregate levels—and hence has been unable to control for important local covariates. To attribute pings to hospitals in New York city, we connect building shapefiles provided by Microsoft to a list of hospitals provided by Homeland Infrastructure Foundation-Level Data using latitudes and longitudes.[4] We include hospitals within New York City that are not long-term care facilities or psychiatric hospitals.[5]

We classify a user to be hospitalized if we observe her pings for more than a day within a hospital. We focus on the first month of the stay-at-home order, between March 22th and April 22nd, when other non-essential hospitalizations were postponed to maximize the probability that our measure corresponds to actual COVID-19 hospitalizations. We restrict our sample to individuals with at least ten observed days during March and April. We also filter out potential hospital workers by excluding those who ping in a hospital immediately after they enter the data. We are left with 53,558 unique individuals and 219 hospitalizations, a hospitalization rate of 0.41%.[6] To verify that our measured hospitalizations line up with other data sources, we compare with actual hospitalization data for COVID-19 from the NYC Department of Health in Appendix Figure A2. Across the period from March 22th – April 22nd, we find a correlation of 0.76 between our measure of hospitalization and actual hospitalizations, suggesting that we are able to accurately estimate individual hospitalized COVID-19 cases.

II.D Census and Occupation Shares

We obtain demographic and occupation data at the ZIP code and census tract level from the American Community Survey (ACS). We include ZIP code median income, average age, racial breakdown, and health insurance status. We also include commuting-related variables: average commute time to work as well as means of transportation.

We also construct the shares of the working-age population employed in different occupation categories. We first divide occupation between flexible and non-flexible occupations. Then, we categorize non-flexible occupations according to their essential definition

[4] See: https://hifld-geoplatform.opendata.arcgis.com/datasets/6ac5e325468c4cb9b905f1728d6fbf0f_0.

[5] For the full list of hospitals see Appendix Section A.

[6] According to official data provided by the DOH, the hospitalization rate for NYC was 0.46%.

and similarity in work environments and social exposure. The summary statistics of demographics and occupations can be found in Appendix A, which also breaks out the occupational groups in the non-flexible category separately.

II.E Aggregating Mobility and Crowding Measures

First, for our individual risk measures, we take a seven-day moving average to reduce the noise of daily raw values and to account for weekly seasonality. Second, we take a two-week lag to account for the delay between exposure and the event of being hospitalized. We keep only the observations whose lagged seven-day moving average do not have any missing dates to eliminate mechanical differences across different days of the week. To aggregate our mobility and crowding metrics to the ZIP code level, we use the geospatial shapes of NYC's census tracts provided by NYC Open Data.[7] Table A1 in the Appendix presents the summary statistics of our aggregate measures.

III Results

III.A Descriptive Analysis

We begin with a descriptive analysis of our sample to highlight the key features of the COVID-19 pandemic in New York City. Appendix Figure A1 shows how our risk measures evolve over time. We observe a decrease in mobility in early March that started prior to the stay-at-home order issued by Governor Cuomo on March 20th. Our finding that mobility responds primarily to the pandemic, rather than the state-imposed order, is consistent with similar nation-wide findings in Goolsbee and Syverson (2020). Mobility in our sample hits a low in early April before recovering later in our sample. On the other hand, we do not see any stark trend for the average number of people per housing unit.

We also contrast the time series of mobility measures in Appendix Figure B1 across different boroughs of NYC. We observe the greatest sheltering response in Manhattan (New York County) and the lowest in Brooklyn (Kings County). The differential patterns across boroughs may reflect the ability of different populations to shelter effectively given

[7] We link to ZIP codes using a crosswalk provided by the Department of Housing and Urban Development. See: https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku for the list of tracts and https://www.huduser.gov/portal/datasets/usps_crosswalk.html for the crosswalk. We select the ZIP and tract mapping that has the highest number of residents residing in the ZIP for a given tract to get a 1:1 mapping of tracts to ZIP codes.

the tendency for frontline jobs to be precarious and non-local. We compare across both the time series and the cross-section in Appendix Figure B2. In the key months of the pandemic, through April and May, measured mobility patterns show sheltering in certain high-income neighborhoods of Manhattan and Brooklyn—while residents in lower-income regions of Brooklyn and other boroughs were much more likely to spend time outside of their home tract. Finally, Appendix Figure B3 shows the spatial variation of the housing crowding measure, averaged across our sample. We tend to observe greater housing crowding in the outer boroughs of the city.

For our cross-section analysis we plot some basic correlations between our risk measures and housing density measures with demographics and occupations. Panel A of Figure B4 shows correlations of mobility with certain neighborhood demographics: the fraction of tract residents who are Black, the fraction who have flexible occupations, and income. We find substantial positive correlations of increased out-of-home mobility in areas with more low-income or Black populations. We also observe a positive correlation between crowded spaces and neighborhoods with a higher share of minorities and lower percentage of flexible workers in Panel B of Figure B4. This last finding highlights the fact that frontline workers are exposed to higher risk through both risk measures.

III.B ZIP Code-Level Analysis

Having established our basic variables, we turn next to a deeper analysis of the relationship between structural inequalities and the incidence of COVID-19. For each building, we estimate the daily number of hospitalizations per unit. We then construct a panel of the daily average of hospitalizations per unit across New York City ZIP codes from March 23rd to April 22nd. For this specification, we estimate the following equation:

$$\text{hospitalizations per unit}_{jt} = \beta_1 \text{mobility}_{jt} + \beta_2 \text{housing density}_{jt} + \gamma X_j + \mu_t + \varepsilon_{jt}$$

where X_j contains demographic and occupational characteristics at the ZIP code level and μ_t is a day fixed effect that controls for the aggregate evolution of the pandemic in New York City. To corroborate that demographic associations of test positivity across ZIP codes are not driven by other factors, New York City official data also show large racial disparities in diagnosis and death rates.[8]

[8] See: https://www1.nyc.gov/site/doh/covid/covid-19-data.page.

Table 1 shows the estimation results of regressing the daily average of hospitalizations per unit across ZIP codes on mobility and housing crowding measures for several specifications that vary in their set of neighborhood controls. The first specification, column (1), includes only basic demographics such as race and income, while column (2) includes only occupations measures. Column (3) only includes our mobility and crowding variables, and Columns (4) and (5) add basic demographics and occupations, respectively. A comparison of column (1) with column (4) shows that the initial racial disparities are partially explained by differences in mobility patterns and housing density, as all coefficients for racial groups shrink towards zero. Similarly, some of the correlation between occupations and hospitalizations is also explained by the risk measures in mobility and crowding. Because demographics are measured only at the ZIP level, this specification relies on an aggregated version of our risk measures. As a result of this attenuation, we do not expect to necessarily account for the entirety of the measured ZIP-level disparities using our risk measures.

For this basic specification, column (3), which presents the interpretation of the magnitudes for our mobility measure, is as follows: if the number of hours outside the home census tract (HCT) increases by 10%, an average level increase of 0.06, the daily average number of hospitalizations per unit increases by 17.6%. On the other hand, a 10% increase in the number of people per unit corresponds to a 6.9% increase in the average hospitalizations per unit.

Table 1: Neighborhood Associations of Hospitalizations

Dependent Variable:	Hospitalizations per Housing Unit											
	(1) Race & Income		(2) Occup.		(3) Mobility		(4) Mobility, Race, & Income		(5) Mobility & Occup.		(6) Mobility, Race, Dem, & Occup.	
Hours out of home tract					0.005^{***}	(0.001)	0.004^{***}	(0.001)	0.004^{***}	(0.001)	0.003^{**}	(0.001)
People per unit					0.002^{***}	(0.000)	0.001^{***}	(0.000)	0.001^{***}	(0.000)	0.001	(0.001)
Log Income	0.003^{***}	(0.001)					0.002^{***}	(0.001)			0.004^{**}	(0.002)
% Black	0.008^{***}	(0.002)					0.006^{***}	(0.002)			-0.002	(0.002)
% Hispanic	-0.000	(0.001)					0.001	(0.001)			-0.006^{**}	(0.002)
% Asian	0.003^{**}	(0.001)					0.002	(0.001)			-0.005^{**}	(0.002)
% Flexible occupations			0.006^{**}	(0.003)					0.012^{***}	(0.003)	-0.002	(0.009)
% Health practitioners			0.023^{*}	(0.013)					0.009	(0.016)	-0.017	(0.029)
% Other health			0.100^{***}	(0.016)					0.096^{***}	(0.016)	0.101^{***}	(0.020)
% Firefighting			-0.020	(0.036)					-0.008	(0.037)	0.018	(0.039)
% Law enforcement			0.128^{***}	(0.038)					0.068^{*}	(0.038)	0.032	(0.037)
% Essential - Service			0.012	(0.009)					0.040^{***}	(0.010)	0.033^{***}	(0.010)
% Non ess. - Service			-0.067^{***}	(0.017)					-0.077^{***}	(0.017)	-0.099^{***}	(0.025)
% Ind. and Construction			0.012	(0.008)					-0.010	(0.009)	-0.032^{**}	(0.014)
% Essential - Technical			-0.047	(0.033)					-0.061^{*}	(0.033)	-0.154^{***}	(0.047)
% Transportation			0.072^{***}	(0.027)					0.050^{*}	(0.027)	0.009	(0.029)
Share $\geq 20, \leq 40$											-0.017^{***}	(0.006)
Share $\geq 40, \leq 60$											0.008	(0.012)
Share ≥ 60											0.001	(0.008)
Share Male											-0.007	(0.010)
Log Household Size											0.006^{**}	(0.003)
% Public Transport											0.003	(0.003)
Log Commute Time											-0.011^{***}	(0.003)
% Uninsured											0.032^{***}	(0.009)
Bronx											-0.000	(0.001)
Brooklyn											-0.001	(0.001)
Queens											-0.000	(0.001)
Staten Island											-0.000	(0.002)
Day FE	✓		✓		✓		✓		✓		✓	
N	4,340		4,340		4,340		4,340		4,340		4340	
adj. R^2	0.08		0.09		0.08		0.09		0.10		0.11	

