**Manuscript Type:** Brief Report

**Title** Widening the gap: greater racial and ethnic disparities in COVID-19 burden after accounting for missing race/ethnicity data

**Authors:** Katie Labgold*[1] & Sarah Hamid*,[1] Sarita Shah, [1,2,3] Neel R. Gandhi,[1,2,3] Allison Chamberlain,[1] Fazle Khan,[4] Shamimul Khan, [4] Sasha Smith, [4] Steve Williams, [4] Timothy L. Lash,[1] Lindsay J. Collin[1,5]

*Co-First Authors

**Affiliations:** [1]Department of Epidemiology, Rollins School of Public Health, Emory University; [2]Department of Global Health, Rollins School of Public Health, Emory University; [3]Division of Infectious Diseases, Emory School of Medicine, Emory University, [4]Fulton County Board of Health; [5]Department of Population Health Sciences, Huntsman Cancer Institute, University of Utah;

**Corresponding Author**: Lindsay J Collin, Department of Population Health Sciences, Huntsman Cancer Institute, University of Utah, 2000 Circle of Hope Drive, Room 4746, Salt Lake City UT, 84112; email: lindsay.collin@hci.utah.edu

**Running head**: Accounting for missing race/ethnicity COVID-19

**Conflicts of Interest**: The authors have no conflicts of interest to declare.

**Financial Support**: This work was supported in part by the US National Institutes of Health F31CA239566 (PI L. J. Collin), R01LM013049 (PI T. L. Lash), and K24AI114444 (PI N. R. Gandhi). It was also supported by a grant from the Robert W. Woodruff foundation (PI A. Chamberlain). K. Labgold is supported in part by the Center for Reproductive Health Research in the Southeast (RISE) Doctoral Fellowship and an ARCS Foundation Award.

**Data Access:** Due to patient confidentiality, data are only available upon request from the Fulton County Board of Health and with IRB approval from the Georgia Department of Public Health. Example code used to perform the imputation and bias-adjustment is available on GitHub.

**Abstract (147/150)**

Black, Hispanic, and Indigenous persons in the United States have an increased risk of SARS-CoV-2 infection and death from COVID-19, due to persistent social inequities. The magnitude of the disparity is unclear, however, because race/ethnicity information is often missing in surveillance data. In this study, we quantified the burden of SARS-CoV-2 infection, hospitalization, and case fatality rates in an urban county by racial/ethnic group using combined race/ethnicity imputation and quantitative bias-adjustment for misclassification. After bias-adjustment, the magnitude of the absolute racial/ethnic disparity, measured as the difference in infection rates between classified Black and Hispanic persons compared to classified White persons, increased 1.3-fold and 1.6-fold respectively. These results highlight that complete case analyses may underestimate absolute disparities in infection rates. Collecting race/ethnicity information at time of testing is optimal. However, when data are missing, combined imputation and bias-adjustment improves estimates of the racial/ethnic disparities in the COVID-19 burden.

**Keywords:** SARS-CoV-2, COVID-19, missing data, bias analysis, race/ethnicity disparities, surveillance

## Introduction

In the United States, early surveillance reports highlight that persons of Hispanic, Black, and American Indian/Alaskan Native race and ethnicity are disproportionately affected by the COVID-19 pandemic.[1] These disparities arise from historical and contemporary social and health inequities that result from systemic racism.[2–4] Racial capitalism in particular produces structurally unequal exposure to (and protection from) SARS-CoV-2 infection in key places of transmission (*e.g.* workplace).[3]

The role of systemic racism in the pandemic motivates the need for accurate surveillance of racial/ethnic disparities in SARS-CoV-2 infection and death. However, there are challenges in estimating COVID-19 racial/ethnic disparities.[5,6] Although reports highlight the unequal burden across racial/ethnic groups, the magnitude of disparities is uncertain because of the large proportion of missing race/ethnicity information in surveillance data. In recent reports, race/ethnicity information was missing in 56% of confirmed infections nationally and in 36% in Georgia.[7,8] Current surveillance estimates are reported as complete case analyses, which exclude cases with missing race/ethnicity.[1,5,8,9] Complete case analyses will bias racial/ethnic disparity estimates if race/ethnicity information is not missing completely at random.[10]

The Department of Health and Human Services issued COVID-19 reporting guidelines in June requiring all labs to report race/ethnicity beginning August 2020.[11] These guidelines seek to address missing data moving forward, but fail to address missing information for case-patients identified before August.

