Distribution Agreement

Signature:

______________________ ______________

Hillary Bonuedie **Date**

Correlates of Missing Information on Race and Ethnicity in the Georgia COVID-19 Surveillance Database

By

Hillary Bonuedie
MPH

Epidemiology

____
Shivani A. Patel, MPH, PhD
Committee Chair

____
Umedjon Ibragimov, MD, MPH, PhD
Committee Member

Correlates of Missing Information on Race and Ethnicity in the Georgia COVID-19 Surveillance Database

By

Hillary Bonuedie

B.A.
Columbia University
2016

Thesis Committee Chair: Shivani A. Patel, MPH, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2021

Abstract

Correlates of Missing Information on Race and Ethnicity in the Georgia COVID-19 Surveillance Database

By Hillary Bonuedie

Objectives: We quantified the amount and pattern of missing race and ethnicity information among COVID-19 cases and deaths in Georgia at the individual and county levels.

Methods: Among confirmed COVID-19 cases and deaths recorded between April 2020 to March 2021 in the Georgia Department of Public Health surveillance database, the percentage of cases and deaths missing race and ethnicity data was calculated. This was compared with the percentage of missing age, sex, zip code, and county of residence among COVID-19 cases and deaths. At the individual level, correlates of missing race and ethnicity information among COVID-19 cases were identified using logistic regression. At the county-level, linear regression was used to identify correlates of differences in the percentage of cases with missing race and ethnicity.

Results: Confirmed COVID-19 cases were missing race and ethnicity information more often than confirmed COVID-19 deaths. The difference in missingness between cases and deaths, respectively, was more pronounced for information on race (18.6% vs 0.8%) and ethnicity (27.6% vs 1.0%) than for age (0.7% vs 0.01%), sex (1.1% vs 0.1%), zip code (2.3% vs 0.8%), or county of residence (2.0% vs 0.26%). At the individual level, in a logistic regression model of COVID-19 cases, males ages 0-17 (OR = 2.50; 95% CI: 2.05, 3.04) and males ages 18-64 (OR = 1.69; 95% CI: 1.49, 1.92) had higher relative odds of missing race information when compared with women ages 65+; age and sex patterns for missing ethnicity information were similar. At the county level, in adjusted linear regression models, the main correlates of the percentage of cases missing race and ethnicity, respectively were case rate (β=0.99; 95% CI: 0.27, 1.72 for race and β=1.47; 95% CI: 0.55, 2.38 for ethnicity) and the percent of the county population reporting multiple races (β=0.76; 95% CI: 0.026, 1.49 for race and β=1.18; 95% CI: 0.25, 2.10 for ethnicity).

Conclusions: The Georgia COVID-19 surveillance system was successful at collecting age and sex information, yet race data were missing for nearly 1 in 5 cases. Both individual demographics and county-level characteristics were informative in predicting missing race and ethnicity information among cases.

Correlates of Missing Information on Race and Ethnicity in the Georgia COVID-19 Surveillance Database

By

Hillary Bonuedie

B.A.
Columbia University
2016

Thesis Committee Chair: Shivani A. Patel, MPH, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2021

Acknowledgements

I would like to thank my committee chair Dr. Shivani Patel for her expertise, time, encouragement, and patience throughout the thesis process. When I decided to start working on a COVID-19 project, there was a lot of uncertainty because we were in the early stages of the pandemic. I'm grateful Dr. Patel was able to make a place for me on her team.

I'd also like to thank my committee member Dr. Umedjon Ibragimov for his contributions and advice throughout the process. Additionally, I'd like to acknowledge the Emory COVID-19 Response Collaborative and the Georgia Department of Public Health.

Finally, I'd like to thank my family and friends who were a source of motivation and cheered me on throughout the thesis process.

Table of Contents

Introduction

Ensuring an equitable national response to the COVID-19 pandemic requires collection of detailed sociodemographic data on confirmed cases across geography. Unfortunately, there have been gaps in US COVID-19 surveillance, largely around race and ethnicity.[1] In June 2020, the federal government responded to criticisms of gaps in race and ethnicity data by issuing guidance to prompt laboratories and testing sites to report race and ethnicity data to state and local health departments no later than August 1, 2020.[2] Yet, on August 28, 2020, 51% of cases were missing data on race and ethnicity.[2] As recently as April 14, 2021, 49% of cases were missing data on race and ethnicity, according to the CDC COVID Data Tracker.[3] The amount of missing data has fluctuated over time and the majority of existing analyses of the incidence of COVID-19 cases by race and ethnicity are limited to reported data.

Gaps in US COVID-19 surveillance data have been widely observed. Spencer et al suggest that many of these gaps arise from under-resourced state, local, and territorial health departments, which leads to slower turnaround time in COVID-19 contact tracing.[4] However, time from positive test results to contact tracing may impact the ability for a case to recall details around COVID-19 exposure source, but time is unlikely to affect a case's ability to report on key demographic information such as age, sex, race, and ethnicity. As such, there is still a knowledge gap on what contributes to better quality data collection.

Invalid estimates of racial disparities could pose a problem because data are used to inform the COVID-19 response in the US. Community advocates and state leaders use this data to advocate for limited federal resources. The existing and rapidly developing body of literature suggests that current estimates of racial disparities may underestimate the impact of COVID-19 on communities of color. However, with so much unknown race and ethnicity information, there is a lot of uncertainty. Therefore, there is a strong need in COVID-19 surveillance to identify and remedy causes for unreported race and ethnicity. Doing so can help direct resources where they are needed most and allow the US to efficiently slow the spread of COVID-19.

To inform approaches to fill gaps in race and ethnicity information in COVID-19 surveillance, we quantified the amount and pattern of missing race and ethnicity information among COVID-19 cases and deaths in Georgia at the individual and county levels.

Literature Review

Impact of COVID-19 on Communities of Color

Evidence on how COVID-19 was affecting American communities indicated that not all communities were equally affected. An early CDC publication found that in June 2020, 96.2% of the 79 counties identified as hotspots had incident COVID-19 racial and ethnic disparities.[5] Findings from this study indicated that racial and ethnic minority groups had higher incidence of COVID-19 infection than would be expected, based on their share of the population. Hispanics/Latinos were the largest minority group disproportionately affected and were found to have COVID-19 incidence rates 4.4 times higher than what would be expected given their representation in the population.[5] Other racial and ethnic minority groups disproportionately affected by COVID-19 were Blacks/African Americans, American Indians/Alaska Natives, Asians, and Native Hawaiians/Pacific Islanders.

A similar phenomenon was identified among individuals who had died from COVID-19. Rossen et al found that at 53.6%, Hispanics experienced the largest percent increase in excess deaths due to COVID-19.[6] Following Hispanics, were Asians (36.6%), Blacks/African Americans (32.9%), and American Indians/Alaska Natives (28.9%). For comparison, during the same time period, Whites experienced an 11.8% average increase in excess deaths.[6] And so, while COVID-19 was ravaging the entire US population, racial and ethnic minorities were bearing the brunt of the disease burden.

Additional research was published which attempted to move beyond surveillance and start disentangling societal mechanisms driving the disproportionate amount of the COVID-19 disease burden among racial and ethnic minority groups. The research found that racial and ethnic minority groups did not have the same flexibility or permissiveness to perform recommended protective behaviors to protect against COVID-19 infection. Occupation, employment benefits, and housing were more often prohibitive.

Based on annual reports from the Bureau of Labor Statistics, we know that within the US, African Americans and Hispanics have lower incomes than Whites.[7] We also know that African Americans and Hispanics are more likely to have jobs that require in-person work.[8] During the pandemic, non-essential workers were able to work remotely, while essential workers continued in-person work. Macias Gil et al note that Hispanics were overrepresented in essential jobs; compared to 31.4% of non-Hispanic workers who could work remotely, only 16.2% of Hispanic workers could engage in remote work.[9] At the same time, Billock et al found that low income workers were nearly three times more likely than high income workers to be prohibited from using preventive measures, such as masks and physical barriers, at work.[10] Further, low income workers were also nearly three times more likely than high income workers to be unable to acquire preventive measures at work.[10] These work conditions heighten the risk for COVID-19 infection. In an investigation of meat and poultry plant processing facilities, Dyal et al found that the risk of COVID-19 infection was higher in congregate work settings. Some structural barriers identified were the inability to physical distance, difficulty with maintaining cleaning procedures, lack of sick pay, and nonadherence to consistent mask use .[11]

In addition to occupation and employment benefits, shared housing, particularly congregate housing, has been identified as a facilitator of the spread of COVID-19 infection.[11,12] Housing, as a social determinant, intersects with race and ethnicity in that racial and ethnic minorities are more likely to live in multi-generational households. Asian-Americans are more likely than White Americans to provide support to elderly family members[13] and Black workers are twice as likely as White workers to live with three or more generations.[14] Additionally, compared to Whites, more Hispanics and African Americans live in houses with at least one essential worker.[15] Multigenerational households are significant factors in COVID-19 incidence because preventive measures, such as physical distancing, quarantining, and isolation, are more difficult to maintain.

Impact of Incomplete Data on COVID-19 Analyses Examining Racial Disparities

Most race and ethnicity data are missing among COVID-19 cases. To account for this, scientists are employing different methods to handle the missing data. Laurencin and McClinton report on the 45% of cases whose data are reported.[16] Moore et al drop 61.5% of the counties from their analysis for failing to have race and ethnicity data on at least half of their cases.[6] Labgold et al. have race and ethnicity on 64% of cases, and so drop 36% of their sample for their complete case analysis (but notably account for missing race and ethnicity in their methodology).[17] Other scientists choose to reclassify missing data. Rossen et al create an "other/unknown race/ethnicity [category], which included non-Hispanic Native Hawaiian or other Pacific Islander, non-Hispanic multiracial, and unknown."[6] This allows them to keep all individuals in their sample, but introduces the possibility for misclassification. Individuals with unknown or missing race and ethnicity may actually belong in another category. Finally, other scientists have employed a mix of complete case analysis and reclassification. Garg et al reported the racial and ethnic distribution of 45% of patients they had race and ethnicity data for. Additionally, they had a category for "other or unknown" race.[18] So while most scientists choose to use traditional complete case analysis, others opt for reclassification.