Spatial HAC Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table shows the results of a regression of daily number of average hospitalizations per building units by ZIP code, scaled by a factor of 100, on different set of covariates over the period March 22rd to April 22nd. Column (1) includes only basic demographics such as race and income, while column (2) includes only occupation-related measures. Column (3) includes only our mobility and crowding measures. Columns (4) and (5) include mobility measures with race and occupation measures, respectively. Column (6) adds other neighborhood and demographic controls.

Examining only demographic variables shows evidence of racial and occupational disparities in hospitalization rates. Incorporating mobility, demographic, and occupational controls lowers the coefficient on fraction Black to zero, and the coefficient on fraction of Asian population becomes negative. This suggests that racial disparities in hospitalization rates, at least for these groups, can be accounted for by variation in background variables related to mobility and occupation.

Part of the variation in hospitalizations explained by occupations can also be linked to our risk measures. On one hand, the coefficients for law enforcement and transportation occupation shares—both essential and frontline, with a positive association with hospitalization rates—turn to zero once we account for mobility and crowding. On the other hand, the coefficient for Essential - Service becomes significant when risk measures are included. These changes imply that these workers are in particular higher risk of infection, even after accounting for crowding and mobility.

III.C Building-Level Analysis

In this section, we move to a more granular level of analysis by focusing on risk measures at the building level. This level of granularity allows us to address the important identification concern outlined in the previous section by exploiting variation at the building level, controlling for variation at the tract-day level. For any building, our outcome variable measures the number of hospitalizations per residential unit, where we classify individuals as being hospitalized if they spend more than 24 hours at a hospital. We focus on the first month after the issuance of the stay-at-home order to maximize the probability that new hospitalizations that we observe in our data are due to COVID-19 and not due to something else.[9]

We first start by reproducing specifications (1)–(6) of Table 1 but at a smaller level of aggregation. Table 2 shows the results of our regression analysis for specifications with different sets of neighborhood controls. Our main regression equation is:

$$\text{hospitalization}_{bt} = \alpha_1 \text{mobility}_{bt} + \alpha_2 \text{housing density}_{bt} + \gamma X_{j(b)} + \mu_t + \varepsilon_{bt},$$

[9] Using data available at `https://github.com/thecityny/covid-19-nyc-data` that constructed total hospitalizations from reports of Governor Cuomo's office, we observe that more than 50% of all hospitalizations for the first three weeks of April were related to COVID-19. This measure includes new hospitalizations as well as as patients with more long-term diseases or patients in palliative care. For the time-series correlation of our measure of hospitalizations and the official numbers, see Appendix A2.

where mobility_{bt} and $\text{housing density}_{bt}$ are respectively the average mobility and housing density measures for date t, $X_{j(b)}$ are demographic and occupational controls for the census tract where the building b is located, and μ_t are date fixed effects.

Next, to move to a more causal estimation of the impact of risk changes on outcomes, we control for census tract by day fixed effects, $\delta_{j(b)t}$, to account for daily factors common to all individuals within a census tract. That is, our regression equation in this case is:

$$\text{hospitalization}_{bt} = \alpha_1 \text{mobility}_{bt} + \alpha_2 \text{housing density}_{bt} + \delta_{j(b)t} + \varepsilon_{bt}.$$

Our identifying assumption for this specification is based on the hypothesis that unobservables with temporal variation that correlate with mobility and housing density measures are common to all residents in buildings in the same census tract and day. Our identifying variation comes from differences in individuals within census tract and day. We can include demographics only at the census tract level because we do not measure demographics for buildings or individuals. In this analysis we cluster standard errors at the building level.

The interpretation of the coefficients is as follows. For our preferred specification, column (7), if a building's residents increase their number of hours outside the HCT by 10%, the number of hospitalizations per occupant in that building increases by 2.5%.[10] Similarly, if a building's number of people per housing unit increases by 10%, we expect to see hospitalizations per occupant increase by 6.8%. We also observe that coefficients on the risk measures are stable across specifications, suggesting that unobservables are not producing meaningful biases in the coefficients (Oster, 2019).

[10] The average number of hospitalizations per unit is 3.8502e-5 and the average number of pings per building is 0.6.

Table 2: Building Level Hospitalizations

Dependent Variable:	Hospitalizations per Unit													
	(1) Race & Income		(2) Occup.		(3) Mobility		(4) Mobility, Race, & Income		(5) Mobility & Occup.		(6) Mobility, Race, Dem, & Occup.		(7) Census Tract × Day	
Hours out of home tract					0.001***	(0.000)	0.001***	(0.000)	0.001***	(0.000)	0.001***	(0.000)	0.001***	(0.000)
People per unit					0.001***	(0.000)	0.001***	(0.000)	0.001***	(0.000)	0.001***	(0.000)	0.001***	(0.000)
Log Income	0.002**	(0.001)					0.001	(0.001)			0.002	(0.001)		
% Black	0.002	(0.001)					0.001	(0.001)			-0.000	(0.002)		
% Hispanic	-0.001	(0.001)					-0.001	(0.001)			-0.002	(0.002)		
% Asian	-0.001	(0.002)					-0.002	(0.002)			-0.004	(0.003)		
% Flexible occupations			-0.003	(0.003)					-0.002	(0.004)	-0.008	(0.006)		
% Health practitioners			0.004	(0.011)					-0.006	(0.011)	-0.016	(0.013)		
% Other health			0.006	(0.008)					0.005	(0.008)	-0.010	(0.010)		
% Firefighting			-0.028	(0.017)					-0.036**	(0.018)	-0.043**	(0.018)		
% Law enforcement			0.042*	(0.024)					0.023	(0.023)	0.022	(0.028)		
% Essential - Service			-0.010	(0.007)					-0.005	(0.007)	-0.005	(0.008)		
% Non ess. - Service			-0.018*	(0.011)					-0.015	(0.011)	-0.019	(0.012)		
% Ind. and Construction			-0.015**	(0.006)					-0.020***	(0.007)	-0.030***	(0.010)		
% Essential - Technical			0.030	(0.027)					0.015	(0.026)	0.001	(0.026)		
% Transportation			0.009	(0.011)					0.003	(0.011)	-0.003	(0.011)		
Share $\geq 20, \leq 40$											0.002	(0.006)		
Share $\geq 40, \leq 60$											0.002	(0.010)		
Share ≥ 60											0.009	(0.006)		
Share Male											-0.006	(0.006)		
Log Household Size											0.004	(0.003)		
% Public Transport											-0.001	(0.003)		
Log Commute Time											0.000	(0.002)		
% Uninsured											0.008*	(0.004)		
Bronx											-0.001	(0.001)		
Brooklyn											-0.000	(0.000)		
Queens											-0.001	(0.001)		
Staten Island											-0.003*	(0.002)		
Day FE	✓		✓		✓		✓		✓		✓			
N	1,555,842		1,555,842		1,555,834		1,555,834		1,555,834		1,555,834		1,555,834	

Standard errors are clustered at the building-level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table shows the results of a building-level regression of mobility measures against hospitalizations per unit. We regress hospitalizations per unit on risk measures and demographics. Column (1) includes only basic demographics such as race and income, while column (2) includes only occupation-related measures. Column (3) includes only our mobility and crowding measures. Columns (4) and (5) include mobility measures with race and occupation measures, respectively. Column (6) adds other neighborhood and demographic controls. Column (7) controls for Census Tract × Day fixed effects.

III.C.1 Building-level weekly analysis

Motivated by the dynamic evolution of different channels of transmission, we estimate the following equation:

$$\text{hospitalization per unit}_{bt} = \alpha_{1,w(t)}\text{mobility}_{bt} + \alpha_{2,w(t)}\text{housing density}_{bt} + \mu_{j(b),y} + \varepsilon_{bt},$$

where coefficients are allowed to change week by week as denoted by subindex $w(t)$.

