Differential Privacy in the 2020 Census Will Distort COVID-19 Rates

Mathew E. Hauer1 and Alexis R. Santos-Lozada2

Abstract
Scholars rely on accurate population and mortality data to inform efforts regarding the coronavirus disease 2019 (COVID-19) pandemic, with age-specific mortality rates of high importance because of the concentration of COVID-19 deaths at older ages. Population counts, the principal denominators for calculating age-specific mortality rates, will be subject to noise infusion in the United States with the 2020 census through a disclosure avoidance system based on differential privacy. Using empirical COVID-19 mortality curves, the authors show that differential privacy will introduce substantial distortion in COVID-19 mortality rates, sometimes causing mortality rates to exceed 100 percent, hindering our ability to understand the pandemic. This distortion is particularly large for population groupings with fewer than 1,000 persons: 40 percent of all county-level age-sex groupings and 60 percent of race groupings. The U.S. Census Bureau should consider a larger privacy budget, and data users should consider pooling data to minimize differential privacy’s distortion.

Keywords
census 2020, differential privacy, COVID-19

As coronavirus disease 2019 (COVID-19) grips the world, scholars, policy makers, and journalists use population data to calculate various population-level COVID-19 rates (incidence or new case rate, prevalence or total case rate, and mortality) to better understand, communicate, address, and inform mitigation efforts of the COVID-19 pandemic (Dowd et al. 2020; Wadhera et al. 2020). Because of these rate calculations, we know that the elderly are more susceptible to COVID-19-related mortality (CDC 2020) and that racial minorities are presently affected at higher rates (Price-Haywood et al. 2020). Accurate COVID-19 rate calculations and estimates are thus paramount to managing this and future pandemics. Inaccurately assessing COVID-19 could lead to misallocation of resources and interventions to mitigate the crisis.

The calculation of any population-level COVID-19 rate is relatively straightforward: one divides the COVID-19 counts (incidence, prevalence, and deaths) by the appropriate population counts from census data. To date, scholars have focused largely on properly counting COVID-19 deaths (Banerjee et al. 2020; Remuzzi and Remuzzi 2020) with a focus on the numbers of cases and deaths. However, scholars and policy makers in the United States must be mindful of population counts in the denominator of COVID-19 rate calculations because of the implementation of differential privacy (DP) in the publication of the 2020 census counts.

A disclosure avoidance system (DAS) will be implemented with the 2020 census tabulations (Mervis 2019), whereby population counts will be subject to noise infusion in an effort to protect respondent privacy. The U.S. Census Bureau is charged with protecting the confidentiality of its respondents. Beginning with the 1970 census, the Census Bureau has used a wide array of disclosure avoidance techniques to protect respondent confidentiality. These techniques include suppression of tables with small cell sizes, swapping or interchanging responses, and suppressing and then imputing responses (Zayatz 2007). Starting with the 2020 census, the Census Bureau plans to “modernize” its disclosure avoidance practices using DP (Ruggles et al. 2019). This is the first large-scale, census-based implementation of DP in the history of this methodology and represents a monumental sea change in population statistics (Garfinkel, Abowd, and Powazek 2018).

1Florida State University, Tallahassee, FL, USA
2Pennsylvania State University, State College, PA, USA

Corresponding Author:
Mathew E. Hauer, Florida State University, Center for Demography and Population Health, 526 Bellamy Building, 113 Collegiate Loop, Tallahassee, FL 32306-2240, USA.
Email: mehauer@fsu.edu
Under the Census Bureau’s proposed DAS using DP, population counts will be subject to noise infusion whereby random numerical values are added or subtracted to “true” population data, drawn from a statistical distribution under a specific privacy budget; the smaller the budget, the greater the noise. The Census Bureau then postprocesses the data to eliminate fractional and negative populations created during the DP process. The differences between the underlying, “true” population counts in the Census Bureau Summary File (SF) and the noise-infused DAS counts could lead to substantial over- or underestimation of COVID-19 rates, dependent on the divergence between the two. The Census Bureau has yet to finalize its DAS algorithm, though it is continually trying to improve the algorithm, and it is unclear how similar the demonstration products are to the final product. Importantly, the Census Bureau could implement less privacy in exchange for less noise and more utility. Scholarly are only beginning to study DP, its accuracy, and its consequences. The extent to which DP would distort the calculation of COVID-19-related rates is currently untested. For the calculation of COVID-19 incidence and prevalence rates, there will be no alternative to differentially private 2020 census data. Given how crucial population counts are for the evaluation and tracking of epidemiological rates, noise-infused population counts could lead to erroneous COVID-19 rate calculations and harm our ability to understand the current pandemic and manage future public health crises. Accurate population counts are just as important as accurate COVID-19-related counts. The COVID-19 rates produced following the implementation of the proposed DAS produce rates that could hinder our understanding of disparities arising from the events such as the illustrative case of the COVID-19 pandemic. Furthermore, the changes in these rates could lead to misstating the impact of COVID-19 across space and population subgroups, within small areas and more noticeably for racial/ethnic minorities.

