Journal of Exposure Science & Environmental Epidemiology (2021) 31:525–537
https://doi.org/10.1038/s41370-021-00329-1

ARTICLE

# Neighborhood characteristics associated with COVID-19 burden—the modifying effect of age

Xueying Zhang[1] · Norah Smith[2] · Emily Spear[1] · Annemarie Stroustrup[1,3]

Received: 14 August 2020 / Revised: 1 April 2021 / Accepted: 7 April 2021 / Published online: 4 May 2021

## Abstract

**Background** Neighborhood characteristics have been linked to community incidence of COVID-19, but the modifying effect of age has not been examined.

**Objective** We adapted a neighborhood-wide analysis study (NWAS) design to systematically examine associations between neighborhood characteristics and COVID-19 incidence among different age groups.

**Methods** The number of daily cumulative cases of COVID-19 by zip code area in Illinois has been made publicly available by the Illinois Department of Public Health. The number of COVID-19 cases was reported for eight age groups (under 20, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80+). We reviewed this data published from May 23 through June 17, 2020 with complete data for all eight age groups and linked the data to neighborhood characteristics measured by the American Community Survey (ACS). Geographic age-specific cumulative incidence (cases per 1000 people) of COVID-19 was calculated by dividing the number of daily cumulative cases by the population of the same age group at each zip code area. The association between individual characteristics and COVID-19 incidence was examined using Poisson regression models.

**Results** At the zip code level, neighborhood socioeconomic status was a more important risk factor of COVID-19 incidence in children and working-age adults than in seniors. Social demographics and housing conditions were important risk factors of COVID-19 incidence in older age groups. We additionally observed significant associations between transportation-related variables and COVID-19 incidences in multiple age groups.

**Significance** We concluded that age modified the association between neighborhood characteristics and COVID-19 incidence.

**Keywords** Children's health · Environmental justice · Geospatial analyses

## Introduction

The Coronavirus Disease 2019 (COVID-19) pandemic has created unprecedented disruptions of public health systems globally, affecting people of all ages. Epidemiologic studies from multiple countries have reported that people over 60 years old are at substantially higher risk of severe COVID-19 disease than other age groups [1–3]. Notably, although children are generally quite vulnerable to viral infections, symptomatic COVID-19 had unexpectedly low prevalence in children and teenagers under 20 years old [1–4] during the early stages of the pandemic. It was at first unclear if children were more likely to experience asymptomatic infections, or simply have a lower infection rate. Population-wide screening conducted in Iceland found no positive infections with SARS-CoV-2, the virus that causes COVID-19, among children under ten years old, whereas people older than 10 years had a 0.8% positive infection rate [5]. In the United States, although population screening was not widely implemented, the low asymptomatic rate (1.5%) in a representative pediatric population suggested

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41370-021-00329-1.

✉ Xueying Zhang
xueying.zhang@mssm.edu

[1] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[2] The Bronx High School of Science, Bronx, NY, USA

[3] Division of Neonatology, Cohen Children's Medical Center, Northwell Health, New Hyde Park, NY, USA

evidence of low rates of infection with SARS-CoV-2 among American children [6]. Though the SARS-CoV-2 infection rate in this age group increased with school re-opening, the pediatric COVID-19 incidence (10%) remains low compared to the proportion of this age group in the general population (20%) [7]. Reasons for disparities in SARS-CoV-2 infection among age groups are unclear. While differing immunological and physiological susceptibilities exist [8, 9], children and adults also experience different levels of environmental exposure to SARS-CoV2 that raise the risk of COVID-19 in adult populations due to differences in daily activities and behaviors [10], indicating that neighborhood environment might contribute to differences in COVID-19 incidence among children and adults.

In addition to the differing infection rates of COVID-19 across age groups, the unequal distribution of COVID-19 cases in neighborhoods with different socioeconomic status (SES) has also been noted, especially in U.S. metropolitan areas [11–16]. The neighborhood characteristics examined in those studies were identified using the American Community Survey (ACS), a tool widely used in epidemiology to investigate disparities among adverse health outcomes [18]. For each study, researchers selected ACS variables as representatives of area-level SES, then linked the selected variables to COVID-19 incidence through geographic units, commonly zip code areas. Results of these studies suggested that neighborhoods marked by lower SES had relatively higher COVID-19 incidence. One multi-city study created an index of neighborhood disadvantage based on six neighborhood characteristics from the ACS [14]. Epidemiological evidence of the link between exposure to lower SES neighborhood characteristics and multiple adverse health outcomes, including COVID-19, has been inconsistent, since different studies select different variables to represent neighborhood SES. Moreover, given that COVID-19 is a new disease, selecting neighborhood characteristics to study using traditional epidemiological approaches risks missing other neighborhood risks of COVID-19 disease which may not have been recognized a priori. We considered that transmission of the SARS-CoV-2 virus and associated COVID-19 disease could be related to neighborhood characteristics not previously investigated, and sought a novel approach to identify neighborhood characteristics important for further study.

The recent development of exposome-based approaches permits agnostic exploration of neighborhood-wide risk factors to diseases [19, 20]. The neighborhood-wide association study (NWAS) framework, inspired by the genome-wide association study (GWAS) approach, is one such approach [21]. While GWAS examines associations between large panels of genes and diseases, NWAS examines associations between all available neighborhood characteristics and disease. For example, two studies applied the NWAS approach to study neighborhood risk factors associated with prostate cancer and poor physical activity, respectively [19, 20]. Lynch et al. [19] developed a multi-phase approach that identified 17 new ACS variables that had not been previously linked to prostate cancer. Mooney et al. [20] used both linear regression and machine learning to identify specific characteristics of extreme poverty associated with a reduced amount of physical activity among residents in the same census tract. The NWAS approaches in Lynch et al. [19] and Mooney et al. [20] both consisted of two major steps. The first step was a pairwise analysis to examine the associations between individual ACS variables and the study outcome. The second step used multivariable analysis to examine the selected variables' adjusted effect on the outcome of interest.

We adapted an NWAS approach to systematically examine associations between neighborhood characteristics and COVID-19 incidence among different age groups. We used the Illinois COVID-19 data and 392 ACS measures to identify neighborhood characteristics associated with COVID-19 in eight age groups ranging from children to seniors. We hypothesized that neighborhood characteristics highly associated with the incidence of COVID-19 vary by age groups. We also hypothesized that in addition to variables widely regarded as representatives of SES, we could identify other neighborhood characteristics related to age-specific COVID-19 incidence.

## Methods

### COVID-19 data

The cumulative number of COVID-19 laboratory-confirmed cases by zip code areas in Illinois was published and updated daily by the Illinois Department of Public Health (IDPH). The compiled dataset was made available by the Chicago Reporter [22]. We retrospectively reviewed the daily cumulative cases for 26 days, May 23 through June 17, 2020. The published COVID-19 data were reported by eight age groups (under 20, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80+). Due to patient privacy concerns, the age groups with less than six cumulative cases were coded as zero cases by IDPH in the downloaded data. Our analysis was conducted based on zip code areas with available data for all of the eight age groups. To maximize the number of zip code areas included in this study, we followed a multi-step approach. First, for each zip code area, we estimated the sum of confirmed cases in all eight age groups. Second, we calculated the difference between the total numbers of confirmed cases reported by IDPH and the estimated sum of all eight age groups. Third, we selected zip code areas in which the reported total confirmed cases and the estimated sum of age-specific cases were same. For the remaining zip code areas, if the difference was less than six and only one age group's case number was 0 (indicating missing) then the zip code areas with case numbers coded as zero were filled with the difference. These zip code areas were also included in our analysis. We focused on the same group of zip code areas with repeated estimates of daily cumulative cases for each of the 26 study days in the analysis. The age-specific incidence of COVID-19 at individual zip code areas were calculated using the daily cumulative number of cases divided by the population at the same age. Population data was downloaded from the latest ACS five-year data (2014–2018 version). Calculated cumulative incidences were scaled to cases/1000 population in the subsequent analyses. The age-specific cumulative incidence (cases/1000 population) was the outcomes of interest in our analyses.