Table 3: Weekly Analysis of Mobility Exposures and Hospitalization

Dependent Variable:	Hospitalizations per Unit			
	(1) Mar 25–31	(2) Apr 1–7	(3) Apr 8–14	(4) Apr 15–21
Hours Outside of Home Tract	0.003**	0.002***	0.001	0.000
	(0.001)	(0.000)	(0.001)	(0.000)
People per Unit	0.001***	0.002***	0.002***	0.001**
	(0.001)	(0.001)	(0.001)	(0.001)
Census Tract × Day FE	✓	✓	✓	✓
N	355,607	353,647	360,603	346,117

Standard errors clustered at the building level

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table shows the results of a building-level regression of mobility measures against hospitalizations per unit. We regress hospitalizations per unit on the risk measures and include census tract interacted with day fixed effects.

Table 3 shows that the coefficients for mobility patterns had a larger impact at early stages of the pandemic and that they decrease in magnitude over time, similar to Glaeser et al. (2020). We also find a similar pattern for housing density; its effect increases in importance as the stay-at-home order kicks in. During week 1 (March 25th to March 31st), a 10% increase in number of hours outside or in housing crowding is associated with a 15.5% increase in the number of hospitalizations per unit. By week 3 (April 8th to April 14th), the coefficient for mobility patterns declines to zero, while a 10% increase in crowding correlates with a 12.4% increase in hospitalizations per unit—suggesting that housing density gained importance with the progression of the pandemic, the issuance of the stay-at-home order, and the large economic shock that led to high unemployment.

III.D Individual-Level Analysis

In this section we present results obtained using anonymized individual-level data. We construct risk measures—mobility and housing crowding—for individual mobile phone users we can track over time. This allows us to see how these individual risk measures correlate with demographics and occupational categories.

However, we face an important challenge of censoring. The event of being hospitalized due to COVID-19 generally happens only once with a probability that increases over time. To appropriately account for this censoring issue, as well as the fact that the probability of hospitalization is not independent of what happened in the past, we borrow tools from the survival analysis literature.

For our survival analysis, we focus again on the days between March 22$^{\text{nd}}$ and April 22$^{\text{nd}}$, where non-essential hospitalizations were postponed to prioritize COVID-19 patients. We take the hazard rate of being hospitalized as the outcome variable for the individuals in our sample. For our risk measures, we take the two-week lag relative to the last time we observe each individual, which can be either the end of our time period, the date that the individuals is hospitalized, or if the individual leaves the sample before those two events. We link each individual to her modal home ZIP code across days to match individuals with demographics. After all of these restrictions, we are left with 23,850 individuals with 81 hospitalizations, which amounts to a hospitalization rate of 0.34%.

We start by plotting Kaplan-Meier graphs with the cumulative probability of failure on a daily time scale in Figure D1. We observe that the probability of being hospitalized increases over time. The other two graphs plot hazard rates for two population groups corresponding to above and below median of number of hours outside HCT and average number of people in the same housing unit plotted in Panel B and Panel C, respectively. We observe that for both of our risk measures, an individual above the median is associated with a higher hazard rate of being hospitalized, with a difference that also increases over time.

III.D.1 Survival Analysis

In this section we present estimation results from a semiparametric Cox regression. We define a failure as the event of being hospitalized for the individuals in our sample. This type of estimation constructs hazard rates of being hospitalized nonparametrically and then uses the log of such hazard rates as the outcome variable in a regression, where covariates can be similarly defined as in any standard linear regression.

Our results from the Cox regression highlight the central role of out-of-home mobility and housing crowding in determining individual hospitalization rates. First, we observe for mobility patterns across columns (1)–(3) of Table 4 a similar pattern as in Table 3: including demographics and occupational controls decreases the magnitude of the coefficient. This suggests that part of the mobility patterns can be mediated by occupations and demographics.

Moreover, we can employ a similar identification strategy as in our aggregate analysis at the census tract level. Unfortunately, Cox regressions do not allow fixed effects at the same level as the temporal unit level, which in this case is days. Hence, we cannot include day fixed effects. To overcome this problem, we estimate time trends at the Community District (CD) level.[11] Our identifying assumption is that unobservables that correlate with our risk measures can be described by time trends at the CD level and our identifying variation comes from daily differences in risk measures for individuals who live in the same CD. However, we do not see stark differences where we compare column (4) to column (5), suggesting that unobservables in captured by CD time trends do not produce meaningful biases.

Reassuringly, Table 4 shows similar results as Table 3: After controlling for demographics and occupations the coefficients on risk measures decrease which can be explained again by the correlation between certain demographics and occupations to risk exposure. Full results for this regression can be seen in Table D1.

In our preferred specification, column (5) of Table 4, a 1% increase number of hours outside HCT increases the hazard rate of being hospitalized by 45%. Similarly, a 1% increase in the number of people per unit corresponds to a hazard rate that is 25% higher.

Given that our survival analysis uses individual data, we can estimate a distribution of exposure to the disease across the individuals of our sample and see how that distribution correlates with demographics and our risk measures. For example, we can construct the 10–90 percentile range the distributions of our risk measures. When we do so, we find that an individual at the 90th percentile of mobility has a hazard rate of being hospitalized that is 4.2 times higher compared to an individual at the 10th percentile. Similarly, an

[11] NYC has 42 CDs. Unfortunately, it is computationally unfeasible to estimate time trends at smaller geographical units such as ZIP codes.

Table 4: Cox Regressions on Risk Measures and Demographics

Dependent Variable:	Hazard Rate of Being Hospitalized				
	(1) Risk	(2) Risk, Race & Income	(3) Risk, Race, Dem. & Occup.	(4) Comm. District Fixed Effects	(5) Comm. District Time Trends
Log hours outside HCT	0.369***	0.360***	0.357***	0.370***	0.369***
	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)
Log people per unit	0.156**	0.154*	0.204*	0.223**	0.221**
	(0.064)	(0.080)	(0.106)	(0.089)	(0.089)
CD Fixed Effect				✓	
CD Time Trend					✓
N	23,756	23,748	23,748	23,553	23,756

Robust Standard Errors in parenthesis.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

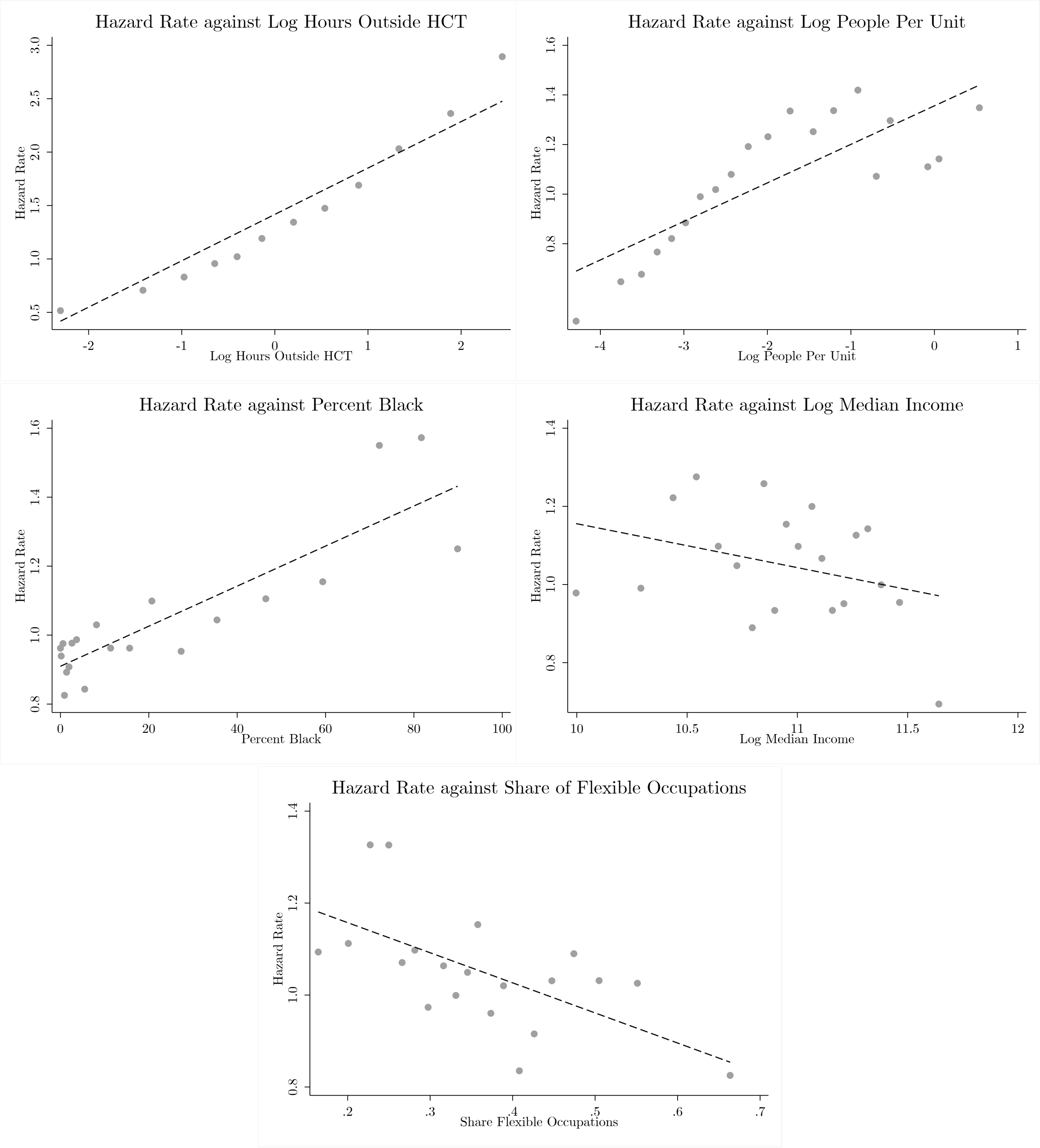
Note: This table shows the results of the regression of the cumulative hazard rate of being hospitalized on different sets of covariates. Column (1) includes only basic demographics such as race and income, while column (2) includes only our risk measures. Column (3) includes all of these covariates together. Columns (4) expands by adding more demographics and occupational shares. Column (5) includes Community District fixed effects and Column (6) adds Community District-specific linear trends.

individual at the 90th percentile of housing crowding has a hazard rate of hospitalization that is 2.2 times higher compared to that of an individual at the 10th percentile.