Collecting race/ethnicity information at time of testing is optimal, especially in surveillance of racial/ethnic health disparities. Until this becomes routine, imputation of missing race/ethnicity combined with quantitative bias-adjustment to account for misclassification of the imputed race/ethnicity can improve estimates of the COVID-19 burden among racial/ethnic groups when

race/ethnicity data are missing.[12] In this study, we calculate SARS-CoV-2 infection, hospitalizations, and case fatality rates by race/ethnicity group and report the absolute racial/ethnic disparities in SARS-CoV-2 infection rates in Fulton County, Georgia after accounting for missing race/ethnicity information.

## Methods

Fulton County, Georgia, includes the city of Atlanta and residents identify as Black (44%), White (40%), Hispanic (7%), Asian (7%), and other races/ethnicities (2%).[13] Between 29 February 2020 and 18 Aug 2020, 19,637 cases of SARS-CoV-2 infection were reported among Fulton County residents. Case reports included the patients' residential address, full name, race/ethnicity, hospitalization (yes/no/unknown), and death (yes/no/unknown). Fulton County Board of Health staff geocoded case-patients' address to census block groups. For this analysis, we categorized reported race/ethnicity as Black, Hispanic, Asian, White or Other.

We accounted for missing race/ethnicity information using a three-step approach: 1) imputation of race/ethnicity for all case-patients, 2) validation of the race/ethnicity imputation by calculating the accuracy of imputation among case-patients with reported race/ethnicity information, and 3) bias-adjustment of race/ethnicity estimates to account for misclassification of imputation among case-patients missing reported race/ethnicity information. Hereafter, we refer to race/ethnicity as reported when provided in case-patient records, *imputed* when referring to the imputed case-patient race/ethnicity, and *classified* when referring to the combined reported and imputed race/ethnicity after bias-adjustment.

First, for all case-patients we predicted their racial/ethnic group using the Bayesian Improved Surname Geocoding method.[14] This method estimates the probability of a person being classified as Black, Hispanic, Asian, White or Other race/ethnic group based on the case-

patient's surname and residential census block group, and the population distribution of race/ethnicity for census block groups and surnames. Imputation was performed using the R package "wru," which includes the 2010 surname census list with corresponding race/ethnicity distribution. Geographic distribution of race/ethnicity came from the 2018 5-year American Community Survey.[15,16] For the 546 (2.8%) case-patients who could not be geocoded, race/ethnicity was imputed using surname only.

Second, we validated the race/ethnicity imputation among case-patients whose race/ethnicity was available in the dataset (n=12,222, 64%). Predictive values (PV) were calculated for each imputed race/ethnic group. The PV is the probability that a person's reported race/ethnicity group classification was correctly imputed.[12]

Third, we used the PV values as bias parameters to quantitatively adjust for the expected misclassification of the imputed race/ethnicity groups. We assigned each race/ethnicity group PV from the validation to a Dirichlet distribution (**Table 1**). We then reclassified the imputed race/ethnicity probabilistically (100,000 iterations).[12] The quantitative bias-adjustment mathematically accounts for inaccurate assignment of case-patients to a race/ethnicity group by the Bayesian Improved Surname Geocoding method. Sampling error was incorporated into the estimates using bootstrap approximation from a standard normal distribution.[12]