The amount of missing race and ethnicity data may or may not present a problem depending on if data are missing differentially by race or ethnicity. If the data were to be missing completely at random, current comparisons, which mostly rely on complete case analysis, of measures of frequency among different racial and ethnic minority groups would be valid.[19] However, publications examining racial disparities imply that the missing data are not missing completely at random.[17,20] Further, while most researchers choose to use complete case analysis, there are inconsistent methods of handling the missing data. These inconsistencies introduce challenges to comparing measures of racial disparities across studies.

Recent publications have attempted to account for the missing data using weighted and unweighted population counts[20] and Bayesian Improved Surname Geocoding.[17] When COVID-19 population data were initially released, CDC did not provide a breakdown by race or ethnicity. Using weighted population counts provided by CDC and unweighted population counts provided by the US Census, Cowger et al compared distributions of COVID-19 mortality by race and ethnicity. They found that the weighted population counts underestimated disease burden among Black and Hispanic individuals. Among Black people, the unweighted method calculated a risk ratio of 1.79, versus 1.23 using the weighted population counts. Among Hispanic people, those risk ratios were .91 and .62, respectively.[20] Cowger et al argue that the weighted population count methodology was over-adjusting; by adjusting for the geographical distribution of racial groups, the methodology ignored that Black, and Hispanic individuals tend to live in the high density, urban locations which were among the hardest hit. Similar conclusions were found months later in Fulton County, GA by Labgold et al. Labgold et al found that complete case analysis underestimated COVID-19 racial disparities. Using imputation and bias adjustment, Labgold et al found that COVID-19 incidence increased by 1.8-fold for Asians, 1.7-fold for Whites, 1.7-fold for Hispanics, 1.6-fold for Other, and 1.5-fold for Blacks.[17]

In the midst of a quickly evolving pandemic, data surveillance infrastructure needs to be set up quickly and efficiently.[21] Understanding whether and which sociodemographic groups are more likely to be missing race and ethnicity information may assist in redressing some of the gaps in these key data points. Furthermore, in the US, state and local government are often at the forefront of policy and decision making. Therefore, it is critical to know whether features of the county environment are related to missing race and ethnicity information in COVID-19 case and death surveillance. Understanding these county features, alongside individual factors, may improve data collection and allow more resources to be preemptively directed to areas that will need greater support. Not only does this approach inform equity considerations, but ensuring the health of the most vulnerable people in society also improves the health of the whole society.

Methods

Datasets

This analysis was performed under the Emory COVID-19 Response Collaborative as public health practice. The principal data source used was the Georgia Department of Public Health (GDPH) COVID-19 case and death surveillance database. Additionally, social variables were obtained using publicly available data files. Namely, the Agency for Healthcare Research and Quality (AHRQ) SDOH Database, the Georgia Secretary of State Voter Registration Statistics, and the United States Census Bureau TIGER/Line Shapefiles.

The GDPH COVID-19 case and death surveillance database contains data from all individuals in Georgia suspected with COVID-19 based on test results and other reports sent to the GDPH. The database includes multiple indicators acquired during the process of case investigation, including information on demographics, residential address, county of testing, pre-existing health conditions, source of exposure, case identification, data collection method, hospitalization status, and death due to COVID-19.

The AHRQ database is a repository of government datasets. Data are organized using SDOH domains: social context, economic context, education, physical infrastructure, and healthcare context. While data are available as early as 2009 in the AHRQ database, this analysis used 2018 data. The American Community Survey provided population counts, racial and ethnic composition, English language proficiency, economic inequality, poverty, insurance status, and median home value variables. The Robert Wood Johnson Foundation County Health Rankings provided the segregation variables, the Centers for Disease Control and Prevention (CDC) provided SVI, the Health Resources and Services Administration (HRSA) included area health resources files which included the health professional shortage area (HPSA) variables.

Information on voter registration, which included the total number of voters by age, sex, and race, was provided by the Georgia Secretary of State office.

Lastly, shapefiles were obtained from the United States Census Bureau. 2020 TIGER/Line Shapefiles were used. These files included county boundaries and FIPS codes.

COVID-19 cases and deaths

At the individual level, the study population included all records of confirmed cases of COVID-19, defined as an individual with a positive molecular test result, from April 2020 to March 2021. These test results were reported through multiple sources including Electronic Lab Reporting (ELR), State Electronic Notifiable Disease Surveillance System (SendSS), faxed case reports, and calls from providers to GDPH. Deaths due to COVID-19 refer to confirmed deaths, which were defined as confirmed COVID-19 cases that were either reported to GDPH as deceased by healthcare providers or medical examiners/coroners, identified by death certificates with COVID-19 indicated as the cause of death, or there was evidence that COVID-19 contributed to the individual's death.[22] Race and ethnicity information are generally obtained for COVID-19 cases at the time of testing; the data available on each individual at the time of testing become part of the case record. If missing, demographic fields are updated at the time of case investigation by GDPH staff. In addition, for deaths, demographic information is also recovered from the death certificate. Finally, GDPH used drivers' licenses and other state-issued identification to recover age, sex, and race information that was missing after case investigation was completed.

The April 2020 start date was chosen to avoid getting data in the first few months of the pandemic, in which testing capacity was low, and there were fewer than 100 confirmed cases of

COVID-19 in Georgia. Moreover, April 5, 2020 is when COVID-19 became a nationally notifiable condition and an updated interim case surveillance definition was published by the Council of State and Territorial Epidemiologists.[23] The March 2021 end date was chosen based on the time needed for GDPH to clean COVID-19 race and ethnicity data. Furthermore, this period represented the first year of a full court press national pandemic response.

Individual-level measures of missing race and ethnicity information among COVID-19 cases and deaths

At the individual level, there were two outcomes of interest – missing race and missing ethnicity information. Three missing indicator variables were created for race and ethnicity: race only, ethnicity only, and both race and ethnicity. The purpose of creating three different indicators for race and ethnicity was to identify in descriptive analysis if both variables were contributing equally or if one of the two variables was contributing more to the overall percentage of missing race and ethnicity data. Each missing indicator variable was coded as a binary variable, with values of 0 (data reported) or 1 (data missing). Race and ethnicity were considered missing if there was no recorded value in the dataset or if the recorded value was “unknown”.

Individual-level measures of missing information

Additionally, missing indicator variables were created for age, sex, county of residence, and zip code. These indicators were created to ascertain the broad context of data collection quality of commonly collected demographic variables and provide a comparison for missing race and ethnicity information.

County-level measures of missing race and ethnicity information among COVID-19 cases and deaths

The percentage of missing race and ethnicity information across cases and deaths were calculated for each of Georgia's 159 counties. This was computed only among cases and deaths residing within a Georgia county. Records missing county of residence and out of state residents were excluded.

Individual-level correlates of missingness

At the individual level, there were two exposure variables of interest – age and sex. Age was categorized into a 3-level variable representing children (ages 0 – 17), adults (ages 18 – 64), and elderly adults (ages 65+) and sex was dichotomous (male and female).

County-level correlates of missingness

The main exposures of interest at the county level were the percent of the population that was Black, the percent of the population that was Hispanic, the percent of the population reporting multiple races, the percent of the population that spoke English less than well, the percent of the population in poverty, and the percent of the population that was uninsured. In addition, we treated the county case rate (per 1000) as an exposure, due to its potential impact on contact tracing load. The case rate numerator was calculated as the total number of confirmed COVID-19 cases from April 2020 to March 2021 in each county, and the case rate denominator was the county's total population; the resulting number was multiplied by 1000 to obtain confirmed cases per 1000 residents.

Other county variables considered for this analysis were the segregation index, median home value, shortage of primary care physicians, average hours per day of licensed staff (registered nurses and licensed practical nurses) per reporting facility, English language proficiency (percentage of population that only spoke English, did not speak English at all, did not speak English well, spoke English very well, and spoke English well), the Gini index, the number of registered drivers, and the number of voters.

Statistical analysis

Statewide percentages of missing race and ethnicity information among cases and among individuals who died from COVID-19 was calculated by month. The month in which a case tested positive was used for this stratification.

We next described the age and sex composition of COVID-19 cases and deaths with and without missing race and ethnicity information. Chi-square tests were conducted to evaluate statistical differences in the demographic composition of records with and without missing race and ethnicity information. To examine county and demographic factors associated with individual-level missing information on race and ethnicity among confirmed COVID-19 cases, we estimated logistic regression models. Logistic regression models were estimated with generalized estimating equations (GEE)[24] to account for clustering of cases within counties. Adjusted models that accounted for all exposures (age, sex, the percent of population that was Black, the percent of the population that was Hispanic, the percent of the population that reported multiple races, the percent of the population that spoke English less than well, the percent of the population in poverty, the percent of the population that was uninsured, and case rate) were reported. We observed an interaction between age and sex categories and missingness of race data; we therefore reported these associations jointly. For consistency, we also estimated the association of age and sex categories jointly in the model of ethnicity.

County-level data missingness analysis included spatial visualization of missing race, ethnicity, zip code, age, and sex. Each visualization was produced by splitting each indicator into quintiles; visualizations were produced to see overall missingness and missingness after accounting for the number of cases in each county. For reference, a visualization was also produced to show the total case rates per county.

County-level missing data were also examined through linear regression. The unadjusted association between each county characteristic (the percent of population that was Black, the percent of the population that was Hispanic, the percent of the population that reported multiple races, the percent of the population that spoke English less than well, the percent of the population in poverty, and the percent of the population that was uninsured) and county-level percentage of missing race and ethnicity was assessed. Then an adjusted model, which included all relevant county characteristics, was used to model missing race and ethnicity. All county-level exposure variables were categorized into quintiles.