Finally, Figure 1 shows how the distribution of predicted hazard rates of hospitalization using model (3) of Table 4 correlate with demographics.[12] We see clear upward trends of predicted hazard against our risk measures. We also observe a positive relationship between the share of a population in a ZIP code that is Black and the hazard rate, even though in our Cox regressions the coefficient on the fraction of a population that is Black is not significant whenever risk measures are included as covariates. Moreover, income and flexible occupations are never significant but nevertheless present a clear positive relationship with the predicted hazard rate of being hospitalized. We take these patterns as evidence that structural inequalities, rather than demographics, expose vulnerable populations to greater risk of contagion through channels of mobility and housing crowding.

[12] We choose model (3) to compare the plotted correlations presented in the graphs with the estimated coefficients on demographics and risk measures of such regression equation.

Figure 1: Predicted Hazard Rate of Hospitalization against Demographics and Risk Measures



Note: These graphs are binscatter plots of the predicted hazard rates according to model (3) of Table 4 against different demographics and risk measures.

IV Conclusion

Our work documents the pathways of COVID-19 risk exposure. We focus on the epicenter of the global pandemic in New York City, showing that infections spread through two channels. Initially, infections spread through essential workers, who continued to commute to establishments.

We demonstrate these links using novel data drawn from cell phones to measure these mobility patterns, which we use to establish a direct link between outside mobility at both neighborhood and individual levels. Our individual-level analysis advances on prior research using geolocation data by directly linking greater mobility for individual workers and presence in hospitals, controlling for other unobserved local factors.

We also connect both cell phone mobility and census data on housing occupancy. We find that housing overcrowding predicts a greater caseload, and we also document more Black, Hispanic, and low-income households reside in overcrowded buildings. Vulnerable populations face disproportionate disease exposure burdens through this housing crowding channel.

As a result, we conclude that important inequities in occupations and housing led to different populations facing different risk exposure during the crisis. Black, Hispanic, and low-income workers are more likely to be employed in essential and frontline occupations and hence exhibit mobility patterns which put them at greater risk of infection in the initial phase of the pandemic. As a consequence, disparities in infections reflect inequalities in access to both jobs and housing.

Our results present a stark contrast to some existing work on the COVID-19 pandemic which highlights the role of static factors such as population density or public transportation. We find that population density per se is not the dominant factor in explaining the cross-section in infections seen throughout this crisis: the densest borough, Manhattan, was less affected. Instead, we find that underlying inequalities in access to jobs and housing explain the racial disparities in outcomes. Crowding at home and exposure at work, rather than density, best explains the inequities of disease burden through the pandemic.

References

Almagro, Milena and Angelo Orane-Hutchinson, "JUE insight: The determinants of the differential exposure to COVID-19 in New York city and their evolution over time," *Journal of Urban Economics*, 2020, p. 103293.

Athey, Susan, Billy Ferguson, Matthew Gentzkow, and Tobias Schmidt, "Experienced Segregation," Working Paper 2019.

Borjas, George J., "Demographic determinants of testing incidence and Covid-19 infections in New York City neighbourhoods," *Covid Economics, Vetted and Real-Time Papers*, 2020.

Carozzi, Felipe, Sandro Provenzano, and Sefi Roth, "Urban Density and Covid-19," *CEP Discussion Paper No 1711*, 2020.

Chen, M. Keith and Ryne Rohla, "The effect of partisanship and political advertising on close family ties," *Science*, 2018, *360* (6392), 1020–1024.

_ , Judith A. Chevalier, and Elisa F. Long, "Nursing Home Staff Networks and COVID-19," Working Paper 27608, National Bureau of Economic Research July 2020.

_ , Kareem Haggag, Devin G. Pope, and Ryne Rohla, "Racial Disparities in Voting Wait Times: Evidence from Smartphone Data," Working Paper 26487, National Bureau of Economic Research November 2019.

Chiou, Lesley and Catherine Tucker, "Social Distancing, Internet Access and Inequality," Working Paper 26982, National Bureau of Economic Research April 2020.

Couture, Victor, Jonathan Dingel, Allison Green, Jessie Handbury, and Kevin Williams, "Measuring movement and social contact with smartphone data: A real-time application to COVID-19," 2020.

Duranton, Gilles and Diego Puga, "The economics of urban density," Technical Report 2020.

García-Lopez, Miquel-Angel and Diego Puga, "Cities reshaped: Mobility patterns and COVID-19's impact on cities," 2020.

Glaeser, Edward L., Caitlin S. Gorback, and Stephen J. Redding, "How Much does COVID-19 Increase with Mobility? Evidence from New York and Four Other U.S. Cities," Working Paper 27519, National Bureau of Economic Research July 2020.

Goolsbee, Austan and Chad Syverson, "Fear, Lockdown, and Diversion: Comparing Drivers of Pandemic Economic Decline 2020," Working Paper 27432, National Bureau of Economic Research June 2020.

Harris, Jeffrey E., "The Subways Seeded the Massive Coronavirus Epidemic in New York City," Working Paper 27021, National Bureau of Economic Research April 2020.

Jing, Qin-Long, Ming-Jin Liu, Zhou-Bin Zhang, Li-Qun Fang, Jun Yuan, An-Ran Zhang, Natalie E Dean, Lei Luo, Meng-Meng Ma, Ira Longini et al., "Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: a retrospective cohort study," *The Lancet Infectious Diseases*, 2020, *20* (10), 1141–1150.

Karaca-Mandic, Pinar, Archelle Georgiou, and Soumya Sen, "Assessment of COVID-19 Hospitalizations by Race/Ethnicity in 12 States," *JAMA Internal Medicine*, 08 2020.

McCormack, Grace, Christopher Avery, Ariella Kahn-Lang Spitzer, and Amitabh Chandra, "Economic Vulnerability of Households With Essential Workers," *JAMA*, 07 2020, *324* (4), 388–390.

McLaren, John, "Racial Disparity in COVID-19 Deaths: Seeking Economic Roots with Census data," Working Paper 27407, National Bureau of Economic Research June 2020.

Oster, Emily, "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 2019, *37* (2), 187–204.

Price-Haywood, Eboni G., Jeffrey Burton, Daniel Fort, and Leonardo Seoane, "Hospitalization and mortality among black patients and white patients with Covid-19," *New England Journal of Medicine*, 2020.

Rentsch, Christopher T., Farah Kidwai-Khan, Janet P. Tate, Lesley S. Park, Joseph T. King Jr., Melissa Skanderson, Ronald G. Hauser, Anna Schultze, Christopher I. Jarvis, Mark Holodniy et al., "Covid-19 by Race and Ethnicity: A National Cohort Study of 6 Million United States Veterans," *medRxiv*, 2020.

Sá, Filipa, "Socioeconomic Determinants of Covid-19 Infections and Mortality: Evidence from England and Wales," *Covid Economics, Vetted and Real-Time Papers*, 2020.

Schmitt-Grohé, Stephanie, Ken Teoh, and Martín Uribe, "COVID-19: Testing inequality in New York City," Working Paper 27019, National Bureau of Economic Research 2020.

Trivedi, Amal N., Alan M. Zaslavsky, Eric C. Schneider, and John Z. Ayanian, "Trends in the quality of care and racial disparities in Medicare managed care," *New England Journal of Medicine*, 2005, *353* (7), 692–700.

Wong, Mitchell D., Martin F. Shapiro, W. John Boscardin, and Susan L. Ettner, "Contribution of major diseases to disparities in mortality," *New England Journal of Medicine*, 2002, *347* (20), 1585–1592.

Online Appendix

A Data Appendix

A.1 Data filters

To isolate the mobility behavior of New York City residents, we employ multiple screens to filter out commuters, visitors, and those who leave the city either temporarily or permanently.

First we separate those who spend the night in New York City from those who spend the night elsewhere. We select from the anonymous users only those who have the majority of their pings between 6pm and 8am (night hours) in New York City (as opposed to any non-New York City county in the US) on at least three different days in a specific month. We then enforce a minimum required data density and keep only those with at least three pings on at least five nights in the data in New York City, with the same requirements during work hours.