For both the complete case and bias-adjusted analyses, we calculated the SARS-CoV-2 infection rates (per 1,000 persons), hospitalization proportions (hospitalized cases/reported cases), and case fatality rates (deaths/reported cases) by race/ethnicity group. We reported 95% confidence intervals (CI) for the complete case analysis and medians with 95% simulation intervals (SI) for the bias-adjusted estimates. We evaluated how accounting for missing race/ethnicity information impacts measures of racial/ethnic disparities by calculating the differences in SARS-CoV-2 infection rates in each race/ethnicity group compared with persons

of White race/ethnicity, among case-patients with reported race/ethnicity information, and among all case-patients after bias-adjustment. All analyses used R v3.6 (Vienna, Austria). The Georgia Department of Health determined this activity to be consistent with public health surveillance, so does not require informed consent or IRB approval.

## Results

Among the 19,637 cases reported in Fulton County from 29 February to 19 August 2020, 7,145 (36%) were missing race/ethnicity information in the case report. Data were more complete among the 1,840 hospitalized case-patients, where only 14 (3.5%) were missing race/ethnicity information. All deceased case-patients (n=456) had complete information on race/ethnicity.

Comparison of reported versus imputed race/ethnicity group showed that the algorithm's imputation accuracy varied by imputed race/ethnicity group (**Table 1**). Of the 5,535 persons who were imputed as Black race/ethnicity, 93% (95%CI: 92%, 93%) were reported as Black in case reports (n=5,118). Among persons imputed as Hispanic ethnicity, 84% (95%CI: 82%, 85%) were reported as Hispanic. The algorithm was less accurate for case-patients with race/ethnicity imputed as Asian (PV=69%, 95%CI: 61%, 74%) and as White (PV=55%, 95%CI: 53%, 56%). The PV estimates for racial/ethnic groups changed over time, likely due to changes in the prevalence of demographic groups affected by the pandemic over time (**Supplemental Table 1**).

In both the complete case and bias-adjusted analyses, the SARS-CoV-2 infection rates were highest among those classified as Other, followed by Hispanic, Black, White, and Asian (**Table 2a and 2b**). Imputation and bias-adjustment yielded higher estimates of infection rates than complete case analysis because more case-patients were included in the numerator. Estimated infection rates increased 1.8-fold for persons classified as Asian, 1.7-fold for White, 1.7-fold for

158 Hispanic, 1.6-fold for Other, and 1.5-fold for Black. Hospitalization proportions and case fatality
159 rates decreased across all race/ethnicity groups with imputation and bias-adjustment compared
160 with the complete case analyses, because more cases were included in the denominator. In
161 both the complete case and bias-adjusted analyses, case-patients who were classified as Black
162 race/ethnicity had the highest hospitalization proportions (complete case: 17%, 95%CI: 16%,
163 18%; bias-adjusted: 12%, 95%SI: 11%, 12%) and case fatality rates (complete case: 4.6%,
164 95%CI: 4.1%, 5.1%; bias-adjusted: 3.1%, 95%SI: 2.8%, 3.4%).
165
166 The magnitude of the absolute disparity—difference in SARS-CoV-2 infection rates for case-
167 patients classified in each race/ethnicity group compared with case-patients classified White—
168 increased in the bias-adjusted analysis relative to the complete case analysis for nearly all
169 race/ethnicity groups (**Table 3**). When comparing bias-adjusted with complete case results, the
170 absolute disparity in infection rates increased 1.3-fold among classified Black and 1.6-fold
171 among classified Hispanic race/ethnicity groups in reference to case-patients classified as
172 White.
173
174 **Discussion**

175 In this study, accounting for missing race/ethnicity information revealed greater differences in
176 SARS-CoV-2 infection rates comparing most racial/ethnic groups with case-patients of White
177 race. These results suggest that national estimates, which exclude case-patients with missing
178 race/ethnicity information, may underestimate the magnitude of absolute racial/ethnic disparities
179 in COVID-19 morbidity and mortality.[6,8]

180 Our results underscore the need for imputation combined with bias-adjustment. In our study
181 population, the PV estimates indicated that imputation alone overestimated infections among

case-patients classified as White and underestimated infections among case-patients classified as Black. Therefore, imputation alone would have been insufficient.