In this short article, we demonstrate the extent to which DP could distort COVID-19 rates by age-sex and by race by combining the most recent Census Bureau DAS demonstration products segmented by age and sex (Van Riper, Kugler, and Schroeder 2020) with empirical COVID-19 age and sex mortality curves from the Centers for Disease Control and Prevention (2020) and a hypothetical 70 percent infection rate, constituting the theoretical herd immunity for the United States (Kwok et al. 2020). This allows us to simulate the difference between hypothetical mortality rate calculations using counts produced with DP from population counts produced using current methods. Although we use mortality rates, COVID-19 incidence and prevalence would be identical in both bias and in their rate calculation.

Methods

We use two primary sources of data in our estimates concerning the denominators for COVID-19 rate calculations and one primary source of data concerning the numerators. For the denominators, we use the 2010 county-level population estimates from traditional disclosure avoidance techniques and 2010 county-level population estimates produced with the proposed DP 2010 demonstration product (Van Riper et al. 2020) from May 27, 2020: the most recent file with age × sex detail. We accessed county-level population counts in 10-year age groups by sex and county-level population counts by race/ethnicity. The 2010 demonstration product simulates the DP algorithm on the 2010 census SF 1 to provide a comparison between traditional disclosure avoidance counts and the new DP counts. The DP demonstration product provides the denominators for calculating the COVID-19 mortality rates but not the numerators.

To calculate the number of anticipated COVID-19 deaths by age and sex, we apply empirical age and sex mortality rates from the Centers for Disease Control and Prevention (2020) to the 2010 Census Bureau SF 1 data that are not produced using DP and assume a 70 percent infection rate before herd immunity halts the spread (Kwok et al. 2020). This allows us to estimate the anticipated mortality for the underlying, “true” population ($D_{i,a,s,SS}$) by county $i$, age group $a$, and sex group $s$. COVID-19 mortality rates are simply calculated as the numerical deaths divided by the population. We calculate the mortality rate under an SF and a DP denominator such that ($m_{i,a,s} = D_{i,a,s,SS} \div P_{i,a,s,c}$), where $P_{i,a,s,c}$ refers to the relevant population and $c$ refers to either SF or DP.

For our race analysis, we apply empirical mortality rates from the Centers for Disease Control and Prevention to each race group $r$ in each county $i$ and a 70 percent infection rate to estimate the COVID-19 mortality rates under SF and DP ($m_{i,r} = D_{i,r,c,SS} \div P_{i,r,c}$). We then calculate a mortality rate ratio (MRR), expressed as the ratio of the DP mortality rate to the SF mortality rate ($[M_{DP}/M_{SF}] - 1$), where values above 1.0 represent a DP mortality rate that exceeds the SF mortality rate.

Reproducible Research

All data and code necessary to reproduce the reported results are licensed under the CC-BY-4.0 license and are publicly available in a replication repository located at https://osf.io/2v7ea/?view_only=443404fc9af041dc876d0617385f9255.

Results

Figure 1a shows the distortion of COVID-19 age-sex-specific mortality rates by population size for U.S. counties using the 2010 demonstration products. We find that smaller age-sex populations have much higher absolute errors than larger populations. These errors are not limited to small areas...
Hauer and Santos-Lozada

or a single age group; rather these errors are present in all age groups. Additionally, using DP as the denominator causes some age-specific mortality rates to impossibly exceed 100 percent (red dots). For example, in the 2010 census, Kent County, Texas, contained 58 women aged 85 and older, but the DP count is 2. If COVID-19 incidence, prevalence, or fatalities exceed 2 individuals in this age-sex group, the COVID-19 calculated rate would impossibly exceed 100 percent. It is particularly worrisome that age-sex groups with fewer than 1,000 persons—more than 40 percent of all

Figure 1. The distortion of coronavirus disease 2019 age-sex-specific mortality rates for U.S. counties. We show only those county age-sex groups with less than 500 percent error. Red dots correspond to county age-sex groups with mortality rates that impossibly exceed 1.0. The blue line is a locally estimated scatterplot smoothing (LOESS; span = 1). (a) Age-sex-specific mortality rates. (b) Race-specific mortality rates. Errors drop precipitously with at least 1,000 persons.
county-level age-sex groupings in the United States—exhibit particularly large errors (Table 1), making any meaningful COVID-19 rate calculation difficult to interpret for large segments of the country.