All statistical analysis was conducted using SAS 9.4 (Cary, NC) and ArcGIS Desktop 10.8 (Redlands, CA).

Results

Percentage of missingness at the state level

840,424 COVID-19 cases and 16,286 deaths from COVID-19 were confirmed in Georgia from April 2020 to March 2021. Among cases, 18.6% had race missing, 27.6% had ethnicity missing, 16.0% had both race and ethnicity missing, 0.7% had age missing, 1.1% had sex missing, 2.3% had zip code missing, and 2.0% had county of residence missing. Among deaths, 0.8% had race missing, 1.0% had ethnicity missing, 0.3% had both race and ethnicity missing, 0.01% had age missing, 0.1% had sex missing, 0.8% had zip code missing, and 0.26% had county of residence missing (Table 1).

Temporal analysis of missing data revealed a slight upward trend in missing race and ethnicity over time among cases (Figure 1). Among deaths, the percentage of missing race and ethnicity stayed relatively constant until around December 2020. From December 2020 to March 2021, there was a steep increase in the amount of missing race and ethnicity data among deaths (Figure 2). Still, on an absolute scale, the percentage of missing race and ethnicity data was still lower at all time points among deaths than among cases.

Spatially, the Atlanta metropolitan area had the largest number of confirmed COVID-19 cases with missing race, ethnicity, zip code, age, and sex information (Figure 3). However, when accounting for the number of cases, there was no one region in Georgia which had high levels of missing race, ethnicity, zip code, age, or sex data (Figure 4). For reference, visualizations of the total case rates have been provided as Supplementary Figures (Figure S1).

Correlates of individual-level missing race and ethnicity information among all confirmed COVID-19 cases

At the individual-level, younger cases had more missing race information (p<.0001) (Table 2a). Among cases with missing race the mean age was 37.9 (SD=18.7), and among cases with reported race the mean age was 41.7 (SD=20.2). Similarly, when considering the association between age and missing ethnicity, younger cases had more missing ethnicity (p<.0001) (Table 2a). The mean age among cases with missing ethnicity was 38.9 (SD=19.1) and 41.7 (SD=20.3) among cases with reported ethnicity (Table 2a).

Considering sex and missing race, male cases were more likely to have missing race information (p<.0001). 19.5% of males were missing race information, while 16.6% of females were missing race information. For ethnicity, males were also more likely to have missing ethnicity information (p<.0001). 28.8% of males were missing ethnicity information, while 25.4% of females were missing ethnicity information (Table 2a).

In the fully adjusted model of missing race among cases, a statistical interaction was found between age and sex (p= 0.0321). Compared with female cases ages 65 and older, males ages 0-17 (OR=2.50; 95% CI: 2.05, 3.04), females ages 0-17 (OR=2.18; 95% CI: 1.90, 2.51), males ages 18-64 (OR=1.69; 95% CI: 1.49, 1.92), females ages 18-64 (OR=1.48; 95% CI: 1.38, 1.59), and males ages 65 and older (OR=1.22; 95% CI: 1.09, 1.37) were more likely to have missing race information. No county factors – percent of the population that was Black, percent of the population that was Hispanic, percent of the population that reported multiple races, percent of the population living in poverty, percent of the population that was uninsured, or case rate – were associated with missing race at the individual level (Table 3).

In the model of missing ethnicity among cases, there was no interaction found between age and sex (p= 0.5014). However, age and sex were independently statistically significant. For comparability and consistency, joint effects of age and sex were reported in the ethnicity model. Compared with female cases ages 65 and older, males ages 0-17 (OR=1.90; 95% CI: 1.62, 2.23), females ages 0-17 (OR=1.68; 95% CI: 1.49, 1.89), males ages 18-64 (OR=1.47; 95% CI: 1.33, 1.62), females ages 18-64 (OR=1.29; 95% CI: 1.22, 1.38), and males ages 65 and older (OR=1.15; 95% CI: 1.05, 1.26) were more likely to have missing ethnicity information. In the ethnicity model, the percent of the population that was uninsured was inversely associated with missingness; cases living in counties in the highest quintile for the percent uninsured had lower odds of missing ethnicity information when compared to counties in the lowest quintile (OR=0.63; 95% CI: 0.40, 0.99) (Table 3). No other factors investigated were associated with missing ethnicity (Table 3).

Correlates of individual-level missing race and ethnicity information among all COVID-19 deaths

Among deaths, there were no sex differences in missing race (p=0.6393) or missing ethnicity (p=0.2515). However, there were age differences. Younger people were more likely to have race and ethnicity information missing than older people (p=0.0011). The mean age of deaths missing race was 67.6 (SD=16.9) whereas the mean age of deaths with reported race was 73.8 (SD=13.9). 1.3% of deceased adults ages 18-64, compared with 0.7% of adults ages 65 and older were missing race information. A similar trend was observed for ethnicity. Deaths missing ethnicity were more likely to be younger (p=0.0002). The mean age of deaths missing race was 66.1 (SD=17.9) whereas the mean age of deaths with reported race was 73.9 (SD=13.9). 1.6% of deceased adults ages 18-64, compared with 0.8% of adults ages 65 and older were missing race information (Table 2b).

Correlates of county-level percentage of missing race and ethnicity among cases

In the unadjusted models with missing race as the outcome variable, the percent of the population that was Hispanic (β=1.26; 95% CI: 0.59, 1.93), the percent of the population reporting multiple races (β=1.15; 95% CI: 0.47, 1.83), and the percent of the population speaking English less than well (β=1.15; 95% CI: 0.47, 1.82) were positively associated with missing race. The percent of the population that was Black (β=-0.93; 95% CI: -1.62, -0.25) and the percent of the population living in poverty (β=-0.76; 95% CI: -1.45, -0.067) were negatively associated with missing race, and there was no association between the percent of the population that was uninsured (β=0.027; 95% CI: -0.67, 0.73) and missing race. In the fully adjusted model, the percent of the population reporting multiple races (β=0.76; 95% CI: 0.026, 1.49) was positively associated with missing race, and all other exposures were found to have no association (Table 4).

In the unadjusted models with missing ethnicity as the outcome variable, the percent of the population that was Hispanic (β=1.21; 95% CI: 0.30, 2.12) and the percent of the population reporting multiple races (β=1.94; 95% CI: 1.07, 2.82) were positively associated with missing ethnicity. The percent of the population that was Black (β=-0.95; 95% CI: -1.86, -0.030), the percent of the population living in poverty (β= -1.93; 95% CI: -2.80, -1.05), and the percent of the population that was uninsured (β=-1.49; 95% CI: -2.39, -0.59), were negatively associated with missing ethnicity, and there was no association between the percent of the population speaking English less than well (β=0.59; 95% CI: -0.33, 1.51) and missing ethnicity. In the fully adjusted model, the percent of the population reporting multiple races (β=1.18; 95% CI: 0.25, 2.10), was positively associated with missing ethnicity, and all other exposures were found to have no association (Table 4).

Discussion, Conclusions, and Public Health Recommendations

As of May 2021, the US is in the 15th month of the COVID-19 pandemic.[25] Although race and ethnicity emerged as a critical dimension of health equity in the covid-19 pandemic, there is limited literature on factors related to the missing race and ethnicity information within

COVID-19 surveillance systems. Within Georgia, we found at the individual level that young men were more likely to have missing race and information. We also found that individuals living in counties with more uninsured residents were less likely to have missing ethnicity information. At the county level, high case rates and percentages of residents reporting multiple races were associated with more missing race and ethnicity data.

Previous research in the primary care setting has found that when race and ethnicity data are missing, individuals for whom the data are missing tend to be older and male.[26] The findings here are only partially consistent with those findings. At the individual level, missing race and missing ethnicity were more common among young men in this sample. In the context of COVID-19 in Georgia, there were two surges of COVID-19: one in summer 2020, and the other in winter 2020. In both of these surges, people ages 18-59 made up the majority of incident cases.[22] Given that this analysis found that counties with higher case rates had more missing data, it's possible that counties which were more impacted by the surges were overwhelmed and did not have the capacity to ensure complete data collection on all cases.

Also, individuals living in counties with higher percentages of uninsured residents were found to be associated with less missing ethnicity information. Previous research has found that undocumented immigrants are more likely to be uninsured.[27] Undocumented immigrants face barriers to healthcare access such as language and fear of deportation. Therefore, undocumented immigrants could be less likely to get tested for COVID-19 than their documented counterparts and would be underrepresented in this sample of confirmed cases.

At the county level, this analysis also found that the percentage of the population reporting multiple races was associated with missing race and missing ethnicity. One reason for this may be confusion around race and ethnicity categorization, particularly among Hispanics and people reporting multiple races.[28,29] Confusion around how to answer the two-question race and ethnicity questions may lead to no response. Notably, at the county level, insurance status was not statistically significant in models for missing race or missing ethnicity. One reason for this may be that COVID-19 testing was free and so did not depend on insurance status.

These findings suggest that greater efforts must be made to capture complete data on young men and encourage individuals living in counties with high percentages of uninsured residents to get tested for COVID-19. Lastly, these findings highlight how difficult it is for counties to maintain high levels of data completion when dealing with high case rates.

Strengths and Limitations

A strength of the study our use of individual records within the GDPH COVID-19 Surveillance Database, which enabled both individual- and county-level analyses of missing race and ethnicity information. These records allowed for: (1) comprehensive examination of race and ethnicity, cases and deaths, and view over time, (2) and merging of social variables into the analysis from other county sources.