Both the complete case analysis and the bias-adjusted estimates demonstrate important absolute racial/ethnic disparities in the infection rates. The bias-adjusted estimates do not change our understanding of the direction of racial/ethnic disparities in the COVID-19 pandemic; however, the magnitude of racial/ethnic disparities changed meaningfully after bias-adjustment. In contrast, the hospitalization proportion and case fatality rate decreased across all classified race/ethnicity groups after accounting for missing race/ethnicity information because few hospitalized or deceased case-patients were missing race/ethnicity information. These results highlight the need for more complete reporting so that health equity and racial justice efforts aimed at addressing these disparities operate on the most accurate data possible.

The imputation of race/ethnicity has limitations. The Bayesian Improved Surname Geocoding algorithm limits the racial/ethnic groups that can be imputed to Black, Hispanic, Asian, White, or Other. The reliance on categories of 'other' is problematic for identifying and addressing disparities in other racial/ethnic populations (*e.g.* indigenous populations). Future studies should explore how accounting for missing race/ethnicity impacts other disease burden measures.

Our findings emphasize the importance of collecting complete race/ethnicity data at the time of testing, for the current pandemic and future outbreaks. When data are missing, Bayesian Improved Surname Geocoding combined with quantitative bias-adjustment provides better estimates of the racial/ethnic disparities in SARS-CoV-2 infection rates, hospitalization proportions, and case fatality rates.

## Tables

**Table 1:** Predictive values (PV) and 95% confidence intervals (CI) of the imputation by race/ethnicity based on residence and surname compared with reported race/ethnic group in the State Electronic Notifiable Disease Surveillance System

| | | Imputed Race/Ethnicity | | | | |
|---|---|---|---|---|---|---|
| | | **Black** | **Hispanic** | **Asian** | **White** | **Other** |
| **Reported Race/Ethnicity** | **Black** | 5118 | 68 | 13 | 1754 | 11 |
| | **Hispanic** | 77 | 1288 | 16 | 230 | 6 |
| | **Asian** | 16 | 15 | 145 | 80 | 4 |
| | **White** | 192 | 103 | 28 | 2827 | 2 |
| | **Other** | 132 | 68 | 12 | 302 | 1 |
| | **Total** | 5,535 | 1,543 | 214 | 5,193 | 24 |
| | **PV % (95% CI)** | 93% (92%, 93%) | 84% (82%, 85%) | 69% (61%, 74%) | 55% (53%, 56%) | 3.8% (0.1%, 15%) |

228
229

**Table 2a:** Complete case estimates of SARS-CoV-2 infection rates, hospitalization proportions, and case fatality rates by race/ethnic group among 12,222 cases reported to Fulton County Board of Health, 29 February – 18 Aug 2020.

| Race/ Ethnicity | Total infections | Hospitalized | Died | At Risk* | Infection rate per 1,000 (95%CI) | Hospitalized percentage (95%CI) | Case Fatality Rate as a percentage (95%CI) |
|---|---|---|---|---|---|---|---|
| **Asian** | 260 | 25 | 5 | 69987 | 3.7 (3.3, 4.2) | 9.6 (6.2, 14) | 1.9 (0.4, 3.8) |
| **Hispanic** | 1617 | 214 | 15 | 74328 | 22 (21, 23) | 13 (12, 15) | 0.9 (0.5, 1.4) |
| **Black** | 6964 | 1195 | 320 | 445992 | 16 (15, 16) | 17 (16, 18) | 4.6 (4.1, 5.1) |
| **White** | 3152 | 312 | 112 | 406755 | 7.7 (7.4, 8.0) | 9.9 (8.9, 11) | 3.6 (2.9, 4.2) |
| **Other** | 515 | 30 | 4 | 6056 | 85 (78, 92) | 5.8 (3.9, 8.0) | 0.8 (0.2, 1.6) |