We repeat this process each month from February to June and exclude those who have been identified as residents in previous months. We use only one month of data at a time to identify residents' home tracts. We then analyze their data in the months after the month that was used to identify their home locations. This gives us a sample population of 647,068 unique users for our base analysis.[13] We also exclude individuals for whom we cannot identify a home census tract (HCT). The resulting data set has 483,698 unique individuals and allows us to measure the mobility responses among NYC residents.

We further restrict our sample to March and April for our main analysis, which starts with the first month of the stay-at-home order in NYC. For these two months, we are able to identify 294,440 unique residents. We keep only the individuals observed for at least ten days during March and April to reduce the noise in individual's modal HCT each date. The noise in an individual's modal HCT comes from an individual staying in different census tracts on different nights. We want to reduce this noise because home geography is how we link demographic data to individuals. This restriction leaves us with 133,891 unique residents.

[13] We find that our estimated mobile phone population correlates with census population at 0.89, suggesting representative sample coverage.

A.2 Spatial merge between GPS data and buildings

We connect the ping data with the geographic data for all building footprints in NYC. Building footprints are created by Microsoft from satellite images.[14] We spatially join these building shapes to land use data from the NYC Department of Planning at the lot level to get the number of residential units in each lot.[15] Multiple lots correspond to each Microsoft building. We then aggregate the lots in each building to arrive at the total number of residential units and residential square footage for each building. We identify 294,971 residential buildings, 32,090 of which constitute a modal building for some mobile phone user in our data.

A.3 Details on Data Set Creation For Building and Individual Level Analysis

A.3.1 Individual panel

Mobility measures

To construct *mobility measures* we follow these steps:

1. Identify residents using filters described above.
2. Keep residents who were identified as residents during February.
3. Identify each resident's modal night building (6:00pm - 8:00am) each date and modal HCT each date.
4. Construct *hours outside HCT* by counting the number of hours a resident spends entirely outside their HCT between 8am and 10pm.
5. This gives a panel of individuals with mobility measures, modal building, and HCT.

Hospitalizations

To identify *hospitalizations* we follow these steps:

1. Using our panel of individuals, identify all pings in hospitals by our residents.
2. Identify the earliest date in a hospital for each resident.

[14] This dataset can be found at: https://github.com/microsoft/USBuildingFootprints.

[15] See: https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page.

3. Keep those who had their first ping in a hospital between 3/22 and 4/22 inclusive.
4. Keep 66 NYC hospital buildings including: Lenox Hill, Montefiore Medical Center, Elmhurst Hospital Center, Mt. Sinai, Weiler Hospital, Brookdale University Hospital Medical Center, Bellevue Hospital, New York Presbyterian hospital system, Staten Island University Hospital, all other Northwell hospitals, the Javits Center field hospital, Flushing Medical Center, NYU Langone, Kingsbrook Jewish Medical Center, Jamaica Hospital Center, Kings County Hospital Center, Wyckoff Heights Medical Center, Coney Island Hospital, Saint Johns Episcopal Hospital, Maimonides Medical Center, Jacobi Medical Center, Richmond University Medical Center, Memorial Sloan Kettering Cancer Center, Queens Hospital Center, Lincoln Medical Center, Woodhull Medical Center, St. Barnabas Hospital, North Central Bronx Hospital, New York Community Hospital, Metropolitan, Harlem Hospital Center, Hospital For Special Surgery, Brooklyn Hospital Center.
5. Exclude residents who live in a tract that has a hospital in it.
6. Make a flag *hospitalized* for residents who ping in the hospital 1 day after their first ping in that hospital, or do not appear in the data the day after their first ping in the hospital.

A.3.2 Housing Crowding

To create our panel of buildings with the variable *housing crowding* we follow these steps:

1. For each building and date, count the unique residents who had that building as their modal night building.
2. Combine with land use features as described above.
3. We divide the number of unique residents for that building x date by the number of residential units in the building to get our *people per unit* measure of housing crowding.

A.3.3 Combining data sets

- For each hospitalized individual, calculate their modal home building pre-hospitalization.
- Merge the individual level hospitalization and building level data sets on modal home building pre-hospitalization for each resident.
- Calculate *hospitalizations per housing unit*: divide hospitalizations by number of units in the building.
- Assign 0's to hospitalizations and number of residents for those buildings and dates for which we cannot identify residents. By construction, the individual level analysis contains only buildings for which we were able to identify a resident.
- This gives us a panel at the individual and date level. It includes information about individual's daily modal building including residential area, number of units, number of residents and hospitalizations per day.

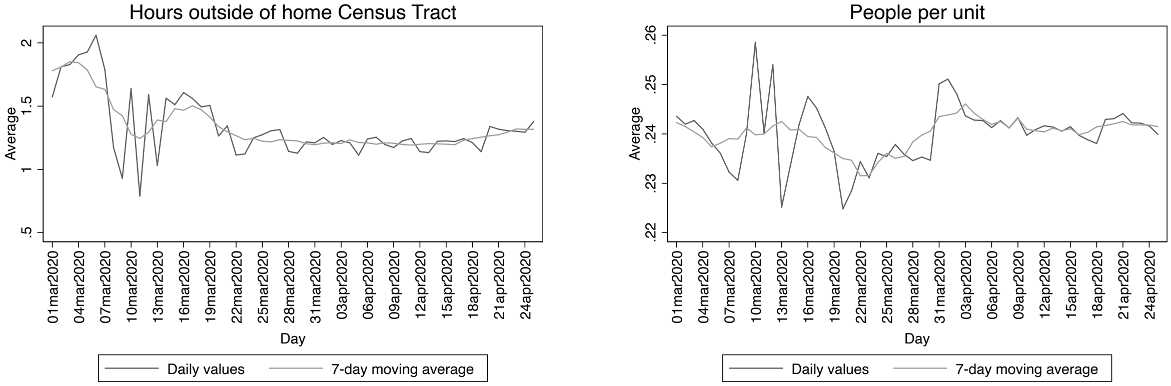
A.4 Summary Statistics

Table A1: Summary Statistics

Variable	**Mean**	**Std. Dev.**	**p10**	**Median**	**p90**
Panel A: Housing Crowding for Building with Pings (Average March 22 – April 22)					
People per Unit	0.470	0.454	0.040	0.364	1.000
Residential Units per Building	21.900	94.817	1.000	3.000	47.000
Residential Area (sq. ft.)	24,230	95,273	1271	3200	46,680
Panel B: Individual Level Mobility (Average March 22 – April 22)					
Number of Hours Entirely Outside of Home Tract	1.303	1.938	0.000	0.544	3.526
Panel C: Other Variables and Local Controls					
Share of Positive Tests	0.563	0.085	0.438	0.583	0.645
Tests per Capita	0.018	0.006	0.012	0.017	0.026
Median Income (in $1000s)	68.604	31.878	34.122	62.202	115.084
Share $\geq 20, \leq 40$	0.323	0.084	0.246	0.308	0.433
Share $\geq 40, \leq 60$	0.258	0.033	0.220	0.261	0.296
Share ≥ 60	0.200	0.079	0.132	0.190	0.276
Share Male	0.477	0.029	0.446	0.479	0.508
Household Size	2.683	0.537	1.930	2.750	3.300
% Black	0.200	0.240	0.010	0.076	0.600
% Hispanic	0.263	0.195	0.078	0.189	0.634
% Asian	0.144	0.139	0.017	0.094	0.335
Density (in 1000s of people per unit)	43.380	31.045	10.784	36.639	90.075
% Public Transport	0.532	0.150	0.312	0.543	0.712
Commuting Time (in mins)	40.647	7.054	27.200	42.100	48.100
% Uninsured	0.089	0.043	0.042	0.084	0.143
% Essential: Professional	0.126	0.089	0.046	0.092	0.285
% Essential: Service	0.065	0.033	0.035	0.060	0.107
% Essential: Technical	0.014	0.009	0.004	0.013	0.022
Non-Flexible Occupations:					
- % Health Practitioners	0.029	0.018	0.009	0.026	0.050
- % Other Health	0.038	0.024	0.010	0.035	0.073
- % Firefighting	0.012	0.009	0.003	0.012	0.023
- % Law Enforcement	0.007	0.007	0.001	0.006	0.014
- % Ind. and Construction	0.054	0.027	0.014	0.056	0.090
- % Transportation	0.029	0.016	0.004	0.032	0.048
- % Non Ess.: Professional	0.279	0.075	0.195	0.271	0.359
- % Science Fields	0.006	0.007	0.001	0.004	0.015
- % Law and Related	0.018	0.026	0.003	0.008	0.049
- % Non Ess.: Service	0.032	0.013	0.016	0.032	0.047