## Neighborhood characteristics

Neighborhood characteristics were examined using the latest ACS five-year data (2014–2018 version). We used all ACS variables included in ACS profiles tables, which consist of more than four hundred ACS variables in four categories (Social, Economic, Housing, and Demographic) to give a broad statistical depiction of a geographic unit ranging from nation to census block group [23]. We downloaded the ACS profile table in the geographic unit of ZIP Code Tabulation Areas (ZCTAs) via the R 3.6.0 [24] package tidycensus [25]. ZCTA has been extensively used as a geographic unit for measuring ACS estimates in epidemiologic studies because it overlaps with the postal zip code [5]. The downloaded data consisted of number estimates and percentage estimates. Only percentage estimates were kept, scaled by one standard deviation (SD), and then linked to age-specific COVID-19 incidences by zip code areas. Those scaled percentage estimates are referred to as "ACS variables" throughout this report.

## Neighborhood-wide association study (NWAS)

### Pairwise analysis

Pairwise analysis was conducted to measure associations between individual ACS variables and age-specific COVID-19 cumulative incidence in a multi-step process. First, Pearson correlation analysis was used to examine bivariate relationships between (a) any two ACS variables and (b) individual ACS variables and the average age-specific COVID-19 cumulative incidence for the 26-day study period. Second, the association between individual ACS variables and daily age-specific COVID-19 cumulative incidence was examined using Poisson mixed-effects regression modeling. Poisson regression was used because disease incidence generally follow a Poisson distribution [26]. We included a random intercept by zip code area to account for repeated daily measurements in the 26-day study period within a zip code area. To adjust the varied association between social variables and population health between urban and rural areas, Poisson models were adjusted with a binary variable indicating whether the zip code area was identified as "urban" by the Census Bureau [17]. The Census Bureau defined urban areas as cities with a population of more than 50,000 residents and adjacent satellite cities/towns with more than 2500 residents [17]. In these analyses, we considered a zip code area to be an urban zip code area if fully covered by a Census-defined urban area. This cutoff categorized half of our study zip code areas as urban zip codes. We conducted Poisson regression for every pair of ACS variables and age-specific COVID-19 cumulative incidence. A cutoff of 2-tailed $P = 0.05$ after Bonferroni correction for multiple comparisons was used to indicate significance.

We illustrated pairwise analyses results using several visualization strategies that have been widely used to depict results after multiple comparisons. Since we were interested in identifying patterns of ACS variables associated with age-specific COVID-19 cumulative incidence, we constructed network maps of ACS variables based on both correlations with other ACS variables and on association with COVID-19 incidence in the Poisson regression analyses. The developed approach included two steps. First, we built a full network using the correlation matrix of all ACS variables. Pairs of ACS variables included in the full network had an absolute value of Pearson correlation coefficient >0.5 and Pearson correlation $P < 0.05$ after Bonferroni correction. Second, we built age-specific networks with the twenty ACS variables with the smallest coefficient $P$ values in Poisson regression models. For two ACS variables with total percent estimates always equal to 100 (e.g., percent of insurance population and percent of uninsured population), we only kept the one positively associated with the COVID-19 incidence (e.g., percent of the uninsured population) to avoid double-counting the same effect. The twenty ACS variables served as nodes in the networks and were linked through their bivariate relationship in the full network. Those ACS variables not included in the full network were disconnected and presented in a table in the graph. Networks were constructed using the R packages tidygraph [27] and ggraph [28]. The estimates and $P$ values of the incidence ratios representing the change in COVID-19 cumulative incidences per one standard deviation increase of ACS variable for age-specific COVID-19 cumulative incidences as well as the direction of correlation (negative or positive) between ACS variables themselves were shown using specific colors and symbol sizes in each network.

### Multivariate analysis

**Variable selection** We applied Elastic Net regression to select neighborhood characteristics most predictive of the age-specific COVID-19 cumulative incidence. Elastic Net is a variable selection approach combining ridge regression and lasso regression. Both ridge and lasso regression are widely applied variable selection approaches in predictive modeling. Lasso regression was used in one of the two previously cited NWAS studies [20]. Elastic Net's algorithm combines the advantage of lasso regression in selecting correlated variables and the advantage of ridge regression in selecting a greater number of variables [29], which has led to Elastic Net becoming a top variable selection approach in environmental health research [30]. In our study, Elastic Net regression was conducted using the R package glmnet [31]. During variable selection, the entire study dataset was randomly divided into a 60% subset for training and a 40% subset for testing. We set the proportional contribution of lasso regression versus ridge regression from 0 to 1 and then selected the model with minimal mean square error (MSE) estimated from the testing subset with ten-fold cross-validation. Variables included in the selected model moved forward to the next step of variable selection.

**Estimate the adjusted association** After Elastic net regression, a diagnostic test was performed using variance inflation factors (VIF) to detect the remaining collinearity between selected ACS variables. The variable with the highest VIF was excluded sequentially from the model until the VIFs of all remaining variables were <2.5. The remaining ACS variables were modeled with the age-specific COVID-19 cumulative incidence using Poisson regression. The coefficient of determination ($R^2$) of each Poisson regression model was calculated as an indicator of the (%) variation of COVID-19 cumulative incidence predicted by the adjusted variables. The incidence ratio (IR) and 95% confidence interval (CI) of ACS variables estimated from Poisson models were reported. The value of the IR indicated the change in COVID-19 cumulative incidence per one standard deviation increase of ACS variable.

## Results

### Descriptive analysis

Two hundred and fourteen zip code areas in the state of Illinois were included in this study. The inclusion of zip code areas were based on the completeness of data in all age groups as described in the "Method" section. Other Illinois' zip code areas were excluded in this study due to a small number of COVID-19 cases—privacy concerns require zip code areas with fewer than six confirmed cases to be coded as "0" cases in the downloaded dataset. Table 1 compares population density and urbanization (percent of urban area) for all zip code areas in Illinois and for the 214 zip code areas included in the analysis. The population density or percentage of the urban area was significantly higher in study zip code areas than in all the zip code areas in Illinois ($P < 0.0001$).

**Table 1** Population density (people/km$^2$) and urbanization (percent of urban area) of all zip code areas in Illinois and of the 214 zip code areas included in the analysis. The *p* value was calculated from the student *t* test.

| | Illinois | This study | *P* value |
|---|---|---|---|
| Number of zip code areas | 1383 | 214 | – |
| Average Population Density (People per km$^2$) | 543.16 | 2340.18 | <0.0001 |
| Median % Urban Area* in zip code areas | 0 | 100.00 | <0.0001 |

"Urban Area" is defined by the Census Bureau [17]. Cities with a population of >50,000 residents and adjacent satellite cities/towns with >2500 residents were categorized as urban area.