**Table 2b:** Bias-adjusted estimates of SARS-COoV-2 infection rates, hospitalization proportions, and case fatality rates including 7,415 cases with imputed race/ethnicity, among 19,637 cases reported to Fulton County Board of Health before 18 Aug 2020.

| Race/ Ethnicity | Total infections (95%SI) | Hospitalized | Died | At Risk* | Infection rate per 1,000 (95%SI) | Hospitalized percentage (95%SI) | Case Fatality Rate as a percentage (95%SI) |
|---|---|---|---|---|---|---|---|
| **Asian** | 456 (439, 474) | 25 | 5 | 69987 | 6.5 (5.9, 7.2) | 5.5 (3.4, 7.6) | 1.1 (0.1, 2.1) |
| **Hispanic** | 2,691 (2,661, 2721) | 214 | 15 | 74328 | 36 (35, 38) | 7.9 (6.9, 9.0) | 0.6 (0.3, 0.8) |
| **Black** | 10,838 (10,327, 10,428) | 1195 | 320 | 445992 | 23 (23, 24) | 12 (11, 12) | 3.1 (2.8, 3.4) |
| **White** | 5,303 (5,250, 5,356) | 312 | 112 | 406755 | 13 (13, 13) | 2.1 (1.7, 2.5) | 2.1 (1.7, 2.5) |
| **Other** | 837 (810, 865) | 30 | 4 | 6056 | 138 (128, 148) | 0.5 (0.0, 0.9) | 0.5 (0.0, 0.9) |

*American Community Survey 5-year 2018 estimates

230
231
232
233
234
235
236
237

**Table 3:** Relative difference (RD) of SARS-CoV-2 infection rates among minority groups compared with non-Hispanic White persons among cases with complete information and after accounting for missing race/ethnicity among 4004 SARS-CoV-2 infected persons reported to Fulton County before 20 May 2020.

| | Complete Case | | Bias-Adjusted | | |
|---|---|---|---|---|---|
| **Race/ Ethnicity** | **Infection rate per 1,000 (95%CI)*** | **RD per 1,000 (95%CI)** | **Infection rate per 1,000 (95%SI)** | **RD per 1,000 (95%SI)** | **Relative change in magnitude of disparity** |
| Asian | 3.7 (3.3, 4.2) | −4.0 (−4.6, −3.5) | 6.5 (5.9, 7.2) | −6.5 (−6.8,−6.2) | 0.6 |
| Hispanic | 22 (21, 23) | 14 (13, 15) | 36 (35, 38) | 23 (23, 23) | 1.6 |
| Black | 16 (15, 16) | 7.9 (7.4, 8.3) | 23 (23, 24) | 10 (10, 11) | 1.3 |
| White | 7.7 (7.4, 7.8) | Reference | 13 (13, 13) | Reference | |
| Other | 85 (78, 92) | 77 (70, 84) | 138 (128, 148) | 125 (121, 130) | 1.6 |

# Appendix

**Supplemental Table 1**: Positive predictive value (PPV) of the imputation by race/ethnicity based on residence and surname compared with the reported race/ethnic group in COVID-19 case report stratified by months (March through May and June through August) of diagnosis

| | | Predicted Race/Ethnicity | | | | |
|---|---|---|---|---|---|---|
| **March–May** | | **Asian** | **Black** | **Hispanic** | **Other** | **White** |
| Reported Race/Ethnicity | **Asian** | 33 | 2 | 6 | 2 | 28 |
| | **Black** | 3 | 1183 | 12 | 3 | 654 |
| | **Hispanic** | 1 | 7 | 156 | 0 | 38 |
| | **Other** | 9 | 17 | 11 | 0 | 45 |
| | **White** | 5 | 69 | 30 | 0 | 608 |
| | **PPV** | 65% | 93% | 73% | 0% | 44% |
| **June–Aug** | | | | | | |
| Reported Race/Ethnicity | **Asian** | 112 | 14 | 9 | 2 | 52 |
| | **Black** | 10 | 3935 | 56 | 8 | 1100 |
| | **Hispanic** | 15 | 70 | 1132 | 6 | 192 |
| | **Other** | 3 | 115 | 57 | 1 | 257 |
| | **White** | 23 | 123 | 73 | 2 | 2219 |
| | **PPV** | 69% | 92% | 85% | 5% | 58% |