Analysis of surveillance data, however, cannot directly elucidate the reasons for missing race and ethnicity, such as a case's reluctance to report or provider hesitancy or unwillingness to collect race and ethnicity information. In addition, results from this analysis may not be generalizable to the whole United States, but could be generalizable to 5 of 10 HHS regions (1, 4, 5, 6, and 10) which have similar levels of data missingness for race and ethnicity. Note that Georgia is part of region 4.[1]

Public Health Recommendations

This analysis was catalyzed by the data released from federal, state, and local governments which indicated that a non-negligible percentage of the race and ethnicity data were missing or unknown. Without the transparency around factors related to availability of critical demographic information to characterize cases and deaths, little would be known about the quality of data surveillance. Peer-reviewed articles commonly exclude the amount of missing race and ethnicity data and surveillance sites often do not include reports of whom the data are missing for; one non-COVID-19 systematic review found that nearly 1 of 2 peer-reviewed articles don’t mention the percent of the sample missing race or ethnicity.[30] The lack of information prevents readers from assessing the risk of bias. Therefore the government, at all levels, should continue releasing comprehensive COVID-19 data reports to the public. Moreover, they should continue to report the percentages of unknown or missing data to give more insight on the validity of estimates.

From the first part of the analysis, we learn that there are more missing data among COVID-19 cases than deaths for race, ethnicity, age, sex, county of residence, and zip code. What’s more, the race and ethnicity data have the largest difference in the amount of missingness between cases and deaths. The difference between cases and deaths is important to note as it informs us on what is working well and where there are areas for improvement within current COVID-19 surveillance. Specifically, from COVID-19 deaths, we learn that the current surveillance system can recover a substantial amount of missing data. For COVID-19 cases, there is still an opportunity to recover data. GDPH recently began an initiative to use drivers’ licenses to recover age, sex, and race information. GDPH’s efforts may explain why Georgia has less missing race and ethnicity data than the nationwide average. Still, given the lag time, there are opportunities to automate or speed up the process. Gold et al contest that automated workflows can reduce the COVID-19 reporting burden on data providers and increase timeliness of reporting.[1]

One recommendation for state and local health departments to speed up their race and ethnicity data recovery time, is to have a department or staff dedicated to recovery. Importantly, part of the responsibilities for this department should include travel to testing sites. While there is a lot that can be ascertained in hindsight using data analysis, there is no substitute for direct, upfront observations of data collection. Insights from site visits could then be used to improve data collection processes and minimize the amount of work needed from state and local health departments to keep up with high levels of missingness. That said, the high levels of data completion for age, sex, and address information demonstrate that it is possible to collect complete data during an active pandemic.

Another key finding was that for overall missingness, 3% of counties account for 40% of the missing race and ethnicity data (Table S1). Notably, these counties are some of Georgia's most populous counties and many are part of the Atlanta Metropolitan Area. From a public health standpoint, this information can be used to advocate for larger, better funded counties to mobilize resources to address data missingness. Targeting the counties with the greatest amounts of data gaps can lead to the largest and quickest improvements in data quality.

Future Research

Missing data present a unique challenge for researchers. Based on the literature, the best method for dealing with missing data depends on how the data is missing. From this analysis, we conclude that the data are missing at random as missing values are associated with observed information. In this situation, reweighted estimating equations are recommended, but multiple imputation may also be sufficient provided that survey weights are applied.[31] Future work may leverage reweighted estimating equations, multiple imputation, or specific algorithms, such as the Bayesian Indirect Surname Geocoding[32] to estimate racial disparities at the local, state, and national levels. These methods will be more adequate methods than complete case analysis. Additionally, future work can further unpack drivers of missing race and ethnicity data. GDPH actively recovers race and ethnicity information using drivers' licenses. A future analysis could examine the impact of missingness before and after drivers' license recovery was implemented.

References

1. Gold JAW, DeCuir J, Coyle JP, et al. COVID-19 Case Surveillance: Trends in Person-Level Case Data Completeness, United States, April 5–September 30, 2020. *Public Health Rep*. Published online March 31, 2021:00333549211006973. doi:10.1177/00333549211006973

2. Krieger N, Testa C, Hanage WP, Chen JT. US racial and ethnic data for COVID-19 cases: still missing in action. *The Lancet*. 2020;396(10261):e81. doi:10.1016/S0140-6736(20)32220-0

3. Centers for Disease Control and Prevention. COVID Data Tracker. Published March 28, 2020. Accessed April 20, 2021. https://covid.cdc.gov/covid-data-tracker

4. Spencer KD, Chung CL, Stargel AS, et al. COVID-19 Case Investigation and Contact Tracing Efforts from Health Departments — United States, June 25–July 24, 2020. *MMWR Morb Mortal Wkly Rep*. 2021;70(3):83-87. doi:10.15585/mmwr.mm7003a3

5. Moore JT, Ricaldi JN, Rose CE, et al. Disparities in Incidence of COVID-19 Among Underrepresented Racial/Ethnic Groups in Counties Identified as Hotspots During June 5–18, 2020 — 22 States, February–June 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69. doi:10.15585/mmwr.mm6933e1

6. Rossen LM. Excess Deaths Associated with COVID-19, by Age and Race and Ethnicity — United States, January 26–October 3, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69. doi:10.15585/mmwr.mm6942e2

7. U.S. Bureau of Labor Statistics. Labor force characteristics by race and ethnicity, 2019. Accessed April 23, 2021. https://www.bls.gov/opub/reports/race-and-ethnicity/2019/home.htm

8. U.S. Bureau of Labor Statistics. *Job Flexibilities and Work Schedules*.; 2019:1-32. Accessed April 23, 2021. https://www.bls.gov/news.release/flex2.toc.htm

9. Macias Gil R, Marcelin JR, Zuniga-Blanco B, Marquez C, Mathew T, Piggott DA. COVID-19 Pandemic: Disparate Health Impact on the Hispanic/Latinx Population in the United States. *J Infect Dis*. 2020;222(10):1592-1595. doi:10.1093/infdis/jiaa474

10. Billock RM, Groenewold MR, Free H, Sweeney MH, Luckhaupt SE. Required and Voluntary Occupational Use of Hazard Controls for COVID-19 Prevention in Non–Health Care Workplaces — United States, June 2020. *MMWR Morb Mortal Wkly Rep*. 2021;70. doi:10.15585/mmwr.mm7007a5

11. Dyal JW, Grant MP, Broadwater K, et al. COVID-19 Among Workers in Meat and Poultry Processing Facilities ― 19 States, April 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69. doi:10.15585/mmwr.mm6918e3

12. Grijalva CG. Transmission of SARS-COV-2 Infections in Households — Tennessee and Wisconsin, April–September 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69. doi:10.15585/mmwr.mm6944e1

13. Wang D, Gee GC, Bahiru E, Yang EH, Hsu JJ. Asian-Americans and Pacific Islanders in COVID-19: Emerging Disparities Amid Discrimination. *J GEN INTERN MED*. Published online October 2, 2020. doi:10.1007/s11606-020-06264-5

14. Millett GA, Jones AT, Benkeser D, et al. Assessing differential impacts of COVID-19 on black communities. *Annals of Epidemiology*. 2020;47:37-44. doi:10.1016/j.annepidem.2020.05.003

15. Selden TM, Berdahl TA. COVID-19 And Racial/Ethnic Disparities In Health Risk, Employment, And Household Composition. *Health Affairs*. 2020;39(9):1624-1632. doi:10.1377/hlthaff.2020.00897

16. Laurencin CT, McClinton A. The COVID-19 Pandemic: a Call to Action to Identify and Address Racial and Ethnic Disparities. *J Racial Ethn Health Disparities*. Published online April 18, 2020. doi:10.1007/s40615-020-00756-0

17. Labgold K, Hamid S, Shah S, et al. Estimating the Unknown: Greater Racial and Ethnic Disparities in COVID-19 Burden After Accounting for Missing Race and Ethnicity Data. *Epidemiology*. 2021;32(2):157-161. doi:10.1097/EDE.0000000000001314

18. Garg S, Kim L, Whitaker M, et al. Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 - COVID-NET, 14 States, March 1-30, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(15):458-464. doi:10.15585/mmwr.mm6915e3

19. Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur J Epidemiol*. 2019;34(1):23-36. doi:10.1007/s10654-018-0447-z

20. Cowger TL, Davis BA, Etkins OS, et al. Comparison of Weighted and Unweighted Population Data to Assess Inequities in Coronavirus Disease 2019 Deaths by Race/Ethnicity Reported by the US Centers for Disease Control and Prevention. *JAMA Netw Open*. 2020;3(7). doi:10.1001/jamanetworkopen.2020.16933

21. Jorden MA, Rudman SL, Villarino E, et al. Evidence for Limited Early Spread of COVID-19 Within the United States, January–February 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69. doi:10.15585/mmwr.mm6922e1

22. Georgia Department of Public Health. COVID-19 Status Report. Accessed May 2, 2021. https://dph.georgia.gov/covid-19-daily-status-report

23. Centers for Disease Control and Prevention. Coronavirus Disease 2019 (COVID-19): 2020 Interim Case Definition, Approved April 5, 2020. Accessed May 7, 2021. /nndss/conditions/coronavirus-disease-2019-covid-19/case-definition/2020/

24. Hubbard AE, Ahern J, Fleischer NL, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*. 2010;21(4):467-474. doi:10.1097/EDE.0b013e3181caeb90

25. Jun S-P, Yoo HS, Lee J-S. The impact of the pandemic declaration on public awareness and behavior: Focusing on COVID-19 google searches. *Technological Forecasting and Social Change*. 2021;166:120592. doi:10.1016/j.techfore.2021.120592

26. Sholle ET, Pinheiro LC, Adekkanattu P, et al. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *J Am Med Inform Assoc*. 2019;26(8-9):722-729. doi:10.1093/jamia/ocz040

27. Beck TL, Le T-K, Henry-Okafor Q, Shah MK. Medical Care for Undocumented Immigrants: National and International Issues. *Prim Care*. 2017;44(1):e1-e13. doi:10.1016/j.pop.2016.09.005

28. Warren RC, Hahn RA, Bristow L, Yu ES. The use of race and ethnicity in public health surveillance. *Public Health Rep*. 1994;109(1):4-6.