The daily laboratory-confirmed COVID-19 cumulative incidence (cases per 1000 people) in the 214 study zip code areas our study period (May 23 to June 17) is available in the Supplementary material, Figure S1. The cumulative incidence of COVID-19 during our study period (May 23–June 17) was reported for eight age groups (under 20, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80+). The youngest study age group (under 20) had the lowest COVID-19 cumulative incidence (2.58–3.66 cases per 1000 people, varying by zip code areas), and the most senior group (80+) had the highest COVID-19 cumulative incidence (23.62–29.93 cases per 1000 people). The difference in the cumulative incidence between the youngest group and oldest group was nearly tenfold, while the incidence of COVID-19 did not differ greatly among the other six age groups (i.e., 20–29, 30–39, 40–49, 50–59, 60–69, 70–79).

### Pairwise analysis

Figure 1 is a heat map showing the Pearson correlation coefficients between 392 ACS variables (rows) and COVID-19 incidence by eight age groups (columns). In the heat map, the color gradient indicates that several ACS variables were consistently correlated with COVID-19 incidence in age groups under 79 years old. Those variables represented educational attainment, non-English speaking status, employment as an "essential worker", health insurance status, residential occupants per room, and Hispanic ethnicity. In contrast, the color gradient for the very senior age group (80+) was distinctly different from

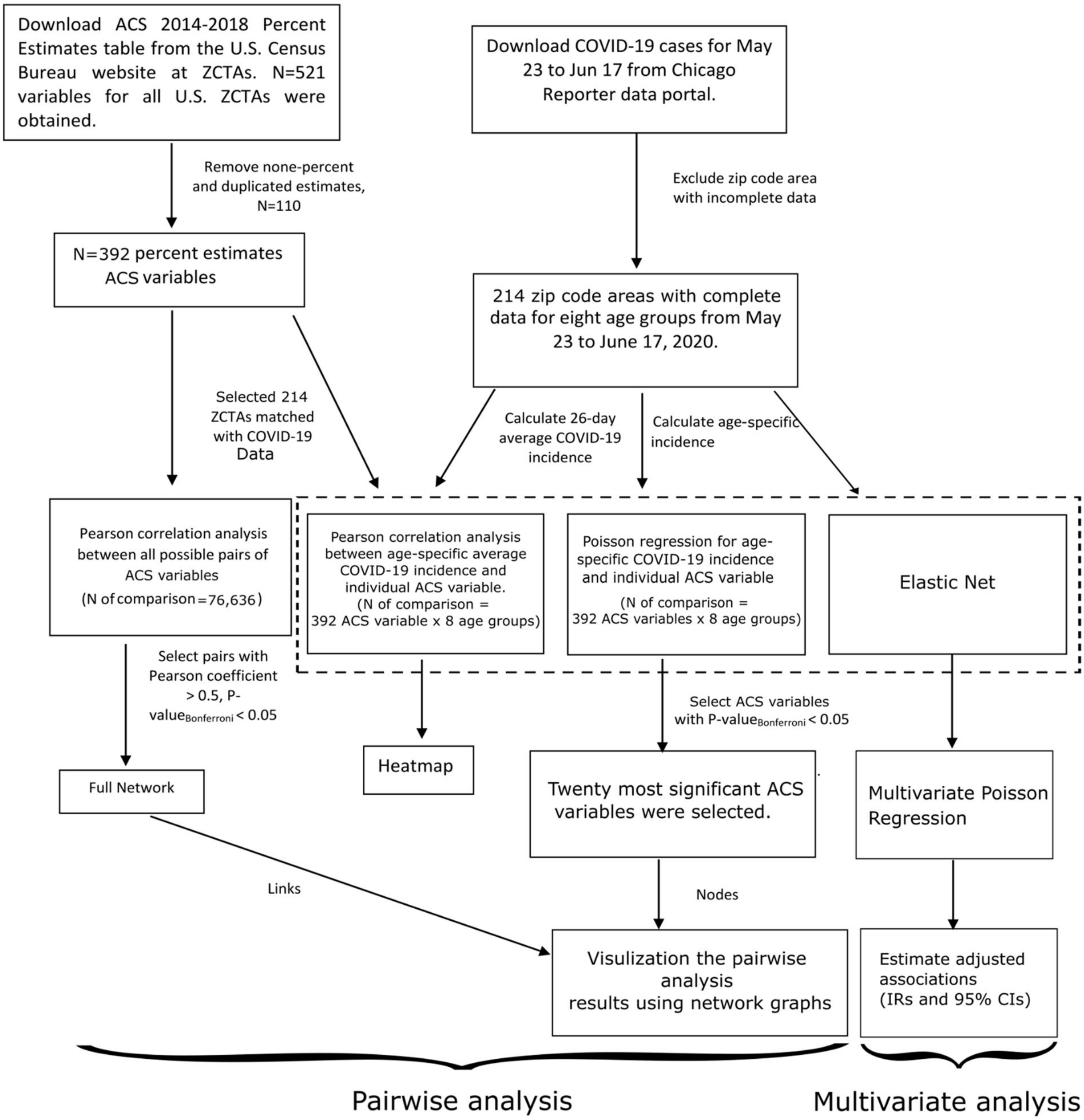
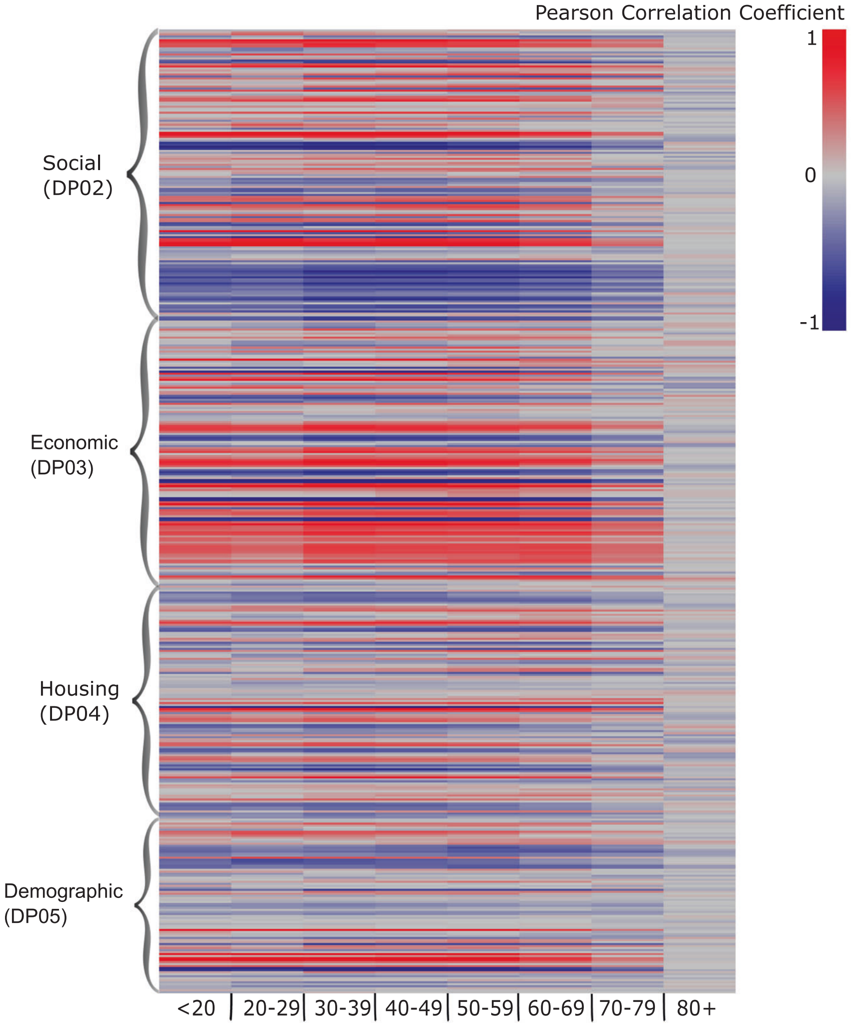