## References

1. Stokes EK, Zambrano LD, Anderson KN, et al. Coronavirus Disease 2019 Case Surveillance - United States, January 22-May 30, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(24):759-765. doi:10.15585/mmwr.mm6924e2
2. Health Equity Considerations & Racial & Ethnic Minority Groups. National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/racial-ethnic-minorities.html. Published 2020. Accessed July 17, 2020.
3. McClure ES, Vasudevan P, Bailey Z, Patel S, Robinson WR. Racial Capitalism within Public Health: How Occupational Settings Drive COVID- 19 Disparities. *Am J Epidemiol*. 2020:113-120.
4. Egede LE, Walker RJ. Structural Racism, Social Risk Factors, and Covid-19 — A Dangerous Convergence for Black Americans. *N Engl J Med*. 2020;383(12):e77(1)-e77(3). doi:10.1056/NEJMp2023616
5. Servik K. 'Huge hole' in COVID-19 testing data makes it harder to study racial disparities. *Science (80- )*. July 2020. doi:0.1126/science.abd7715
6. 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):e2016933. doi:10.1001/jamanetworkopen.2020.16933
7. Georgia Department of Public Health COVID-19 Daily Status Report. https://dph.georgia.gov/covid-19-daily-status-report. Published 2020. Accessed July 18, 2020.
8. Oppel R, Gebelhoff R, Lai K, Wright W, Smith M. The Fullest Look Yet at the Racial Inequity of Coronavirus. New York Times. https://www.nytimes.com/interactive/2020/07/05/us/coronavirus-latinos-african-

americans-cdc-data.html?campaign_id=2&emc=edit_th_20200706&instance_id=20039&nl=todaysheadlines&regi_id=71026656&segment_id=32674&user_id=c99fb3a6b3b754c. Published 2020. Accessed July 18, 2020.

9. Wu SL, Mertens AN, Crider YS, et al. Substantial underestimation of SARS-CoV-2 infection in the United States. *Nat Commun*. 2020. doi:10.1038/s41467-020-18272-4

10. Perkins NJ, Cole SR, Harel O, et al. Principled Approaches to Missing Data in Epidemiologic Studies. *Am J Epidemiol*. 2017;187(3):568-575. doi:10.1093/aje/kwx348

11. *The Coronavirus Aid, Relief, and Economic Security (CARES) Act*. United States; 2020.

12. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. (Gail M, Krickeberg K, Samet J, Tsiatis A, Wong W, eds.). New York: Springer; 2009. doi:10.1007/978-0-387-87959-8

13. American Community Survey: Hispanic or Latino Origin by Race. The United States Census Bureau. https://data.census.gov/cedsci/table?t=Race and Ethnicity&g=0500000US13121&tid=ACSDT5Y2018.B03002&moe=false&hidePreview=true. Published 2020. Accessed August 19, 2020.

14. Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res*. 2008;43(5 P1):1722-1736. doi:10.1111/j.1475-6773.2008.00854.x

15. About the American Community Survey. US Census Bureau. https://www.census.gov/programs-surveys/acs/about.html. Published 2020. Accessed July 8, 2020.

16. Khanna K, Imai K. Package ‘ wru ’: Who are You? Bayesian Prediction of Racial Category Using Surname and Geolocation. 2019. https://cran.r-project.org/web/packages/wru/wru.pdf.

298