29. Sondik EJ, Lucas JW, Madans JH, Smith SS. Race/ethnicity and the 2000 census: implications for public health. *Am J Public Health*. 2000;90(11):1709-1713.

30. Long JA, Bamba MI, Ling B, Shea JA. Missing Race/Ethnicity Data in Veterans Health Administration Based Disparities Research: A Systematic Review. *Journal of Health Care for the Poor and Underserved*. 2006;17(1):128-140. doi:10.1353/hpu.2006.0029

31. Henry AJ, Hevelone ND, Lipsitz S, Nguyen LL. Comparative methods for handling missing data in large databases. *Journal of Vascular Surgery*. 2013;58(5):1353-1359.e6. doi:10.1016/j.jvs.2013.05.008

32. Fremont A, Weissman JS, Hoch E, Elliott MN. When Race/Ethnicity Data Are Lacking. *Rand Health Q*. 2016;6(1). Accessed November 19, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5158280/

Tables

Table 1: Demographic Data Missingness Status Among Confirmed COVID-19 Cases and Deaths Occurring Between April 2020 - March 2021 in Georgia

	Cases (N=840424)	**Deaths (N=16286)**
Missing Race		
No	684526 (81.5)	16150 (99.1)
Yes	155898 (18.6)	136 (0.8)
Missing Ethnicity		
No	608615 (72.4)	16128 (99.0)
Yes	231809 (27.6)	158 (1.0)
Missing Both Race and Ethnicity		
No	706284 (84.0)	16230 (99.7)
Yes	134140 (16.0)	56 (0.3)
Missing Age		
No	834645 (99.3)	16285 (99.99)
Yes	5779 (0.7)	1 (0.01)
Missing Sex		
No	831040 (98.9)	16270 (99.9)
Yes	9384 (1.1)	16 (0.1)
Missing Zip Code		
No	820876 (97.7)	16159 (99.2)
Yes	19548 (2.3)	127 (0.8)
Missing County of Residence		
No	823620 (98.0)	16243 (99.7)
Yes	16804 (2.0)	43 (0.26)

Table 2a: Age and Sex Characteristics of Confirmed Cases With and Without Missing Race and Ethnicity Information (Individual-level analysis)

	Missing Race	**Reported Race**		**Missing Ethnicity**	**Reported Ethnicity**		**Overall**
Age							
Mean (SD)	37.9 (18.7)	41.7 (20.2)		38.9 (19.1)	41.7 (20.3)		
Age			**p < .0001**			**p < .0001**	
0-17	21,687 (23.4)	71,176 (76.7)		29,564 (31.8)	63,299 (68.2)		92,863 (11.1)
18-64	118,038 (18.8)	510,697 (81.2)		176,124 (28.0)	452,611 (72.0)		628,735 (75.3)
65+	13,595 (12.0)	99,452 (88.0)		23,305 (20.6)	89,742 (79.4)		113,047 (13.5)
Sex			**p < .0001**			**p < .0001**	
Male	74,746 (19.5)	309,126 (80.5)		110,520 (28.8)	273,352 (71.2)		383,872 (46.2)
Female	74,386 (16.6)	372,782 (83.4)		113,655 (25.4)	333,513 (74.6)		447,168 (53.8)

Table 2b: Age and Sex Characteristics of Confirmed Deaths With and Without Missing Race and Ethnicity Information (Individual-level analysis)

	Missing Race	**Reported Race**		**Missing Ethnicity**	**Reported Ethnicity**		**Overall**
Age							
Mean (SD)	67.6 (16.9)	73.8 (13.9)		66.1 (17.9)	73.9 (13.9)		
Age			**p = .0011**			**p = .0002**	
0-17	0 (0.0)	10 (100.0)		0 (0.0)	10 (100.0)		10 (0.06)
18-64	49 (1.3)	3,698 (98.7)		58 (1.6)	3689 (98.5)		3747 (23.0)
65+	86 (0.7)	12,442 (99.3)		100 (0.8)	12428 (99.2)		12,528 (76.9)
Sex			p = 0.6393			p = 0.2515	
Men	73 (0.8)	8,595 (99.2)		88 (1.0)	8,580 (99.0)		8,668 (53.3)
Women	59 (0.8)	7,543 (99.2)		64 (0.8)	7,538 (99.2)		7602 (46.7)

Table 3: Adjusted Associations of Demographic and Social Variables with Missing Race and Ethnicity (n = 804,400)

	Missing Race	Missing Ethnicity
	OR (95% CI)	**OR (95% CI)**
Age and Sex		
Male, 0-17	2.50 (2.05, 3.04)	1.90 (1.62, 2.23)
Male, 18-64	1.69 (1.49, 1.92)	1.47 (1.33, 1.62)
Male, 65+	1.22 (1.09, 1.37)	1.15 (1.05, 1.26)
Female, 0-17	2.18 (1.90, 2.51)	1.68 (1.49, 1.89)
Female, 18-64	1.48 (1.38, 1.59)	1.29 (1.22, 1.38)
Female, 65+	ref	ref
% Population Reporting Black Race		
1st Quintile	ref	ref
2nd Quintile	0.94 (0.85, 1.04)	1.02 (0.92, 1.12)
3rd Quintile	0.89 (0.73, 1.09)	1.04 (0.86, 1.25)
4th Quintile	0.84 (0.62, 1.13)	1.05 (0.79, 1.40)
5th Quintile	0.79 (0.53, 1.18)	1.07 (0.73, 1.57)
% Population Reporting Hispanic Ethnicity		
1st Quintile	ref	ref
2nd Quintile	1.02 (0.88, 1.17)	1.02 (0.83, 1.25)
3rd Quintile	1.03 (0.78, 1.37)	1.04 (0.69, 1.57)
4th Quintile	1.05 (0.69, 1.60)	1.06 (0.57, 1.96)
5th Quintile	1.06 (0.60, 1.87)	1.08 (0.47, 2.45)
% Population Reporting Multiple Race		
1st Quintile	ref	ref
2nd Quintile	1.01 (0.94, 1.09)	1.04 (0.95, 1.15)
3rd Quintile	1.03 (0.88, 1.20)	1.09 (0.90, 1.32)
4th Quintile	1.04 (0.83, 1.31)	1.13 (0.85, 1.51)
5th Quintile	1.06 (0.78, 1.43)	1.18 (0.81, 1.73)
% Population Speaking English Less Than Well		
1st Quintile	ref	ref
2nd Quintile	1.15 (0.99, 1.33)	1.10 (0.88, 1.39)
3rd Quintile	1.31 (0.97, 1.77)	1.22 (0.77, 1.94)
4th Quintile	1.50 (0.96, 2.35)	1.35 (0.67, 2.71)
5th Quintile	1.72 (0.95, 3.12)	1.49 (0.59, 3.78)

% Population in Poverty		
1st Quintile	ref	ref
2nd Quintile	1.02 (0.93, 1.12)	1.00 (0.87, 1.39)
3rd Quintile	1.04 (0.86, 1.25)	0.99 (0.77, 1.95)
4th Quintile	1.06 (0.79, 1.41)	0.99 (0.67, 2.71)
5th Quintile	1.07 (0.73, 1.57)	0.99 (0.59, 3.79)
% Population Uninsured		
1st Quintile	ref	ref
2nd Quintile	0.97 (0.89, 1.05)	0.89 (0.79, 1.00)
3rd Quintile	0.94 (0.80, 1.11)	0.79 (0.63, 0.99)
4th Quintile	0.91 (0.71, 1.17)	0.70 (0.50, 0.99)
5th Quintile	0.88 (0.63, 1.23)	0.63 (0.40, 0.99)
Case Rate		
1st Quintile	ref	ref
2nd Quintile	1.02 (0.89, 1.05)	1.07 (0.99, 1.15)
3rd Quintile	1.05 (0.80, 1.11)	1.14 (0.98, 1.33)
4th Quintile	1.07 (0.71, 1.17)	1.22 (0.97, 1.54)
5th Quintile	1.10 (0.63, 1.23)	1.31 (0.96, 1.78)

Table 4: Unadjusted and Adjusted County-level Associations of Social Variables with Percentage of Missing Race and Ethnicity (n=159 GA counties)