The DAS distorts general mortality rates for racial/ethnic minorities (Santos-Lozada, Howard, and Verdery 2020), and Figure 1b shows the distortion of COVID-19 race-specific mortality rates by population size for U.S. counties. Such as with age-sex-specific mortality, error increases substantially as population size decreases for all race groups. Only white non-Hispanic exhibits the lowest error; all other race groups, including pooling all nonwhite groups together, exhibit large errors as population size decreases. Race groups with fewer than 1,000 persons—more than 60 percent of all county-race groupings—exhibit the largest errors.

Balancing Data Privacy and Utility

We highlight how the planned 2020 census data under DP will significantly alter our understanding of COVID-19 via noise-infused population counts. Using age-sex-specific COVID-19 mortality curves from the CDC, we show that DP will introduce substantial errors in COVID-19 expected age-sex-specific mortality rates—sometimes causing age-specific mortality rates to exceed 100 percent—hindering our ability to understand the pandemic. These errors are particularly large for approximately 40 percent of county age-sex groupings and 60 percent of county-race groupings containing fewer than 1,000 persons. Overall, DP will introduce significant challenges in our understanding of the COVID-19 global pandemic expected to last well into 2021.

Age groups with fewer than 1,000 persons can occur in counties with relatively large total populations. Autauga County, Alabama, has four age-sex groups with fewer than 1,000 persons (men 85 and older, women 85 and older, total 85 and older, and men 75 to 84) yet had a total population in 2010 of 109,000. Autauga County is one of approximately 500 U.S. counties with more than 100,000 people (putting it in the top 15th percentile of population size), yet it still contains age groups with fewer than 1,000 people. Thus, large distortions are not limited to small, rural counties but can be found in relatively large, urban counties as well.

How are we to understand this pandemic if the very foundation upon which we calculate the most basic rates contains significant distortion? How will cities, states, and the federal government effectively manage the current or future pandemics if crucial denominators are untrustworthy? The populations most at risk for DP distortion, namely, elderly and minority populations, are the very groups COVID-19 harms the most and are in need of the most targeted interventions. If we cannot parse out the noise from the true values, we are left with a muddied vision of the pandemic, and our responses will further reflect that uncertainty. To provide some guidance, we offer recommendations for the Census Bureau and those calculating COVID-19 rates.

The Census Bureau is still fine-tuning its DP algorithm and has previously expressed concern about the trade-off between privacy and utility (Abowd and Schmutte 2019). A second run of the DP algorithm dealt with numerous concerns of the data user community (U.S. Census Bureau 2020), yet its utility still needs assessment. However, the Census Bureau release of the second run of the DP algorithm only contains race-sex breakdowns, making it impossible to conduct such an assessment. Census data are foundational to many kinds of analyses, including some analyses the Census Bureau probably never envisioned, and unfortunately the COVID-19 pandemic arose in the midst of the Census Bureau’s privacy changes. Because the Census Bureau DP demonstration products are so new, deep analysis of the impact these changes will have on the utility of public health

Table 1. Absolute Percentage Errors by Population Size for Age-Sex Groups and for Race/Ethnic Groups.

| Population Size | Median Absolute Percentage Error | Mean Absolute Percentage Error | n     | Percentage of County-Age-Sex Groups |
|----------------|----------------------------------|--------------------------------|-------|-------------------------------------|
| Age-sex        |                                  |                                |       |                                     |
| <1,000         | 13.4                             | 24.4                           | 18,991| 42.1                                |
| <2,500         | 8.3                              | 17.7                           | 29,147| 64.7                                |
| <5,000         | 6.4                              | 15.1                           | 35,318| 78.4                                |
| <10,000        | 5.4                              | 13.7                           | 39,518| 87.7                                |
| <20,000        | 4.8                              | 13.0                           | 42,089| 93.4                                |
| All            | 4.2                              | 12.1                           | 45,062| 100.0                               |
| Race/ethnicity |                                  |                                |       |                                     |
| <1,000         | 18.1                             | 46.6                           | 16,275| 60.7                                |
| <2,500         | 13.3                             | 41.0                           | 18,650| 69.5                                |
| <5,000         | 10.5                             | 37.6                           | 20,392| 76.0                                |
| <10,000        | 8.4                              | 34.7                           | 22,140| 82.6                                |
| <20,000        | 6.9                              | 32.4                           | 23,723| 88.5                                |
| All            | 4.5                              | 28.6                           | 26,819| 100.0                               |

Note: Pop refers to populations less than or equal to a given value.
data is yet to be undertaken. As we show, the DP algorithm, as proposed, sacrifices the usefulness of basic COVID-19 calculations in many counties and population groups.