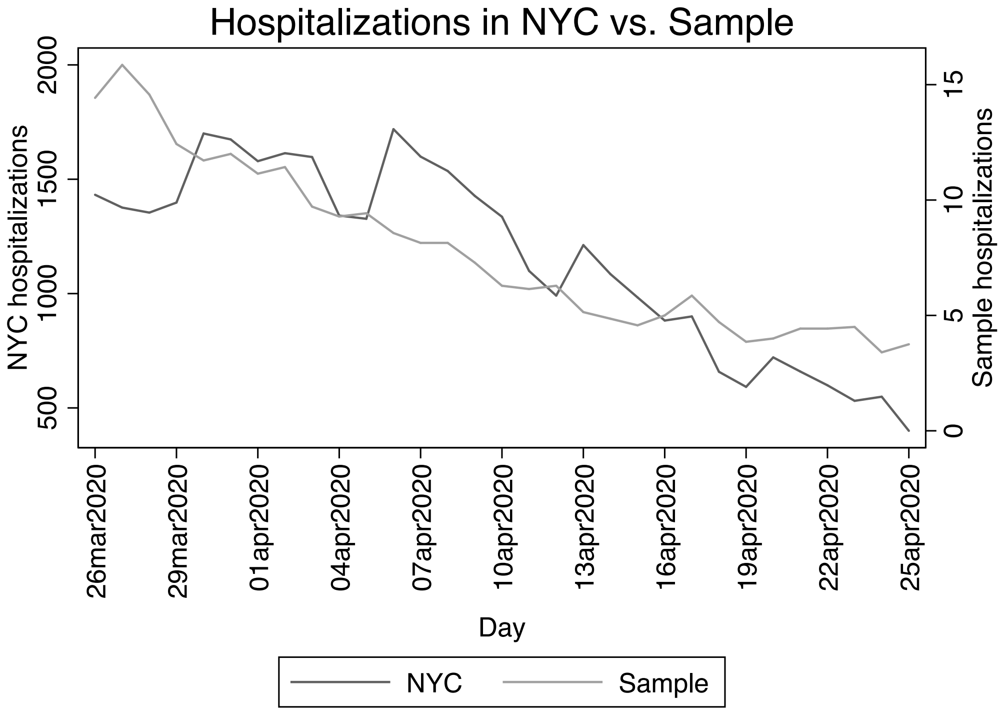
A.5 Descriptive Analysis

Figure A1: Time series of mobility patterns and housing crowding



Note: These graphs present the time series of our risk measures: hours outside HCT and number of people per housing unit.

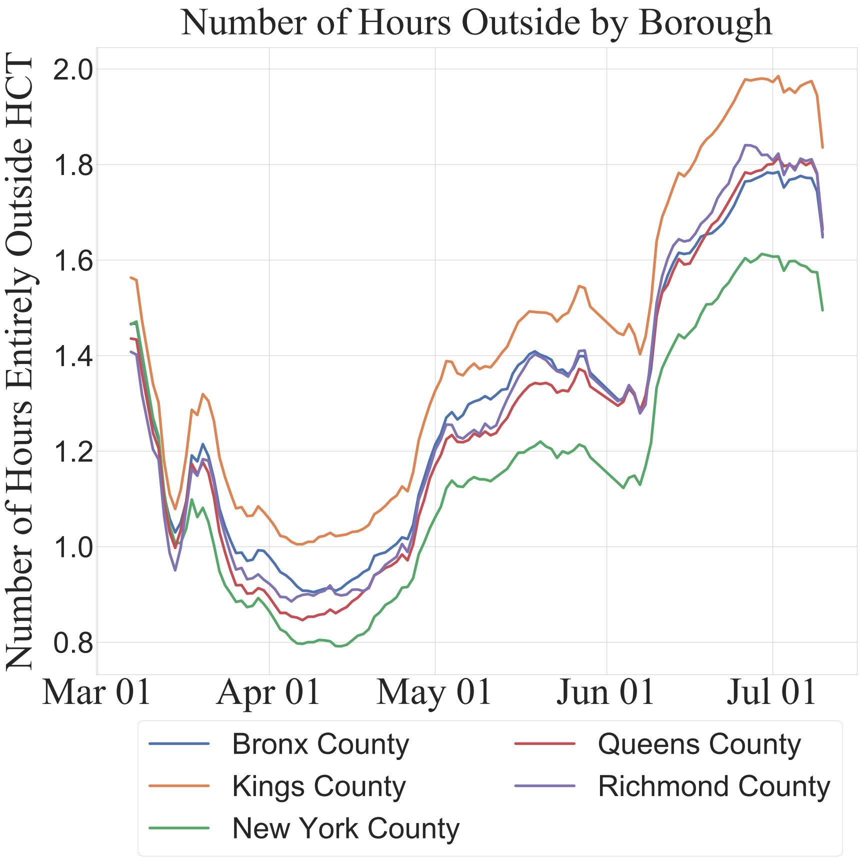
Figure A2: Comparing Hospitalizations in Mobile Phone Sample



Note: These graphs plots the time series for our individual-level mobility-derived measure of hospitalization in comparison with the official figures provided by the DOH. The correlation between the two is 0.76

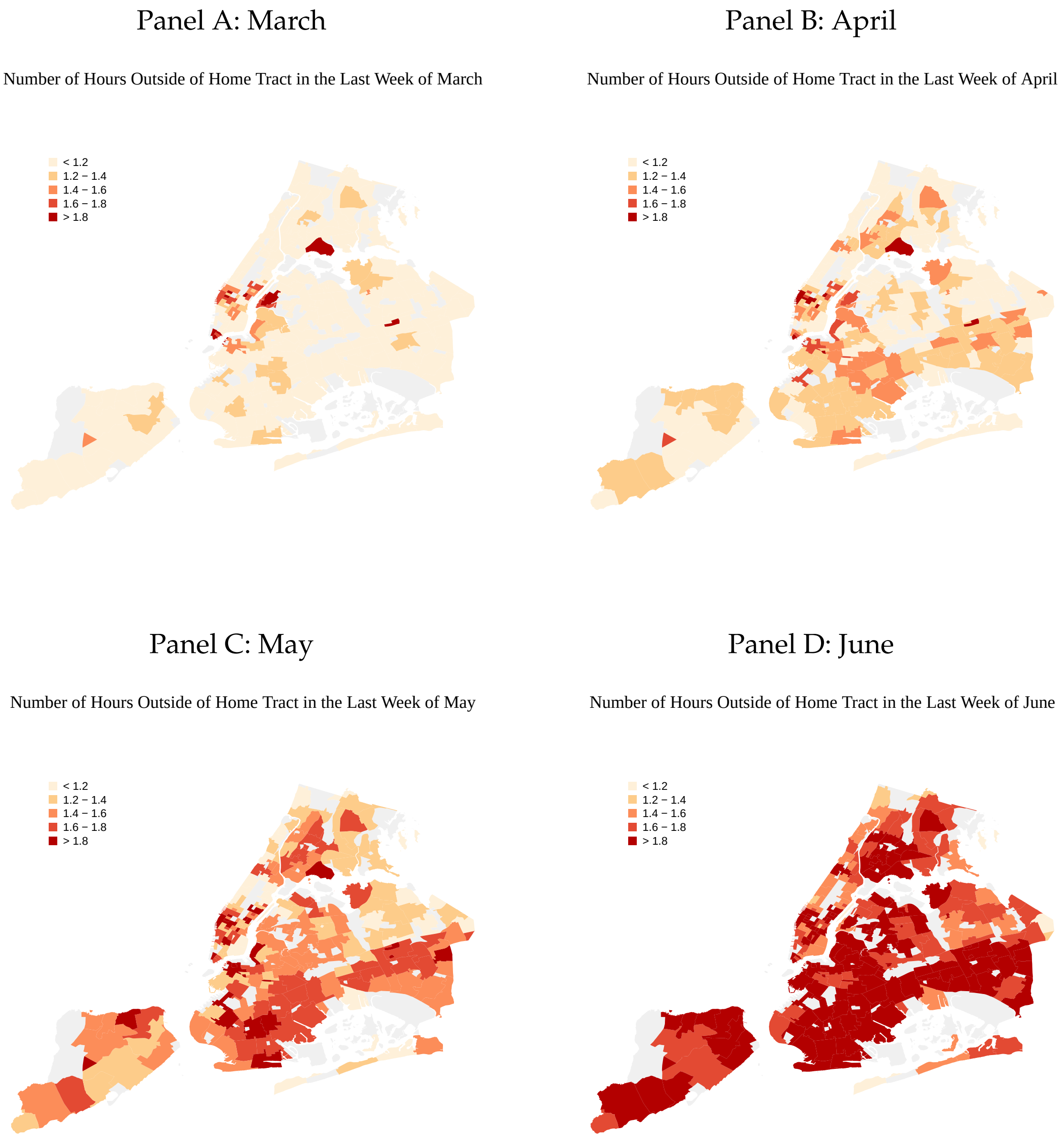
B Time Series and Cross-Section of Mobility measures

Figure B1: Time Series of Mobility Measures



Note: This graph presents the time series of our mobility measure: the average number of hours a user spends entirely outside of their HCT. This is a 7 day moving average to account for weekly variation. This also excludes ZIP codes with average incomes over $200,000. The ZIP codes are aggregated from census tracts. We exclude census tracts that contain bridges, tunnels, or highways to decrease noise.