	% Cases Missing Race		**% Cases Missing Ethnicity**	
	Unadjusted Model	**Adjusted Model***	**Unadjusted Model**	**Adjusted Model***
County Characteristics	**B (95% CI)**	**B (95% CI)**	**B (95% CI)**	**B (95% CI)**
% Population Reporting Black Race				
1st Quintile (lowest)	ref	ref	ref	ref
2nd Quintile	-0.93 (-1.62, -0.25)	-0.30 (-1.12, 0.53)	-0.95 (-1.86, -0.030)	0.56 (-0.48, 1.61)
3rd Quintile	-1.87 (-3.24, -0.50)	-0.60 (-2.25, 1.05)	-1.89 (-3.72, -0.06)	1.13 (-0.96, 3.21)
4th Quintile	-2.80 (-4.86, -0.74)	-0.90 (-3.37, 1.58)	-2.84 (-5.58, -0.089)	1.69 (-1.44, 4.82)
5th Quintile (highest)	-3.73 (-6.48, -0.99)	-1.20 (-4.50, 2.10)	-3.78 (-7.45, -0.12)	2.25 (-1.92, 6.43)
% Population Reporting Hispanic Ethnicity				
1st Quintile (lowest)	ref	ref	ref	ref
2nd Quintile	1.26 (0.59, 1.93)	0.46 (-0.63, 1.55)	1.21 (0.30, 2.12)	1.20 (-0.17, 2.58)
3rd Quintile	2.52 (1.17, 3.86)	0.92 (-1.25, 3.10)	2.42 (0.61, 4.24)	2.41 (-0.34, 5.16)
4th Quintile	3.78 (1.76, 5.79)	1.39 (-1.88, 4.65)	3.63 (0.91, 6.36)	3.61 (-0.51, 7.74)
5th Quintile (highest)	5.03 (2.34, 7.73)	1.84 (-2.51, 6.19)	4.84 (1.21, 8.48)	4.82 (-0.68, 10.31)
% Population Reporting Multiple Race				
1st Quintile (lowest)	ref	ref	ref	ref
2nd Quintile	1.15 (0.47, 1.83)	**0.76 (0.026, 1.49)**	1.94 (1.07, 2.82)	**1.18 (0.25, 2.10)**
3rd Quintile	2.30 (0.95, 3.66)	**1.52 (0.053, 2.99)**	3.89 (2.14, 5.64)	**2.35 (0.50, 4.20)**
4th Quintile	3.45 (1.42, 5.49)	**2.28 (0.079, 4.48)**	5.83 (3.20, 8.46)	**3.53 (0.74, 6.31)**
5th Quintile (highest)	4.60 (1.89, 7.31)	**3.04 (0.11, 5.97)**	7.78 (4.27, 11.28)	**4.70 (0.99, 8.41)**
% Population Speaking English Less Than Well				
1st Quintile (lowest)	ref	ref	ref	ref
2nd Quintile	1.15 (0.47, 1.82)	0.18 (-0.91, 1.26)	0.59 (-0.33, 1.51)	-0.93 (-2.30, 0.45)
3rd Quintile	2.29 (0.94, 3.65)	0.35 (-1.82, 2.53)	1.18 (-0.67, 3.02)	-1.85 (-4.60, 0.90)
4th Quintile	3.44 (1.41, 5.47)	0.53 (-2.73, 3.79)	1.77 (-1.00, 4.53)	-2.78 (-6.90, 1.34)
5th Quintile (highest)	4.59 (1.88, 7.30)	0.71 (-3.64, 5.05)	2.35 (-1.34, 6.04)	-3.70 (-9.20, 1.79)
% Population in Poverty				
1st Quintile (lowest)	ref	ref	ref	ref
2nd Quintile	-0.76 (-1.45, -0.067)	-0.13 (-1.08, 0.82)	-1.93 (-2.80, -1.05)	-1.10 (-2.30, 0.11)
3rd Quintile	-1.52 (-2.90, -0.13)	-0.25 (-2.15, 1.65)	-3.85 (-5.60, -2.10)	-2.19 (-4.59, 0.21)
4th Quintile	-2.28 (-4.35, -0.20)	-0.38 (-3.23, 2.47)	-5.78 (-8.41, -3.15)	-3.29 (-6.89, 0.32)
5th Quintile (highest)	-3.03 (-5.80, -0.27)	-0.51 (-4.31, 3.30)	-7.70 (-11.21, -4.19)	-4.38 (-9.19, 0.42)
% Population Uninsured				
1st Quintile (lowest)	ref	ref	ref	ref
2nd Quintile	0.027 (-0.67, 0.73)	0.062 (-0.75, 0.88)	-1.49 (-2.39, -0.59)	-0.94 (-1.97, 0.090)
3rd Quintile	0.054 (-1.35, 1.46)	0.12 (-1.50, 1.75)	-2.98 (-4.78, -1.19)	-1.88 (-3.93, 0.18)
4th Quintile	0.081 (-2.02, 2.19)	0.19 (-2.26, 2.63)	-4.48 (-7.17, -1.78)	-2.81 (-5.90, 0.27)
5th Quintile (highest)	0.11 (-2.70, 2.91)	0.25 (-3.01, 3.50)	-5.97 (-9.56, -2.38)	-3.75 (-7.87, 0.36)
Case Rate				
1st Quintile (lowest)	ref	ref	ref	ref

2nd Quintile	1.39 (0.73, 2.06)	**0.99 (0.27, 1.72)**	1.57 (0.67, 2.46)	**1.47 (0.55, 2.38)**
3rd Quintile	2.79 (1.46, 4.12)	**1.99 (0.54, 3.43)**	3.13 (1.34, 4.92)	**2.93 (1.11, 4.76)**
4th Quintile	4.18 (2.19, 6.18)	**2.98 (0.81, 5.15)**	4.70 (2.01, 7.38)	**4.40 (1.66, 7.13)**
5th Quintile (highest)	5.58 (2.91, 8.25)	**3.97 (1.08, 6.86)**	6.26 (2.68, 9.84)	**5.86 (2.21, 9.52)**

*Adjusted models account for all county characteristics shown in the table

Figures

Figure 1. Missing Data Among COVID-19 Cases for Race and Ethnicity from April 2020 to March 2021, in Georgia

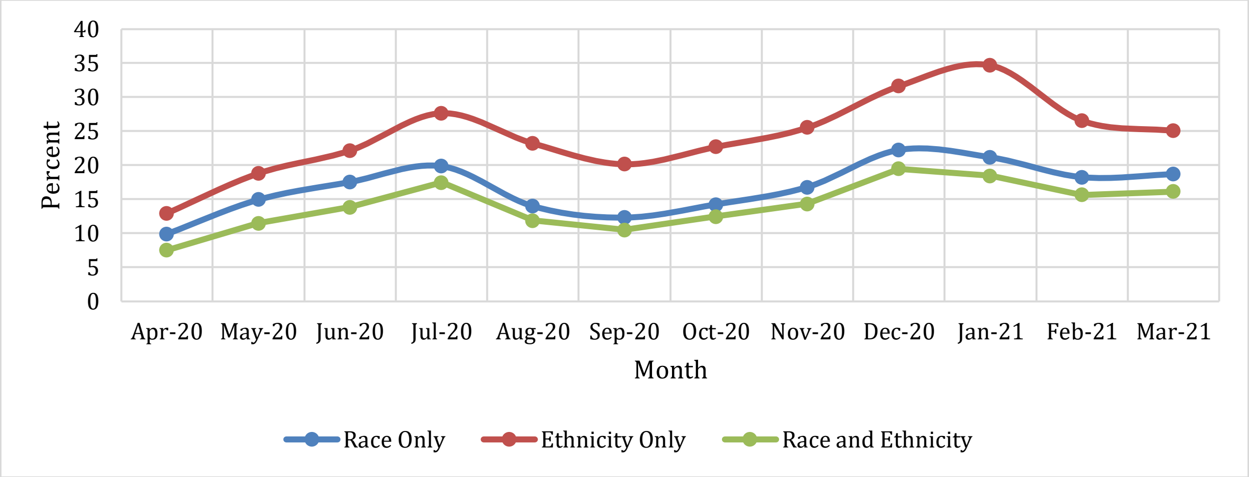


Figure 2. Missing Data Among COVID-19 Deaths for Race and Ethnicity from April 2020 to March 2021, in Georgia

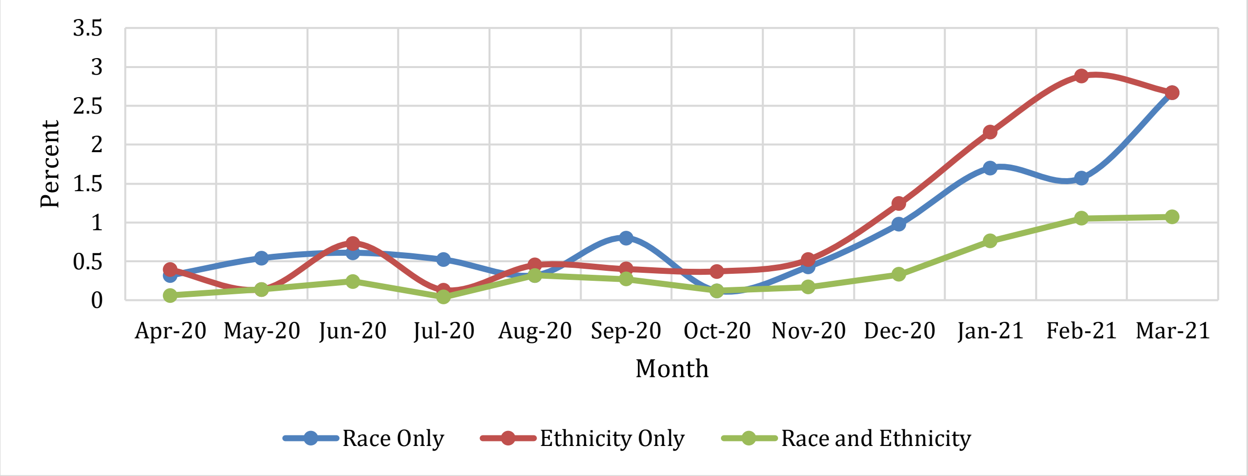


Figure 3. Percent of Total Missing Demographic Variables by County of Residence from April 2020 to March 2021