Download ACS 2014-2018 Percent Estimates table from the U.S. Census Bureau website at ZCTAs. N=521 variables for all U.S. ZCTAs were obtained.
Download COVID-19 cases for May 23 to Jun 17 from Chicago Reporter data portal.
Remove none-percent and duplicated estimates, N=110
Exclude zip code area with incomplete data
N=392 percent estimates ACS variables
214 zip code areas with complete data for eight age groups from May 23 to June 17, 2020.
Selected 214 ZCTAs matched with COVID-19 Data
Calculate 26-day average COVID-19 incidence
Calculate age-specific incidence
Pearson correlation analysis between all possible pairs of ACS variables (N of comparison=76,636)
Pearson correlation analysis between age-specific average COVID-19 incidence and individual ACS variable. (N of comparison = 392 ACS variable x 8 age groups)
Poisson regression for age-specific COVID-19 incidence and individual ACS variable (N of comparison = 392 ACS variables x 8 age groups)
Elastic Net
Select pairs with Pearson coefficient > 0.5, P-value$_{Bonferroni}$ < 0.05
Select ACS variables with P-value$_{Bonferroni}$ < 0.05
Full Network
Heatmap
Twenty most significant ACS variables were selected.
Multivariate Poisson Regression
Links
Nodes
Visulization the pairwise analysis results using network graphs
Estimate adjusted associations (IRs and 95% CIs)
Pairwise analysis
Multivariate analysis

**Fig. 1** Workflow of the Neighborhood-Wide Association Study (NWAS) method in this study.

the other age groups. No clear cluster of variables was observed for this age group, and the colors also demonstrate relatively small correlations between neighborhood characteristics and COVID19 incidence. The Pearson Correlation Coefficient tables used to create this heat map are available on this study's GitHub site [32].

Figure 2 shows the example network graphs of two age groups: children under 20 and seniors in 70–79 years old. The network graphs of other age groups are available in the Supplementary material, Figure S2–9. In each network graph, ACS variables are shown by their IR (change in COVID-19 cumulative incidence per one SD increase of ACS variable) estimated using Poisson mixed-effect regression models. ACS variables are linked through their bivariate relationship with other ACS variables estimated using Pearson correlation analysis. As shown in Fig. 2 and S2–9, the percent of people with European ancestries (i.e., English, Scottish, Irish, Swedish, and German) was negatively associated with COVID-19 incidence across all age groups. Besides that, networks differed between age groups. In the network for children under 20, we observed five clusters of ACS variables: health insurance, race/ethnicity,

**Fig. 2 Heatmap of Correlation Analysis Results.** The Pearson correlation coefficients between 392 neighborhood characteristics measured by the American Community Survey and age-specific COVID-19 cumulative incidence.



and educational attainment connecting the former two clusters, and a detached cluster dominated by employment as an essential worker. Among them, clusters of health insurance, race/ethnicity, and education attainment were found in the networks of age groups under 20 to 59. For two senior age groups (60–79), ACS variables related to marital status (percent of unmarried people or percent of people living with a spouse) were on the networks. The details of ACS variables significantly associated with age-specific COVID-19 incidence are available in supplementary material, Table S1.

The US census bureau categorizes all ACS variables into 43 "subjects". We identified subjects of ACS variables that were generally associated or not associated with age-specific COVID-19 incidence by the proportion of significantly associated ACS variables within the subject, results were available in supplementary material, Table S1. Variables representing computer access, fertility (number of children), and residential occupants per room were highly associated with COVID-19 incidence for age groups under 79. Overall, COVID-19 in adults between 40 and 59 was associated with a greater number of ACS variables than other age groups. Only one ACS variable (percent of people who take public transportation to work) was significantly associated with COVID-19 incidence in the very senior group (80+).

## Multivariate analysis

Table 2 shows the ACS variables identified as significant predictors of age-specific COVID-19 cumulative incidence in each step of variable selection. In the first step, Elastic net selected zero to 41 ACS variables that were highly predictive of age-specific COVID-19 cumulative incidence. The number of selected ACS variables differed by age group. Among the eight age groups, the 30–40-year-old age group had the most ACS variables ($n = 41$) selected by Elastic net regression, followed by the pediatric age group (<20) ($n = 40$). Only one ACS variable (no phone in household) was selected as a predictor of COVID-19

**Table 2** ACS variables identified as significant predictors of age-specific COVID-19 cumulative incidences in the 214 study zip code areas.

| | <20 | 20–29 | 30–39 | 40–49 | 50–59 | 60–69 | 70–79 | 80+ |
|---|---|---|---|---|---|---|---|---|
| Variable selection step 1: Elastic net results | | | | | | | | |
| Num. of ACS variables selected | 40 | 30 | 41 | 23 | 23 | 19 | 1 | 0 |
| ACS Variables | Living with grandchild for 1–2 years; <9th grade education; 9–12th grade education; Bachelors degree; High school grad +; Native, born abroad; Foreign Born, naturalized; Foreign Born, not US citizen; Born in Latin America; Age >5, English only; Non-English speaker; Limited English speaking; Speak Spanish; Only speak Spanish; German ancestry; Irish ancestry; Take carpool to work; Employed; Service workers; Production/transportation/shipment employees; Manufacturing employees; Family income, $15–25k; Family income, $25–35k; Family income, $150–200k; With any health insurance; With private health insurance; No health insurance; Employed with health insurance; Employed with private health insurance; Employed without health insurance; Not labor force, private health | Householder; Not living with spouse or children; <9th grade education; 9–12th grade education; High school grad +; Native, born abroad; Born in Asia; Born in Latin America; Limited English speaking; Speak Spanish; Only speak Spanish; German ancestry; Take carpool to work; Family income, $25–35k; With any health insurance; No health insurance; Employed with health insurance; Employed with private health insurance; Employed without health insurance; Married couple family in poverty; House built 1950–1959; Age <18, total; Age >16, total; Age >18; One race, other; Other race; Hispanic/Latino; Mexican; Not Hispanic/Latino; Non-Hispanic White | Single female, with children; Householder; Not living with spouse or children; <9th grade education; Bachelors degree; High school grad +; Native, born abroad; Foreign Born, naturalized; Foreign Born, not US citizen; Born in Asia; Born in Latin America; Speak Spanish; Only speak Spanish; English ancestry; German ancestry; Irish ancestry; Italian ancestry; Scotch-Irish ancestry; Scottish ancestry; Take carpool to work; Family income, $25–35k; With any health insurance; No health insurance; Employed with health insurance; Employed with private health insurance; Employed without health insurance; 2 units per building; House built 2000–2009; House built 1950–1959; <1 person/room; 1–1.5 people/room; Housing cost (mortgage) $1.5–2k; Housing cost (mortgage)/income <20%; Housing cost (mortgage)/income 20–25%; Housing cost | Not living with spouse or children; <9th grade education; High school grad +; Speak Spanish; Only speak Spanish; English ancestry; German ancestry; Take carpool to work; Family income, $25–35k; With any health insurance; No health insurance; Employed with health insurance; Employed without health insurance; 2 units per building; <1 person/room; 1–1.5 people/room; Housing cost (mortgage)/income >35%; One race, other; Other race; Hispanic/Latino; Mexican; Not Hispanic/Latino; Non-Hispanic; White | Living with spouse; Male, unmarried; <9th grade education; High school grad +; Speak Spanish; Only speak Spanish; English ancestry; German ancestry; Family income, $25–35k; With any health insurance; No health insurance; Employed with health insurance; Employed with private health insurance; Employed without health insurance; 2 units per building; Age 55–59; Age 60–64; One race, other; Other race; Hispanic/Latino; Mexican; Not Hispanic/Latino; Non-Hispanic White | Male, unmarried; Female, never married; <9th grade education; High school grad +; Speak Spanish; Only speak Spanish; English ancestry; German ancestry; With any health insurance; No health insurance; Employed with health insurance; Employed without health insurance; 2 units per building; Moved in current house 2010–14; Age 60–64; Age 65–74; Hispanic/Latino; Not Hispanic/Latino; Non-Hispanic White | No phone in household | - |