There is still time for the Census Bureau to continue refining its DP algorithm or improve the privacy budget to allow more stable estimates in more population groups. The first 2020 census data products were originally slated for release in December 2020, but with the updated 2020 census timeline, the first products should be released by April 2021. The Centers for Disease Control and Prevention lags health and mortality data, making detailed COVID-19-related analyses very likely reliant on 2020 census noise-infused population counts rather than population counts produced using traditional methods. If the DP algorithm continues to produce distorted COVID-19 rates, data users might turn to outdated population estimates released prior to DP in their COVID-19 calculations.

The Census Bureau should consider alternative data sets specifically tailored for COVID-19 analyses, alternative DASs, or a larger privacy budget during this historic pandemic. It is entirely possible that future scientists, during the next major pandemic, will turn to the remnants of the COVID-19 data to understand their own pandemic—data that DP will certainly distort. The decisions the Census Bureau makes now will have long-term repercussions for what we can learn about COVID-19. Scholars, policy makers, and journalists turn toward the last major global pandemic, the 1918 influenza, to draw important parallels from the historical clues left behind in photographs, newspapers, and scientific articles. Those parallels play a powerful role in shaping public discourse, even with their historical patina. When we look back on COVID-19 during the next major global pandemic, any statistical measures arising from the United States will be far less meaningful because of the infusion of noise in the very building blocks of COVID-19 rates.

When, and not if, the Census Bureau releases DP data, all data users analyzing COVID-19 need to be aware of these limitations in using DP data for COVID-19 analyses. On the basis of our findings, we offer three recommendations to scholars and policy makers. First, we suggest a minimum cell size of 1,000 persons for the calculation of any COVID-19 rates (incidence, prevalence, and mortality). The distortion in COVID-19 rates rapidly shrinks as population sizes increase, especially in sizes larger than 1,000 persons. Second, scholars and policy makers can combine areas to create larger cell sizes via regions, sacrificing geographic detail for population specificity. The Census Bureau uses this approach for its public use microdata samples, and we recommend a similar approach for COVID-19 analyses. Third, scholars can pool data together in either wider age intervals (i.e., 20-year age intervals rather than 10-year age intervals) or wider race classifications (i.e., using the Office of Management and Budget’s minimum race classifications rather than the fully detailed nine-race classification). These strategies, either in isolation or in combination, will minimize the distortion in COVID-19 rate calculations.

The Census Bureau’s demonstration product presently contains only age-sex-county and race-county breakdowns and does not contain age-sex-race-county. Yet race differentials in COVID-19 mortality are an important aspect of the pandemic (Hooper, Nápoles, and Pérez-Stable 2020). The potential errors in COVID-19 mortality by age and sex are already significantly large, and we believe that analyzing COVID-19 mortality by age-sex-race would further reduce cell sizes, ensuring an even greater number of combinations with fewer than 1,000 persons, the identified threshold with the largest errors.

As the pandemic continues, scholars, policy makers, and journalists should embrace minimum standards for COVID-19 analyses using the 2020 census and subsequent data products. Recent visualizations by the New York Times and the Centers for Disease Control and Prevention (CDC 2020; Oppel et al. 2020) concerning racial/ethnic disparities in COVID-19 demonstrate the intense hunger for detailed COVID-19 analysis. Future analyses should be, at a minimum, informed of the issues of using noise-infused population counts and should incorporate the strategies outlined above to ensure analyses accurately reflect their chosen measurement and the social phenomenon of interest.

Acknowledgments

We gratefully acknowledge early comments and feedback from B. Jarosz, J. Howard, M. Taylor, and D. Van Riper.