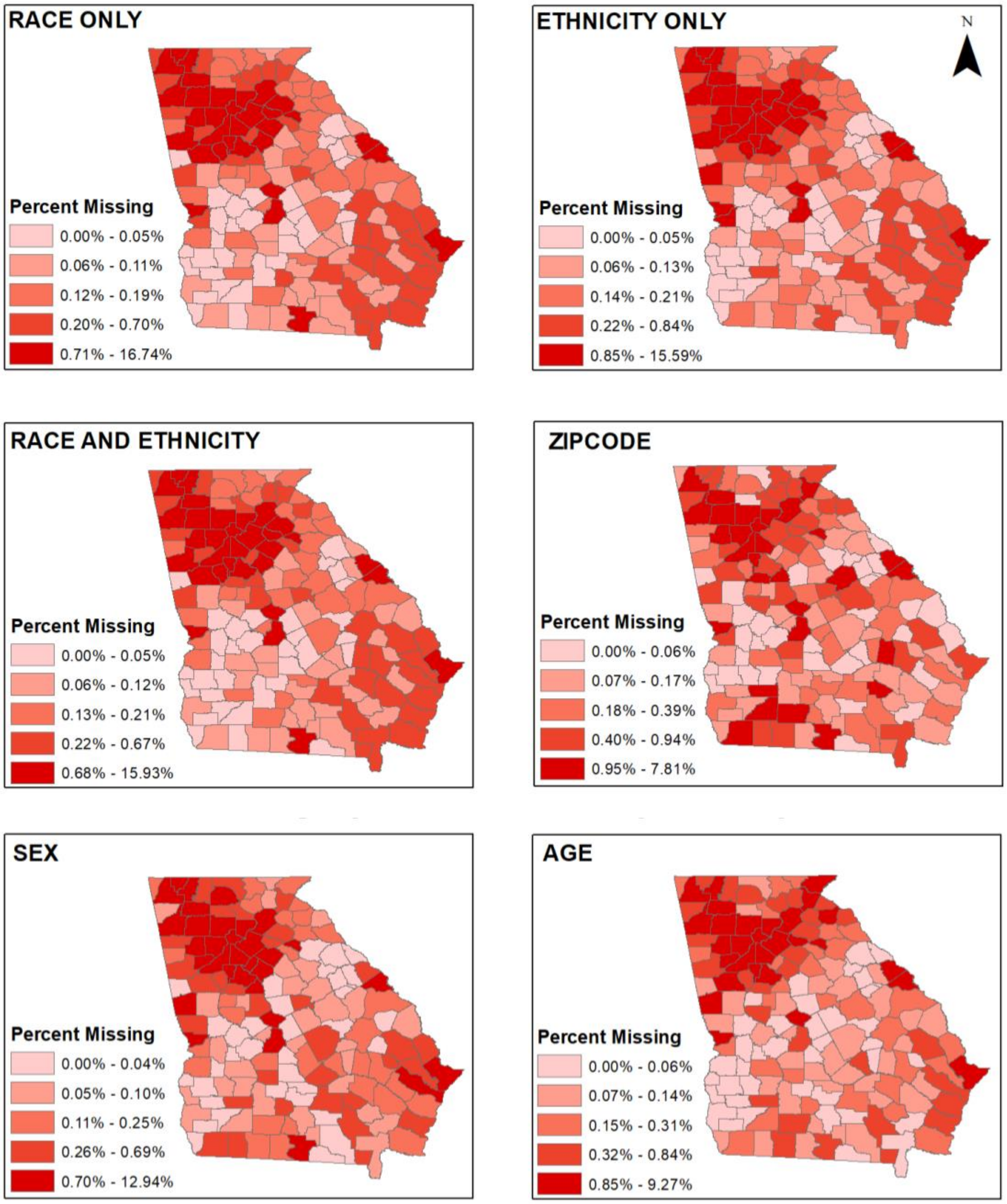
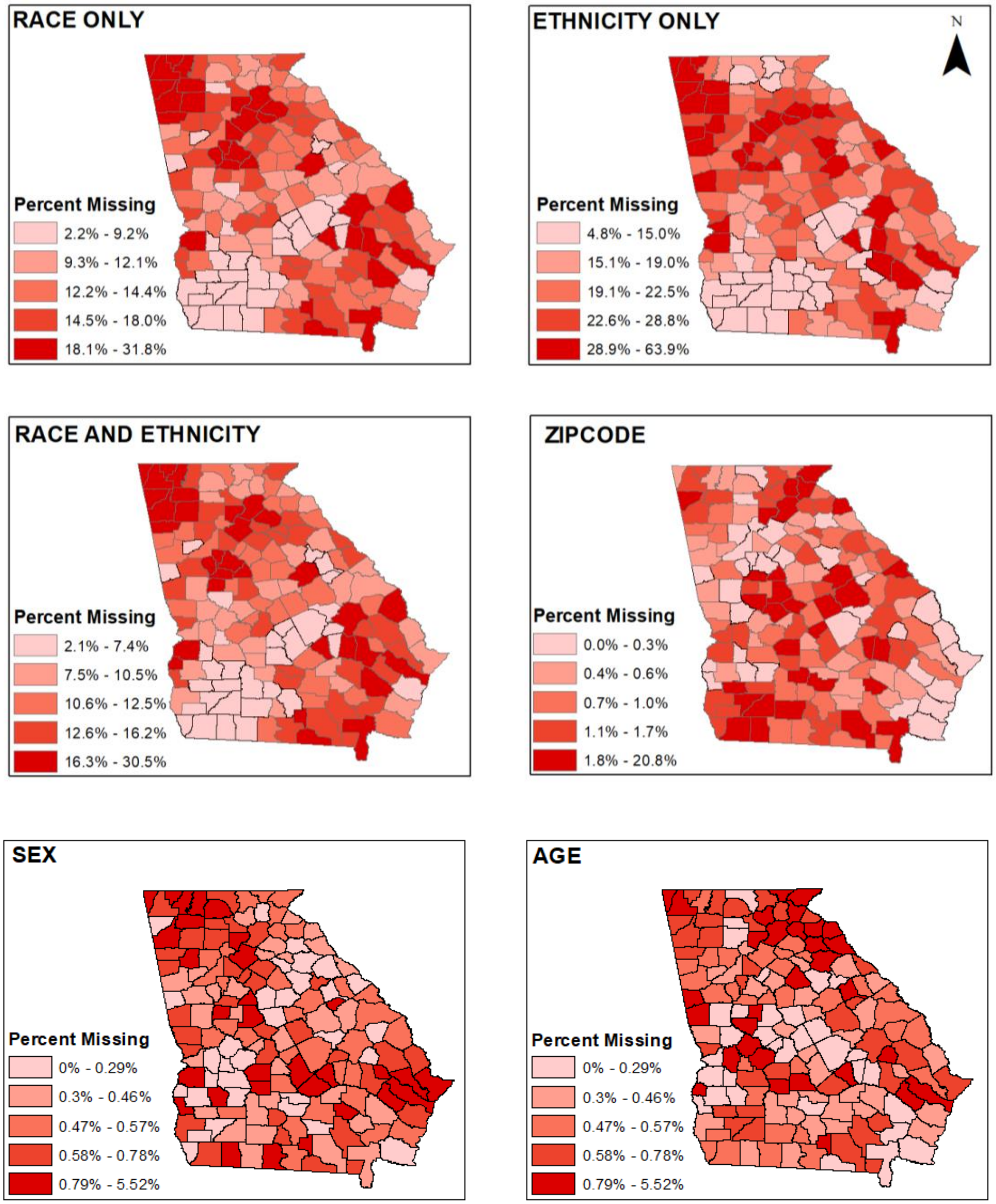


Figure 4. Percentage of Missing Demographic Variables by County of Residence from April 2020 to March 2021



Supplementary Tables

Table S1: Percentage and Percent of Total Missing Race and Missing Ethnicity by County from April 2020 to March 2021

	Percentage of Missing Race	Percentage of Missing Ethnicity	Total Percent Missing Race	Total Percent Missing Ethnicity
Counties				
Appling	18.0	28.8	0.24	0.25
Atkinson	14.8	16.5	0.08	0.06
Bacon	15.9	12.9	0.14	0.08
Baker	6.6	10.5	0.01	0.01
Baldwin	11.7	26.2	0.31	0.46
Banks	13.6	25.9	0.15	0.19
Barrow	20.0	30.8	1.19	1.20
Bartow	22.3	34.5	1.65	1.68
Ben Hill	16.1	19.4	0.17	0.13
Berrien	10.0	11.5	0.07	0.06
Bibb	12.5	19.7	1.15	1.18
Bleckley	4.0	9.8	0.02	0.04
Brantley	13.5	18.2	0.09	0.08
Brooks	13.4	20.9	0.09	0.09
Bryan	14.6	22.4	0.27	0.27
Bulloch	16.4	20.5	0.60	0.49
Burke	10.1	23.2	0.12	0.18
Butts	21.2	38.6	0.32	0.38
Calhoun	7.0	14.9	0.02	0.03
Camden	12.5	17.1	0.28	0.25
Candler	16.1	17.2	0.08	0.06
Carroll	14.1	34.0	0.71	1.12
Catoosa	29.0	37.0	1.06	0.89
Charlton	28.5	32.7	0.21	0.16
Chatham	11.0	19.5	1.53	1.78
Chattahoochee	9.2	63.9	0.20	0.93
Chattooga	24.9	33.4	0.37	0.33
Cherokee	16.0	21.9	2.43	2.18
Clarke	15.6	23.6	1.36	1.36
Clay	15.7	17.5	0.02	0.01
Clayton	23.6	33.7	3.69	3.45
Clinch	14.6	20.7	0.07	0.07
Cobb	17.6	25.7	7.21	6.88