**Table 2** (continued)

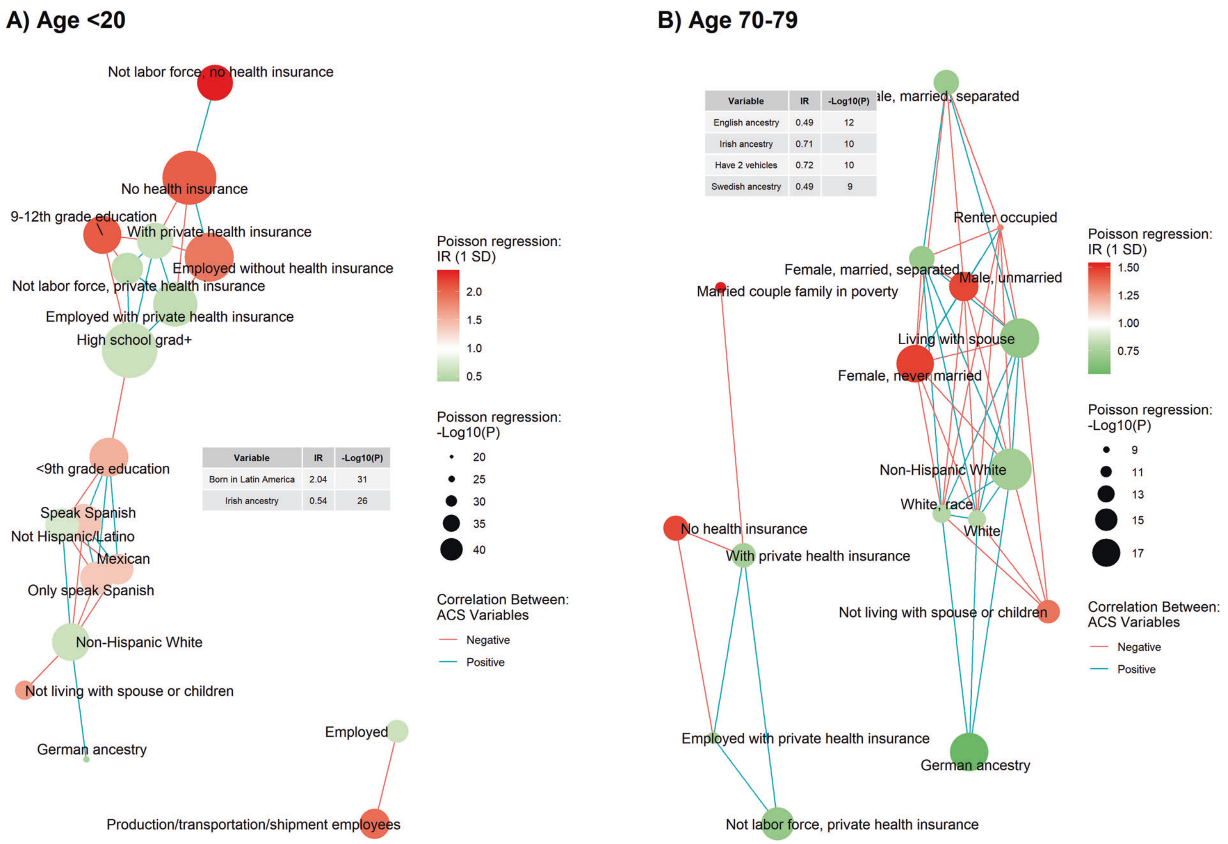
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | insurance; Married couple family in poverty; Family in poverty with kid <18 y; Housing cost $250–400; One race, other; Other race; Hispanic/Latino; Mexican; Not Hispanic/Latino; Non-Hispanic White | | (mortgage)/income >35%; One race, other; Other race; Hispanic/Latino; Mexican; Not Hispanic/Latino; Non-Hispanic White | | | | | |
| Variable selection step 2: Remove variables with VIF > 2.5 | | | | | | | | |
| Num. of variables in the final model | 8 | 7 | 11 | 4 | 5 | 5 | 1 | 0 |
| Variables in the final model | Living with grandchild for 1–2 years; Native, born abroad; Foreign Born, naturalized; Irish ancestry; Take carpool to work; Manufacturing employees; Family in poverty with kid <18 y; Housing cost $250–400 | Native, born abroad; Born in Asia; German ancestry; Take carpool to work; Married couple family in poverty; House built 1950–1959; Age >16, total | Householder; Native, born abroad; Foreign Born, naturalized; Born in Asia; Italian ancestry; Scotch-Irish ancestry; Take carpool to work; House built 1950–1959; Housing cost (mortgage) $1.5–2k; Housing cost (mortgage)/income <20%; Housing cost (mortgage)/income 20–25% | English ancestry; Take carpool to work; 2 units per building; 1–1.5 people/room | Male, unmarried; English ancestry; 2 units per building; Age 60–64; Other race | English ancestry; 2 units per building; Moved in current house 2010–14; Age 60–64; Hispanic/Latino | No phone in household | |
| $R^2$ of final model | 0.69 | 0.73 | 0.80 | 0.75 | 0.73 | 0.66 | 0.24 | - |

The name of each ACS variable indicates the percent (%) of people comparing to the reference population in a zip code area. For example, the “2 units per building” means the percent (%) of people living in a building with two units among all people whose housing information was available.

cumulative incidence in the 70–79 age group. No ACS variable was identified as a significant predictor in the oldest group (80+). In the second step of multivariate analysis, after removing the ACS variables with higher VIF (>2.5), between one and 11 ACS variables were selected as predictors of COVID-19 incidence in different age groups. The relative number of ACS variables in the final model corresponded to the number of variables selected by Elastic Net regression, where the 30–40 age group had the most ACS predictors of COVID-19 cumulative incidence. This age group also had the highest coefficient of determination ($R^2$) (0.80), which indicates that the 11 ACS variables adjusted in the final model explained 80% of the variation in COVID-19 incidence among people 30–40 years old. Based on $R^2$, the ability of ACS variables to explain the variation of COVID-19 cumulative incidence increased from the <20-year-old to the 30–40 age group, then decreased as age increased. The estimated IRs and 95% CIs of ACS variables in the final models are shown in Fig. 3. The oldest age group (80+) results were not included in Fig. 3 because no ACS variables were retained after variable selection.