ORCID iDs

Mathew E. Hauer https://orcid.org/0000-0001-9390-5308
Alexis R. Santos-Lozada https://orcid.org/0000-0002-8437-4493

References

Abowd, John M., and Ian M. Schmutte. 2019. “An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices.” American Economic Review 109(1):171–202.
Banerjee, Amitava, Laura Pasea, Steve Harris, Arturo Gonzalez-Izquierdo, Ana Torralbo, Laura Shalleross, Mahdad Noursadeghi, et al. 2020. “Estimating Excess 1-Year Mortality Associated with the COVID-19 Pandemic According to Underlying Conditions and Age: A Population-Based Cohort Study.” The Lancet 395(10238):1715–25.
CDC (Centers for Disease Control and Prevention). 2020. “Coronavirus Disease 2019 (COVID-19).” Retrieved July 8, 2020. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/racial-ethnic-minorities.html.
Dowd, Jennifer Beam, Liliana Andriano, David M. Brazel, Valentina Rotondi, Per Block, Xuejie Ding, Yan Liu, and Melinda C. Mills. 2020. “Demographic Science Aids in Understanding the Spread and Fatality Rates of COVID-19.” Proceedings of the National Academy of Sciences 117(18):9696–98.
Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. “Issues Encountered Deploying Differential Privacy.” Pp. 133–37 in Proceedings of the 2018 Workshop on Privacy in the Electronic Society. New York: Association for Computing Machinery.

Hooper, Monica Webb, Anna Maria Nápoles, and Eliseo J. Pérez-Stable. 2020. “COVID-19 and Racial/Ethnic Disparities.” *JAMA* 323(24):2466–67.

Kwok, Kin On, Florence Lai, Wan In Wei, Samuel Yeung Shan Wong, and Julian W. T. Tang. 2020. “Herd Immunity—Estimating the Level Required to Halt the COVID-19 Epidemics in Affected Countries.” *Journal of Infection* 80(6):e32–33.

Mervis, Jeffrey. 2019. “Can a Set of Equations Keep U.S. Census Data Private?” *Science*. Retrieved January 29, 2021. https://www.sciencemag.org/news/2019/01/can-set-equations-keep-us-census-data-private.

Oppel, Richard A., Jr., Robert Gebeloff, K. K. Rebecca Lai, Will Wright, and Mitch Smith. 2020. “The Fullest Look Yet at the Racial Inequity of Coronavirus.” *The New York Times*, July 5.

Price-Haywood, Eboni G., Jeffrey Burton, Daniel Fort, and Leonardo Seoane. 2020. “Hospitalization and Mortality among Black Patients and White Patients with COVID-19.” *New England Journal of Medicine* 382(26):2534–43.

Remuzzi, Andrea, and Giuseppe Remuzzi. 2020. “COVID-19 and Italy: What Next?” *The Lancet* 395(10231):1225–28.

Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. “Differential Privacy and Census Data: Implications for Social and Economic Research.” *AEA Papers and Proceedings* 109:403–408.

Santos-Lozada, Alexis R., Jeffrey T. Howard, and Ashton M. Veredery. 2020. “How Differential Privacy Will Affect Our Understanding of Health Disparities in the United States.” *Proceedings of the National Academy of Sciences* 117(24):13405–12.

U.S. Census Bureau. 2020. “Developing the DAS: Progress Metrics and Data Runs.” Retrieved July 8, 2020. https://www.census.gov/programs-surveys/decennial-census/2020-census/plan-nmg-2020-census-data-products/2020-das-metrics.html.

Van Riper, David, Tracy Kugler, and Jonathan Schroeder. 2020. “IPUMS NHGIS Privacy-Protected 2010 Census Demonstration Data, Version 20200527.” Minneapolis, MN: IPUMS National Historical Geographic Information System.

Wadhera, Rishi K., Priya Wadhera, Prakriti Gaba, Jose F. Figueroa, Karen E. Joynt Maddox, Robert W. Yeh, and Changhai Chen. 2020. “Variation in COVID-19 Hospitalizations and Deaths across New York City Boroughs.” *JAMA* 323(21):2192–95.

Zayatz, Laura. 2007. “Disclosure Avoidance Practices and Research at the US Census Bureau: An Update.” *Journal of Official Statistics* 23(2):253.

**Author Biographies**

**Mathew E. Hauer** is an assistant professor in the Department of Sociology and an affiliate of the Center for Demography and Population Health at Florida State University. His research focuses on the intersection between climate change and demographic processes to better understand the current and projected impacts of climate change on human society.

**Alexis R. Santos-Lozada** is an assistant professor in the Department of Human Development and Family Studies and an affiliate of the Population Research Institute, and social disparities faculty at the Social Sciences Research Institute at Pennsylvania State University. His research focuses on stress, well-being, and mortality with emphasis on the implication of measurements to our understanding of health disparities with emphasis in minority populations.