Figure B2: Cross-Section of Hours Outside HCT Mobility



Panel A: March

Panel B: April

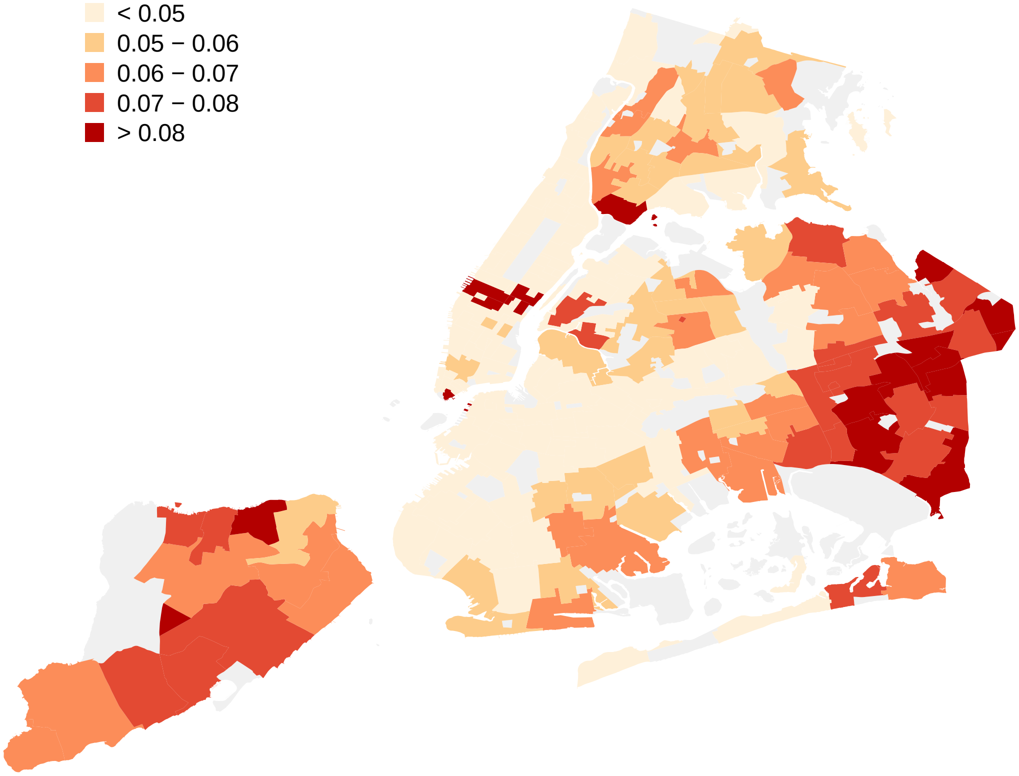
Panel C: May

Panel D: June

Number of Hours Outside of Home Tract in the Last Week of March
Number of Hours Outside of Home Tract in the Last Week of April
Number of Hours Outside of Home Tract in the Last Week of May
Number of Hours Outside of Home Tract in the Last Week of June
< 1.2
1.2 – 1.4
1.4 – 1.6
1.6 – 1.8
> 1.8

Note: These maps present the spatial distribution at the ZIP code level of our mobility measure: the average number of hours a user spends entirely outside of their HCT. This is averaged across the last week of the month. The maps exclude census tracts with average incomes above $150,000, or population density below 1000 people per square mile.

Figure B3: Cross-Section of Housing Crowding Measure

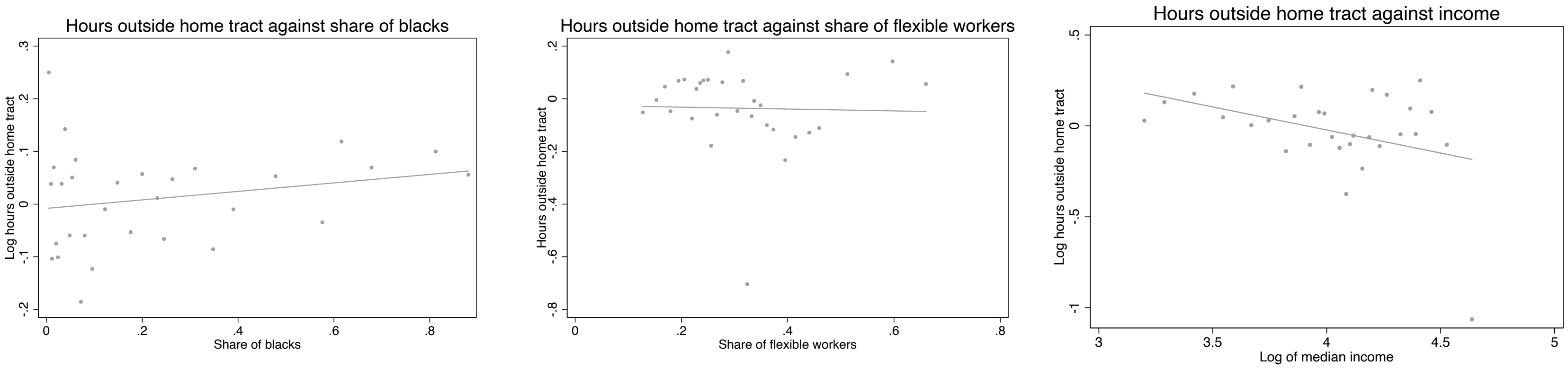


Mean People Per Unit

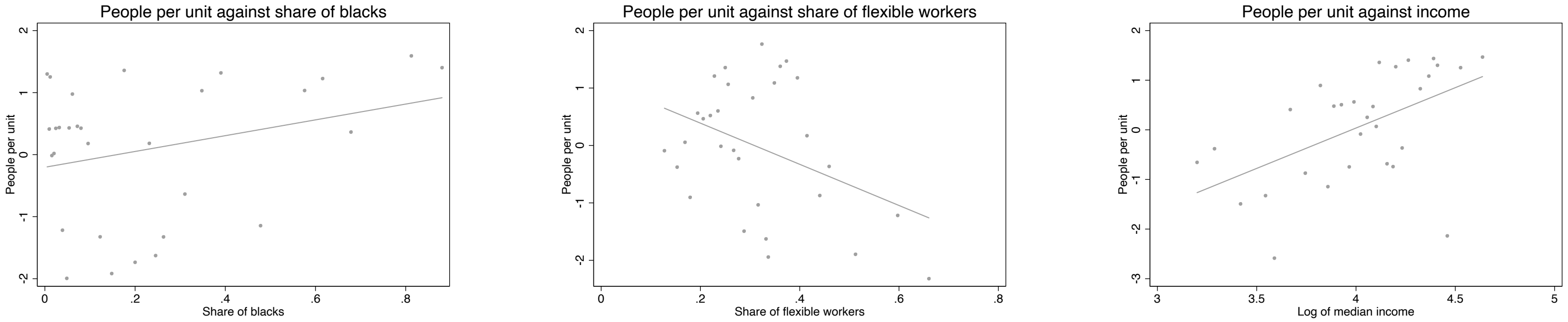
Note: These maps present the spatial distribution at the ZIP code level of our housing crowding measure: mean people per residential unit. This is from March 1st to June 30th. The map excludes census tracts with average incomes above $150,000, or population density below 1000 people per square mile.

Figure B4: Demography and Mobility Measures

Panel A: Binscatter plots of share of mobile phone pings in home tract and demographics



Panel B: Binscatter plots of crowded spaces and demographics



Note: These graphs are binscatter plots showing correlations between demographics and the residuals of a regression of risk measures on time. This last exercise is performed to control for time-varying risk behavior independent of the neighborhood demographic composition.

C ZIP Code Level Analysis with NYC Test Data

In this section we reproduce our analysis but using official measures of tests performed and positive results across ZIP codes. Our source of incidence rates of COVID-19 and number of tests performed is the NYC Department of Health (DOH) data release.[16] The DOH releases (almost) daily data on the cumulative count of COVID-19 cases and the total number of residents who have been tested, divided by the ZIP code of residence. We have collected data covering the months of April and May.[17] In our analysis, we drop the first week of April due to these missing dates, and also because the first few days in our sample appear very noisy.

Table C1 shows our regression results, using similar specifications as in our main analysis but with the outcome variable defined as the daily share of positive tests per ZIP code. Although statistical significance varies according to specification, we find positive correlation between both of our risk measures and the share of positive tests. For example, in the specification with only demographic variables, a 10% increase in the number of hours out of home tract is associated with a 1.2% increase in the share of positive tests.

[16] See: https://github.com/nychealth/coronavirus-data.

[17] Unfortunately April 2nd and April 6th are missing from our sample as these data have never been made publicly available.