We observed several common ACS variables in the final models across different age groups (Table 2 and Fig. 3). For example, from the youngest to the 40–50-year-old age group, the percentage of people who carpool to get to work was the most important predictor of COVID-19 cumulative incidence. Older housing (built 1950–1959) and crowded housing (e.g., multiple units in a building, more than one resident per room) were also significant predictors of COVID-19 cumulative incidence in adult groups (>20 years old). Several protective factors were also identified. ACS variables related to European-ancestry were associated with a lower incidence of COVID-19, ranging from the youngest age group to the 60–70-year-old age group. In addition, multiple ACS variables related to low housing cost compared to income (e.g., housing <25% of income) were associated with lower COVID-19 cumulative incidence in the 30–40-year-old age group.

We compared the results of pairwise analysis and multivariate analysis of the ACS variables and their associations with age-specific COVID-19 cumulative incidence. We identified several factors in common between the results of



| Variable | IR | -Log10(P) |
|---|---|---|
| Born in Latin America | 2.04 | 31 |
| Irish ancestry | 0.54 | 26 |

| Variable | IR | -Log10(P) |
|---|---|---|
| English ancestry | 0.49 | 12 |
| Irish ancestry | 0.71 | 10 |
| Have 2 vehicles | 0.72 | 10 |
| Swedish ancestry | 0.49 | 9 |

**Fig. 3 Network of two age groups (<20 and 70–79).** Each graph consisted of the twenty American Community Survey (ACS) variables most significantly associated with the age-specific COVID-19 incidence. The associations between each ACS variable and the age-specific COVID-19 cumulative incidence were presented as specific color and size of node. Network was built based on the correlation between ACS variable. The network graphs of two age groups (**A**) <20 and (**B**) 70–79 were shown here. The networks for all age groups were available in the supplementary material (Figure S2-S9).

pairwise analysis and multivariate analysis results: (a) Hispanic ethnicity was significantly associated with incidence of COVID-19 disease; (b) Population-level marital status was significantly associated with COVID-19 incidence in older ages (e.g., older than 50) but not in younger ages; (c) ACS variables directly related to neighborhood socioeconomic status (e.g., family in poverty) had a stronger association with COVID-19 incidence in children (<20) than in older adults. Several inconsistencies were also observed: (a) variables related to health insurance status and educational attainment were strongly associated with COVID-19 incidence in the pairwise analysis and in the preliminary variable selection by Elastic net regression, but were excluded by VIF-based selection in the multivariate analysis; (b) the pairwise analysis and multivariable analysis identified different significant ACS factors related to income — housing cost as a percentage of income was highly associated with COVID-19 incidence in multivariable analysis, while the family income variables, which also represent the family financial situation, were associated with COVID-19 cumulative incidence in pairwise analysis.

## Discussion

Using a NWAS method and zip code-level COVID-19 and neighborhood characteristics data, we found that the incidence of COVID-19 disease early in the pandemic was unequally distributed among neighborhoods based on differing social, economic, demographic, and housing factors. This finding contributes to the evidence that SES-disadvantaged neighborhoods are associated with increased risk of a number of health problems, including COVID-19 [11, 14]. Second, age modified the associations between neighborhood characteristics and the COVID-19 incidence. To our knowledge, this is the first study to examine the cross-sectional association between neighborhood characteristics and COVID-19 incidence by age groups.

The age composition of the COVID-19 cases found in our study zip code areas was similar to that reported in China [2] and elsewhere in the U.S. [33] early in the pandemic. Among all COVID-19 cases, the proportion of people under 20 years old ranged from 1% to 10%, which was disproportionally lower compared to the proportion of people in this age range in the general population [7, 34]. In addition to the biological mechanisms highlighted in previous studies (e.g., ACE receptor expression [8]) our findings suggested that neighborhood environment might also contribute to differing COVID-19 incidences between children and other ages. In our study, COVID-19-associated neighborhood characteristics were more diverse and more significant in middle-aged adults than in children and young adults. This might be explained by greater exposure to certain neighborhood factors among middle-aged adults compared to younger people after social "lockdown" measures were enacted. Children's daily activities are usually restricted to their local community and school district, and became even more restricted after schools and community activities closed. Middle-aged adults' daily activities were more diverse before pandemic closures than those of children, including commuting, working, grocery shopping, etc., and many of these persisted by necessity despite broad lockdown orders.

Our study further confirmed that disadvantaged neighborhoods and older age jointly contribute to the elevated incidences of COVID-19 disease. In our study, zip code areas characterized by lower educational attainment (less than high school) and Hispanic ethnicity had a disproportionately higher incidence of COVID-19. These influences were more significant for children and middle-aged adults compared to older adults. Fewer ACS-measured neighborhood characteristics were associated with COVID-19 in adults older than 70, suggesting that older age remains the most important risk factor for COVID-19 [33, 35, 36]. Although fewer ACS variables were associated with COVID-19 in seniors than in younger ages, more than one-third of study neighborhood characteristics were still statistically significantly associated with COVID-19 incidence in people aged 60–79 years in the pairwise analysis. In multivariate analysis, neighborhood characteristics selected from an objective, multi-step approach contributed 66% and 24% of the variation in COVID-19 incidences in the 60–70 and 79–80 age groups, respectively. Considering the mortality risk of senior COVID-19 patients, the impact of this group's neighborhood environment cannot be ignored.

Our study identified several neighborhood characteristics that have not been linked with COVID-19 in previous studies. For example, in both pairwise analysis and multivariable analysis, residents' marital status was a significant neighborhood characteristic associated with the COVID-19 incidence early in the pandemic among people older than 50. COVID-19 incidence was positively associated with the percent of unmarried people and percent of people living alone at zip code levels. Likewise, COVID-19 incidence was inversely associated with the percentage of people living with a spouse. Notably, as depicted in the network graphs, zip codes with a higher percentage of married people also had a greater white population, which has been widely used as an indicator of neighborhood SES [37] and has been previously noted as a protective factor of COVID-19 [13, 14]. We originally thought that the effect of zip-code-level marriage rate might be instead due to the effect of SES. However, no marital variables were found in the networks of highly associated neighborhood characteristics of age groups younger than 50 years old in the pairwise analysis. Also, the multivariate analysis suggested a significant association between zip-code-level marriage rates and COVID-19 incidence in the 50–60-year-old age group after adjusting for other selected SES variables. These results suggest that the effect of marriage on COVID-19 incidence was distinct from SES among older adults. This is particularly interesting considering the infection risk conferred by the reduced possibility of social distancing among married couples cohabitating compared to unmarried individuals. A possible explanation is that locations with higher rates of older adults living with their spouses would likely have fewer people living in nursing homes, where residents are at high risk of highly infectious diseases like COVID-19. In Illinois, about half of counties reported COVID-19 outbreaks in long-term care facilities [38]. Living with a spouse might be a marker of not living in a long-term care facility, which may be why the zip code areas with more married seniors have lower COVID-19 incidence. Future analysis could include the determination of housing type (group living situation or not), which was not possible in the present study.