Coffee	13.9	19.9	0.41	0.39
Colquitt	6.8	10.9	0.16	0.17
Columbia	14.2	30.1	1.07	1.50
Cook	10.8	12.2	0.09	0.06
Coweta	17.3	21.5	1.02	0.84
Crawford	11.6	22.9	0.04	0.06
Crisp	11.7	27.7	0.11	0.17
Dade	27.6	33.4	0.22	0.17
Dawson	16.4	19.6	0.30	0.23
Decatur	6.9	13.8	0.10	0.13
Dekalb	18.0	29.9	6.94	7.56
Dodge	4.3	11.0	0.03	0.05
Dooly	12.0	21.3	0.06	0.07
Dougherty	4.3	12.8	0.14	0.27
Douglas	8.7	25.5	0.70	1.34
Early	8.6	12.4	0.06	0.05
Echols	21.8	18.7	0.05	0.03
Effingham	10.5	17.0	0.27	0.29
Elbert	13.2	23.7	0.14	0.17
Emanuel	25.4	29.2	0.30	0.23
Evans	13.6	15.5	0.07	0.05
Fannin	12.6	17.5	0.18	0.17
Fayette	23.5	29.6	1.06	0.87
Floyd	25.4	34.3	1.72	1.52
Forsyth	21.9	28.6	2.57	2.21
Franklin	12.4	19.9	0.19	0.20
Fulton	12.3	21.4	6.77	7.75
Gilmer	9.7	13.4	0.16	0.15
Glascock	9.1	25.0	0.01	0.02
Glynn	10.1	13.1	0.46	0.39
Gordon	25.0	32.4	1.08	0.92
Grady	5.1	9.6	0.05	0.07
Greene	12.7	28.8	0.13	0.19
Gwinnett	28.4	40.3	16.74	15.59
Habersham	14.4	22.3	0.46	0.46
Hall	20.6	23.8	3.46	2.62
Hancock	23.0	33.3	0.13	0.13
Haralson	11.6	36.0	0.14	0.28
Harris	13.3	20.4	0.19	0.19
Hart	10.7	17.3	0.12	0.13

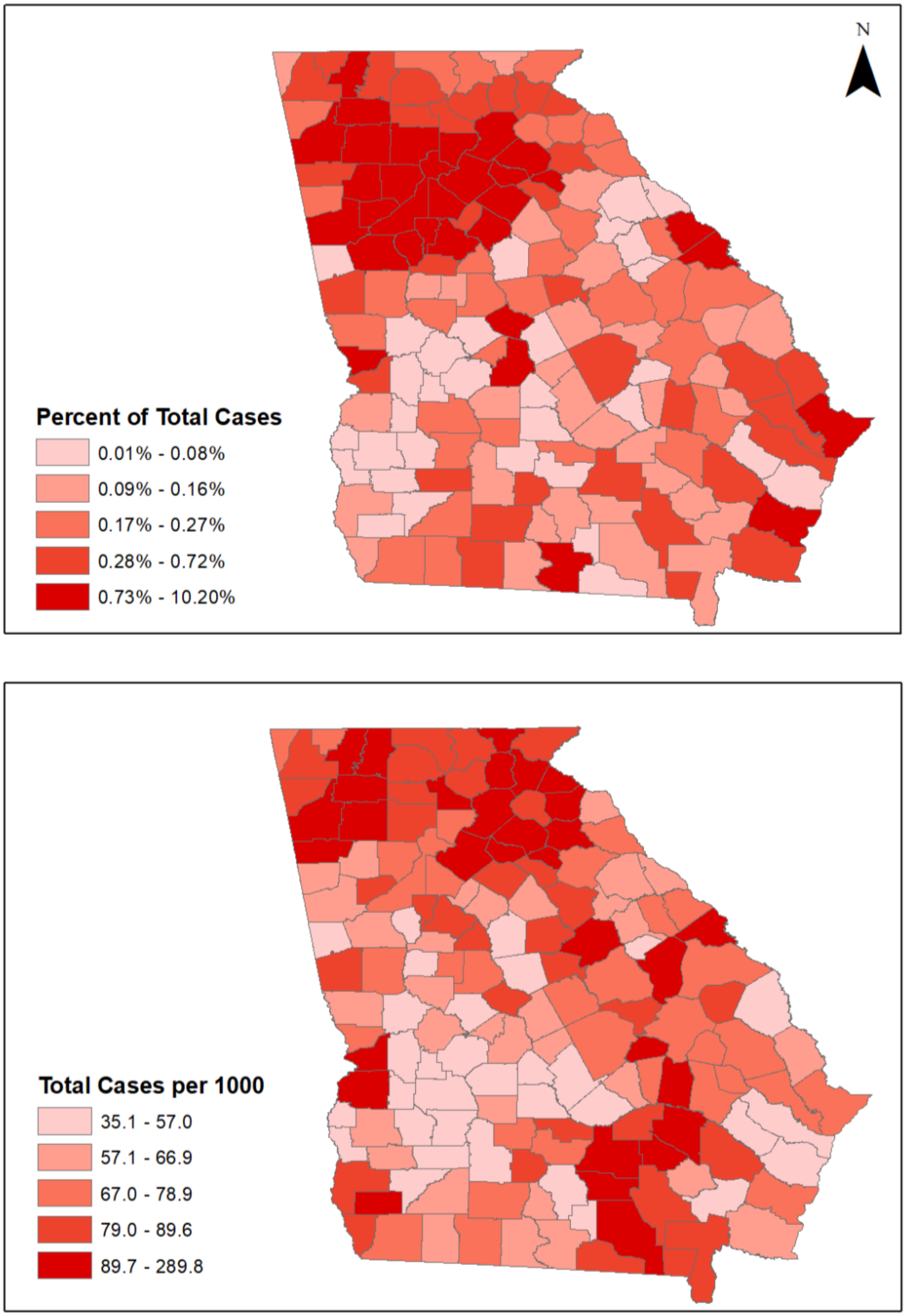
Heard	7.5	19.8	0.03	0.06
Henry	20.6	28.4	2.64	2.40
Houston	17.2	25.9	1.18	1.16
Irwin	15.1	14.9	0.07	0.05
Jackson	20.5	33.6	1.20	1.29
Jasper	12.5	18.7	0.06	0.06
Jeff Davis	9.8	14.2	0.09	0.08
Jefferson	11.1	17.9	0.12	**0.13**
Jenkins	13.8	19.2	0.07	**0.06**
Johnson	4.1	14.9	0.02	**0.05**
Jones	10.5	19.4	0.11	**0.14**
Lamar	13.5	24.2	0.12	0.14
Lanier	23.1	25.2	0.08	0.06
Laurens	4.8	11.1	0.12	0.19
Lee	5.7	15.3	0.06	0.10
Liberty	22.5	37.5	0.52	0.57
Lincoln	16.1	24.5	0.05	0.05
Long	14.2	22.1	0.07	0.07
Lowndes	15.9	18.8	0.84	0.65
Lumpkin	11.5	12.7	0.22	0.16
Mcduffie	13.6	21.8	0.06	0.06
Mcintosh	7.8	12.4	0.29	0.38
Macon	15.1	20.3	0.04	0.03
Madison	15.6	30.8	0.14	0.15
Marion	13.3	16.5	0.04	0.04
Meriwether	9.8	22.5	0.10	0.15
Miller	4.4	5.3	0.02	0.02
Mitchell	4.2	11.8	0.04	0.08
Monroe	14.8	21.5	0.19	0.18
Montgomery	7.1	12.6	0.04	0.04
Morgan	13.4	25.8	0.11	0.14
Murray	15.2	17.8	0.43	0.33
Muscogee	17.4	23.7	1.63	1.46
Newton	14.5	23.2	0.73	0.77
Oconee	14.3	24.7	0.30	0.34
Oglethorpe	15.3	27.8	0.12	0.15
Paulding	15.6	35.0	1.14	1.68
Peach	14.4	21.0	0.18	0.17
Pickens	9.0	18.5	0.16	0.21
Pierce	12.6	18.9	0.11	0.10

Pike	18.0	25.0	0.13	0.12
Polk	21.4	26.6	0.58	0.47
Pulaski	5.9	19.7	0.02	0.05
Putnam	12.5	26.5	0.15	0.21
Quitman	16.3	21.3	0.01	0.01
Rabun	15.9	19.9	0.16	0.13
Randolph	13.2	17.2	0.04	0.03
Richmond	11.3	22.1	1.52	1.96
Rockdale	14.4	22.0	0.58	0.58
Schley	8.5	12.5	0.01	0.01
Screven	20.8	24.1	0.12	0.09
Seminole	2.2	4.8	0.01	0.02
Spalding	20.6	32.9	0.56	0.58
Stephens	9.4	16.9	0.19	0.23
Stewart	31.8	28.9	0.17	0.10
Sumter	11.6	18.0	0.14	0.14
Talbot	10.5	16.1	0.03	0.03
Taliaferro	5.2	16.7	0.00	0.01
Tattnall	24.0	28.8	0.31	0.24
Taylor	10.5	15.0	0.04	0.03
Telfair	11.9	18.4	0.06	0.06
Terrell	2.5	8.2	0.01	0.02
Thomas	3.7	10.4	0.09	0.17
Tift	9.1	9.1	0.21	0.14
Toombs	20.0	29.1	0.41	0.39
Towns	13.1	16.3	0.10	0.08
Treutlen	7.4	15.0	0.03	0.04
Troup	14.8	31.4	0.60	0.84
Turner	11.0	13.0	0.05	0.04
Twiggs	12.0	20.4	0.04	0.05
Union	10.6	14.4	0.15	0.13
Upson	7.0	16.7	0.09	0.14
Walker	30.9	39.3	1.29	1.08
Walton	17.6	28.1	0.97	1.02
Ware	14.3	22.8	0.30	0.31
Warren	9.6	17.3	0.02	0.03
Washington	11.4	20.5	0.13	0.15
Wayne	27.0	31.1	0.51	0.39
Webster	5.3	6.3	0.00	0.00
Wheeler	31.8	35.8	0.10	0.08

White	12.0	15.9	0.24	0.21
Whitfield	22.0	21.6	2.27	1.46
Wilcox	6.6	16.1	0.02	0.04
Wilkes	10.2	15.6	0.05	0.05
Wilkinson	12.6	23.2	0.06	0.08
Worth	5.1	10.9	0.04	0.06

Supplementary Figures

Figure S1. Total Cases and Case Rates of COVID-19 by County of Residence from April 2020 to March 2021 in Georgia


N
Percent of Total Cases
0.01% - 0.08%
0.09% - 0.16%
0.17% - 0.27%
0.28% - 0.72%
0.73% - 10.20%
Total Cases per 1000
35.1 - 57.0
57.1 - 66.9
67.0 - 78.9
79.0 - 89.6
89.7 - 289.8