Table C1: Neighborhood Associations of Positive Tests

Dependent Variable:	Daily Share of Positive Tests											
	(1) Race & Income		(2) Occup.		(3) Mobility		(4) Mobility, Race, & Income		(5) Mobility & Occup.		(6) Mobility, Race, Dem, & Occup.	
Log hours outside HCT					0.068***	(0.013)	0.055**	(0.019)	0.003	(0.013)	0.012	(0.012)
Log people per unit					0.002	(0.003)	0.008*	(0.003)	0.011	(0.008)	0.010*	(0.004)
Log Income	-0.009	(0.010)					-0.010	(0.008)			-0.083*	(0.040)
% Black	0.151***	(0.015)					0.137***	(0.019)			-0.004	(0.035)
% Hispanic	0.218***	(0.015)					0.225***	(0.016)			0.191***	(0.022)
% Asian	0.235***	(0.013)					0.237***	(0.014)			-0.046*	(0.020)
% Flexible occupations			0.059	(0.084)					0.099	(0.109)	0.320	(0.218)
% Health practitioners			-0.985***	(0.263)					-1.226**	(0.395)	-0.625*	(0.318)
% Other health			1.106***	(0.132)					1.121***	(0.134)	1.392***	(0.406)
% Firefighting			0.533	(0.368)					0.713	(0.430)	2.154	(1.140)
% Law enforcement			-3.732***	(0.352)					-4.273***	(0.609)	-2.481***	(0.541)
% Essential - Service			0.569***	(0.135)					0.781***	(0.102)	0.093	(0.124)
% Non ess. - Service			-0.102	(0.280)					-0.233	(0.229)	0.033	(0.218)
% Ind. and Construction			0.298	(0.368)					0.096	(0.245)	-0.660***	(0.173)
% Essential - Technical			0.368	(0.309)					0.227	(0.369)	-1.277*	(0.548)
% Transportation			1.573***	(0.452)					1.385***	(0.340)	0.659	(0.368)
Share $\geq 20, \leq 40$											0.355	(0.204)
Share $\geq 40, \leq 60$											0.611	(0.330)
Share ≥ 60											0.676**	(0.206)
Share Male											0.922***	(0.098)
Log Household Size											0.254**	(0.091)
% Public Transport											-0.097*	(0.048)
Log Commute Time											-0.013	(0.026)
% Uninsured											0.370***	(0.079)
Bronx											-0.013	(0.022)
Brooklyn											0.062**	(0.020)
Queens											0.038*	(0.017)
Staten Island											-0.039	(0.021)
Day FE	✓		✓		✓		✓		✓		✓	
N	2660		2660		2660		2660		2660		2660	
adj. R^2	0.95		0.95		0.94		0.95		0.95		0.96	

Spatial HAC Standard errors in parentheses

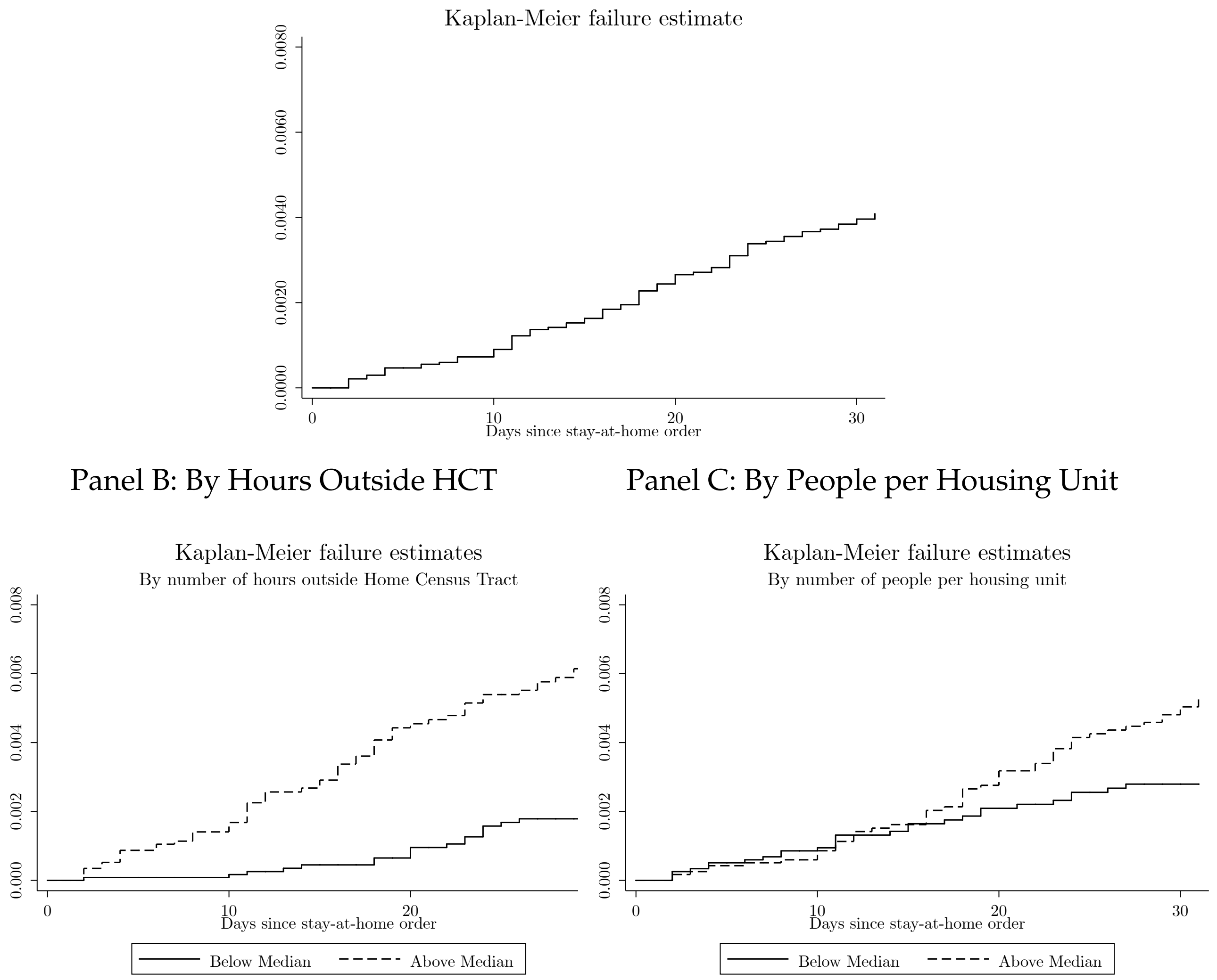
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table runs daily share of positive tests by ZIP code on different sets of covariates, where share of positive tests is defined as the fraction of all tests in the ZIP code that turn positive over the period April 8th to April 22th. Column (1) includes only basic demographics such as race and income, while column (2) includes only our mobility and housing density measures. Column (3) includes all of these covariates together. Columns (4) and (5) add additional demographic and occupational controls.

D Survival Analysis

D.1 Kaplan-Meier Graphs

Figure D1: Kaplan-Meier graphs of survival probability



Panel A: Whole Sample

Panel B: By Hours Outside HCT

Panel C: By People per Housing Unit

Note: These graphs show hazard rates of hospitalizations over days for the first month after the stay-at-home order of NYC issued on March 22nd. Panel A shows the hazard rate for all individuals. Panel B and C show hazard rates for two groups defined as above and below the median for our risk measures, hours outside HCT and people per housing unit.

Table D1: Cox Regression on Risk Measures and Demographics

Dependent Variable:	Hazard rate of being hospitalized											
	(1)		(2)		(3)		(4)		(5)		(6)	
	Race & Income		Risk		Risk, Race & Income		Risk, Race, Dem, & Occup.		Comm. District FE		Comm. District Time Trend	
Log hours outside HCT			0.369***	(0.051)	0.360***	(0.051)	0.357***	(0.051)	0.370***	(0.051)	0.369***	(0.051)
Log people per unit			0.156**	(0.064)	0.154*	(0.080)	0.204*	(0.106)	0.223**	(0.089)	0.221**	(0.089)
Log Income	0.300	(0.323)			0.158	(0.346)	0.430	(0.548)				
% Black	0.009*	(0.005)			0.007	(0.005)	-0.002	(0.008)				
% Hispanic	0.004	(0.007)			0.004	(0.007)	-0.009	(0.010)				
% Asian	-0.001	(0.009)			-0.001	(0.009)	-0.012	(0.013)				
Share $\geq 20, \leq 40$							1.162	(3.002)				
Share $\geq 40, \leq 60$							-1.100	(4.109)				
Share ≥ 60							-3.912	(3.188)				
Share Male							1.378	(3.586)				
Log Household Size							1.015	(1.033)				
% Flexible occupations							-1.460	(2.440)				
% Health practitioners							2.986	(5.370)				
% Other health							6.159*	(3.637)				
% Firefighting							-9.700	(10.130)				
% Law enforcement							-25.022**	(10.994)				
% Essential - Service							-5.327	(4.259)				
% Non ess. - Service							7.468	(6.074)				
% Ind. and Construction							-1.490	(4.299)				
% Essential - Technical							5.807	(9.354)				
% Transportation							0.943	(5.855)				
% Public Transport							0.741	(1.227)				
Log Commute Time							-0.620	(1.032)				
% Uninsured							2.443	(3.123)				
CD FE									✓			
CD Time Trend											✓	
N	23,761		23,756		23,748		23,748		23,553		23,756	

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table runs the cumulative hazard rate of being hospitalized on different sets of covariates. Column (1) includes only basic demographics such as race and income, while column (2) includes only our risk measures. Column (3) includes all of these covariates together. Columns (4) expands by adding more demographics and occupational shares. Column (5) includes Community District fixed effects and Column (6) adds Community District-specific linear trends.