Transportation was found to be an influential factor of COVID-19 incidence in our study. In the pairwise analysis, the only neighborhood characteristic found to be significantly associated with COVID-19 incidence in people older than 80 was the percent of people who take public transportation to work in a zip code area. In multivariate analysis, the percent of people who take carpool to work was widely associated with increased COVID-19 incidences in age groups ranging from children to middle age. Public

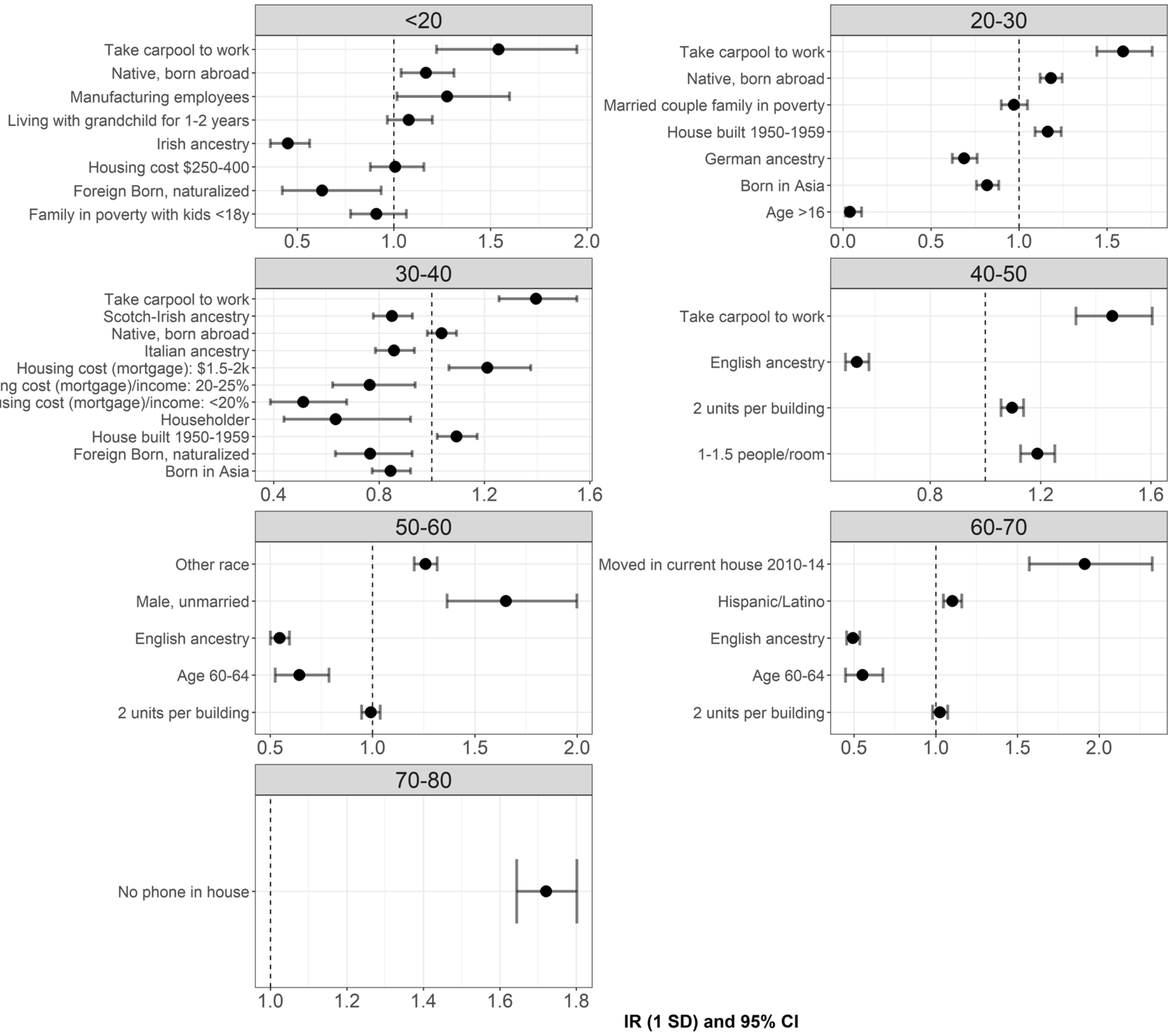


**Fig. 4 Results of multivariate analysis.** Incidence ratio and 95% CI between American Community Survey (ACS) variables and age-specific COVID-19 cumulative incidence estimated by multivariate Poisson regression. ACS variables adjusted in each Poisson model were selected using Elastic net regression and variance inflation factor (VIF). Unit of incidence ratio: change in the age-specific COVID-19 cumulative incidence (cases/1000 people) per one standard deviation (SD) of ACS variable. The name of each ACS variable indicates the percent (%) of people compared to the reference population in the study area. For example, the "2 units per building" means the percent (%) of people living in a building with two units among all people whose housing information was available in the study area.

transportation could be a hub of transmission for the SARS-CoV-2 virus. Reduced immune response and existing chronic diseases could elevate the risk of COVID-19 infection and severe syndromes in older adults [35, 36]. In general, carpools are considered less risky than public transportation, given people are contacting less frequently in carpools than in public transportations. But the increased risk of COVID-19 among zip codes with more people takes carpools indicated that the reduction of contact at this extent could not efficiently prevent COVID-19.

Our study has several limitations. First, our ecological study design cannot make causal inferences between neighborhood characteristics and COVID-19 disease. Second, although ZCTAs have been widely used in public health studies to estimate sociodemographic status at matched postal zip code levels [24], this study is subject to possible minor misclassification because the neighborhood characteristics in our study were measured by ACS at the geographic unit ZCTA, while COVID-19 cases were reported by postal zip codes. Another limitation which universally exists in studies using the pairwise analysis and NWAS framework is that the study is an agnostic hypothesis-generating rather than hypothesis-testing approach. This limitation is inherent in our exploratory study design, which was to identify new neighborhood characteristics whose associations with age-specific COVID-19 incidence could have been missed by methodologies.

Despite these limitations, our study applied a novel method to examine the association between COVID-19 infection rates and neighborhood characteristics and age. We systematically examined associations between 392 ACS variables and age-specific COVID-19 incidence. The visualization tools we used (heatmap and networks) proved to be powerful ways of exhibiting the association between neighborhood characteristics and COVID-19 incidence across age groups. The variable selection and multivariate regression analysis estimated associations between neighborhood characteristics and age-specific COVID-19 incidence, adjusting for other neighborhood characteristics as possible confounders. The identified neighborhood characteristics associated with age-specific COVID-19 incidence highlight opportunities for further study of high-risk populations by their age and living environment.

## Conclusion

In the present study, we systematically reviewed the association between 392 ACS-measured neighborhood characteristics and age-specific COVID-19 incidence in 214 Illinois zip code areas. COVID-19 cases were disproportionally distributed in neighborhoods characterized by lower educational attainment, crowded households, and higher numbers of essential workers. Lower neighborhood SES was a significant risk factor for COVID-19 across age groups, particularly among children. However, neighborhood social characteristics such as marital status and housing conditions were additional risk factors for COVID-19 among seniors. Further study to better understand the epidemiology of COVID-19 should consider both the neighborhood environment and the populations' age.

## Code availability

The entire analytic procedure is depicted in Fig. 1. R code and detailed analysis results are available on GitHub [32].

**Acknowledgements** We thank David Eads and his team for preparing the download repository of the COVID-19 zip code-level data published by IDPH. We thank Dr. Elena Colicino at Icahn School of Medicine at Mount Sinai for consulting the GWAS statistical method. Dr. Annemarie Stroustrup's research is supported by the National Institute of Environmental Health Sciences (NIEHS) mentored award K23ES022268. Dr. Xueying Zhang is supported by grants 5UG3OD023320-02 and UH3OD023337 from the National Institutes of Health for the Environmental Influences on Child Health Outcomes (ECHO) program.

## Compliance with ethical standards

**Conflict of interest** All authors declare no competing interests.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

1. Wu JT, et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nat Med. 2020;26:506–10.
2. Pan A, et al. Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China. JAMA. 2020;323:1915–23.
3. Onder G, Rezza G, Brusaferro S. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. JAMA. 2020;323:1775–6.
4. Ludvigsson JF. Systematic review of COVID-19 in children shows milder cases and a better prognosis than adults. Acta Paediatrica. 2020;109:1088–95.
5. Grubesic TH, Matisziw TC. On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. Int J Health Geogr. 2006;5: https://doi.org/10.1186/1476-072X-5-58.
6. CDC. Coronavirus Disease 2019 in Children—United States, February 12-April 2, 2020. MMWR Morb Mortal Wkly Rep. 2020;69:422–6.
7. American Academy of Pediatrics. Children and COVID-19: State-Level Data Report. Retrieved October 27, 2020, from https://services.aap.org/en/pages/2019-novel-coronavirus-covid-19-infections/children-and-covid-19-state-level-data-report/.
8. Li W, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature. 2003;426:450–4.

9. Bunyavanich S, Do A, Vicencio A. Nasal gene expression of angiotensin-converting enzyme 2 in children and adults. JAMA. 2020;323:2427–9.
10. Moya J, Bearer CF, Etzel RA. Children's behavior and physiology and how it affects exposure to environmental contaminants. Pediatrics. 2004;113:996–1006.
11. Carrion D, et al. Assessing capacity to social distance and neighborhood-level health disparities during the COVID-19 pandemic. medRxiv. 2020. https://doi.org/10.1101/2020.06.02.20120790.
12. Zhang CH, Schwartz GG. Spatial disparities in coronavirus incidence and mortality in the United States: an ecological analysis as of May 2020. J Rural Health. 2020;36:433–45.
13. Maroko AR, Nash D, Pavilonis BT. COVID-19 and inequity: a comparative spatial analysis of New York City and Chicago hot spots. J Urban Health. 2020;97:461–70.
14. Bilal U, Barber S, Diez-Roux AV. Spatial Inequities in COVID-19 outcomes in 3 US Cities. medRxiv (2021): 2020-05. https://doi.org/10.1101/2020.05.01.20087833.
15. Sy K, Martinez ME, Rader B, White L, Socioeconomic Disparities in Subway Use and COVID-19 Outcomes in New York City, Am J Epidemiol, 2020; kwaa277, https://doi.org/10.1093/aje/kwaa277.
16. Figueroa JF, Wadhera RK, Lee D, Yeh RW, Sommers BD. Community-level factors associated with racial and ethnic disparities in COVID-19 rates in Massachusetts. Health Aff. 2020;39:1984–92.
17. US Census Bureau. 2010 census urban and rural classification and urban area criteria. Retrieved October 27, 2020, from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html.
18. National Research Council. Using the American Community Survey: benefits and challenges. Washington, DC: The National Academies Press https://doi.org/10.17226/11901.
19. Lynch SM, et al. A neighborhood-wide association study (NWAS): example of prostate cancer aggressiveness. PloS One. 2017;12: https://doi.org/10.1371/journal.pone.0174548.
20. Mooney SJ, Joshi S, Cerdá M, Kennedy GJ, Beard JR, Rundle AG. Contextual correlates of physical activity among older adults: a Neighborhood Environment-Wide Association Study (NE-WAS). Cancer Epidemiol Biomark Prev. 2017;26:495–504.
21. Lynch SM. Book Chapter: towards systematic methods in an era of big data: neighborhood wide association studies. In Berger NA, Berrigan D, editors. Geospatial approaches to energy balance and breast cancer. Springer, 2019.
22. Eads DM, COVID-19 in Illinois—ZIP code lookup. Retrieved October 27, 2020, from https://www.chicagoreporter.com/tracking-coronavirus-cases-in-illinois-daily/.
23. US Census Bureau. American community survey design and methodology report. Retrieved October 27, 2020, from https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html.
24. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2014.
25. Walker K. tidycensus: Load US Census boundary and attribute data as 'tidyverse'and 'sf'-ready data frames. R package version 0.4. 1. 2018. Last time download: Feburary 6, 2021.
26. Frome EL, Checkoway H. Use of poisson regression models in estimating incidence rates and ratios. Am J Epidemiol. 1985;121:309–23.
27. Pedersen TL Retrieved Feburary 6, 2021, from https://www.data-imaginist.com/2017/introducing-tidygraph/.
28. Pedersen TLR Package 'ggraph'. Last time download: Feburary 6, 2021.
29. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101:1418–29.
30. Braun JM, Gennings C, Hauser R, Thomas WF. What can epidemiological studies tell us about the impact of chemical mixtures on human health? Environ Health Persp. 2016;124: https://doi.org/10.1289/ehp.1510569.
31. Friedman J, et al. glmnet: Lasso and elastic-net regularized generalized linear models. Retrieved Feburary 7, 2020. from https://cran.r-project.org/web/packages/glmnet/index.html.
32. GitHub repository COVID19_IDPH: Retrieved October 27, 2020, from https://github.com/zxy1219/COVID19_IDPH.
33. CDC. Severe outcomes among patients with coronavirus disease 2019 (COVID-19) — United States, February 12–March 16, 2020. Morb Mortal Wkly Rep. 2020;69;343–6.
34. Kaiser Family Foundation. Population distribution by age. Retrieved October 27, 2020, from https://www.kff.org/other/state-indicator/distribution-by-age/%currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D.
35. Alpert A, et al. A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. Nat Med. 2019;25:487–95.
36. Koff WC, Williams MA. Covid-19 and immunity in aging populations—a new research agenda. N. Eng J Med. 2020;383:804–5.
37. Yen IH, Michael TL, Perdue L. Neighborhood environment in studies of health of older adults: a systematic review. Am J Pre Med. 2009;47:455–63.
38. IDPH Long-Term Care Facility Outbreaks COVID-19. Retrieved October 27, 2020, from https://www.dph.illinois.gov/covid19/long-term-care-facility-outbreaks-covid-19.