**Differential privacy in the 2020 Census distorts COVID-19 rates**

Mathew E. Hauer1,2\* & Alexis R. Santos-Lozada3,4

**Affiliations:**

1Department of Sociology, Florida State University, Tallahassee, Florida, USA.

2Center for Demography and Population Health, Florida State University.

3Human Development and Family Studies, Pennsylvania State University, State College, PA, USA.

4Population Research Institute, Pennsylvania State University.

\*[mehauer@fsu.edu](mailto:hauer@uga.edu)

As the coronavirus disease 2019 (COVID-19) grips the world, scientists, policy makers, and journalists use population data calculate various CVOID-19 rates (mortality, incidence or the new case rate, and prevalence or the total case rate) to better understand, communicate, address, and inform mitigation efforts of the COVID-19 pandemic 1. Because of these rate calculations, we know that the elderly are more susceptible to COVID-19 related mortality2 and that racial minorities are presently affected at higher rates [NYT citation]. Accurate COVID-19 rate calculations and estimates are thus paramount to managing the pandemic and illuminating how to manage future pandemics. Inaccurately assessing COVID-19 could lead to misallocation of resources and interventions to mitigate the crisis.

The calculation of any 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, scientists have largely focused on properly counting COVID-19 deaths 3,4 with a focus on the numeric amount of cases and deaths. However, scientists and policy makers in the United States will need to pay closer attention to population counts due to the implementation of differential privacy (DP) in the publication of Census 2020 counts. DP is a disclosure avoidance system, implemented with the 2020 Census tabulations 5, where population counts will be subject to noise injection in an effort to protect respondent privacy. Scientists are only beginning to study DP’s implementation and the extent to which DP, as proposed, 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 DP Census 2020 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 for the calculation of COVID-19 rates and after the release of Census 2020 data we fear DP will render most COVID-19 rates confused at best and highly inaccurate at worst. The implementation of DP, as proposed, will substantially reduce our understanding of the pandemic’s dynamics for rural areas, racial/ethnic minorities, and age groups.

To demonstrate the extent to which DP could distort COVID-19 rates by age-sex and by race, we combine the US Census Bureau’s DP demonstration products 8 with empirical COVID-19 age and sex mortality curves from Italy 2 and the United Kingdom [CITE]. This allows us to simulate the difference between hypothetical mortality rate calculations using counts produced with DP from population counts produced using current methods.

The US Census Bureau is charged with protecting the confidentiality of its respondents. Beginning with Census 1970, the US Census Bureau employed 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 6. Starting with Census 2020, the US Census Bureau plans to “modernize” its disclosure avoidance practices using DP 7. This is the first large-scale, Census based implementation of differential privacy in the history of this methodology and represents a monumental sea-change in population statistics [CITE].

Under the Census Bureau’s proposed DP algorithm, population counts will be subject to noise infusion, drawn from a statistical distribution. The DP algorithm operates by adding or subtracting random numerical values to “true” population data, under a specific privacy budget. To illustrate how DP alters population counts, **Figure 1** shows age-sex structures from 2010 population counts (gray pyramids) and 2010 counts from the DP demonstration product for six sample US counties (in red). For some county age-sex groups, the changes in age-sex structure are negligible or marginal. Some counties, however, show either one age-sex category or the entire pyramid substantially altered by DP. The differences between the underlying, “true” population counts in the Census Summary File and the noise-infused DP counts could lead to substantial over/under estimation of COVID-19 rates, dependent on the divergence between the two.

**Figure 2** shows the distortion of COVID-19 age-sex specific mortality rates by population size for US counties using the 2010 demonstration products. We calculate absolute error for each county-age-sex combination. We find that smaller age-sex populations have much higher absolute errors than larger populations. These errors are not limited to small areas 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% (red dots). For example, Census 2010’s Lincoln County Georgia contained 183 women aged 80+ but the DP count is 5. If the COVID-19 incidence, prevalence, or fatality, exceeds 5 individuals in this age-sex group, the COVID-19 calculated rate would impossibly exceed 100%. It is particularly worrisome that age-sex groups with fewer than 1000 persons -- more than 40% of all county-age-sex groupings -- exhibit particularly large errors (**Table 1**) making any meaningful COVID-19 rate calculation difficult to interpret for large segments of the country.

The distortion in COVID-19 rates is not limited to age-sex population groupings but impacts race-specific analyses too**.** DP distorts general mortality rates for racial/ethnic minorities 9 and **Figure 3** shows the distortion of COVID-19 race specific mortality rates by population size for US counties. Much like with age-sex specific mortality, error increases substantially as population size decreases for all race groups. Only White, Non-Hispanic exhibit the lowest error; all other race groups – including pooling all non-white groups together – exhibit large errors as population size decreases. Race-groups with fewer than 1000 persons – more than 60% of all county-race groups – exhibit the largest errors.

**BALANCING DATA PRIVACY AND UTILITY**

We highlight how the planned, noise-infused U.S. Census data will significantly alter our understanding of COVID-19 via noise-infused population counts. Using age-sex specific COVID-19 mortality curves from Italy and Wuhan, we show that differential privacy will introduce significant errors in COVID-19 expected age-sex specific mortality rates – sometimes causing age-specific mortality rates to exceed 100% - hindering our ability to understand the pandemic. These errors are particularly large for approximately 40% of county age-sex groupings and 60% of county-race groupings containing fewer than 1000 persons. Overall, differential privacy will introduce significant challenges in our understanding of mortality amid a global pandemic expected to last well into 2021.

The Census Bureau’s demonstration product currently only contains age-sex-county and race-county breakdowns and does not contain age-sex-race-county. Yet race differentials in COVID mortality are an important aspect of the pandemic 10. The potential errors in COVID mortality by age and sex are already significantly large and we believe analyzing COVID mortality by age-sex-race would further reduce cell sizes, ensuring an even greater number of combinations with fewer than 1000 persons – the identified threshold with the largest errors. How are we to understand this pandemic if the very foundation upon which we calculate the most basic rates contain significant errors? How will cities, states, and the federal government effectively manage the current or future pandemics if important denominators are untrustworthy? 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, navigating this error with little, if any, guidance. The Census Bureau should publish suggested guidance on using DP data far in advance of the release of DP products to minimize their disruption.

To provide some guidance, we offer recommendations for the Census Bureau and those calculating COVID-19 rates.

The Census Bureau is still fine tuning their DP algorithm and has previously expressed concern about the trade off between privacy and utility 11. A second run of the DP algorithm dealt with numerous concerns of the data user community [CITE], yet its utility still needs assessment. The currently proposed algorithm sacrifices the usefulness of basic COVID-19 calculations in most counties. Census data are foundational to many kinds of analyses – some analyses the Census Bureau probably never envisioned – and unfortunately the COVID-19 pandemic arose during disclosure avoidance modernization. Because the Census Bureau DP demonstration products are so new, deep analysis of the impact this disclosure modernization will have on the utility of public health data are yet to be determined.

The first Census 2020 data products were originally slated for release in December 2020 but with the updated Census 2020 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-related analyses in the next 12 months very likely reliant on Census 2020 noise-infused population counts rather than population counts produced using traditional methods. There is still time for the Census Bureau to continue refining their DP algorithm or improve the privacy budget to allow more stable estimates in more population groups. Otherwise data users might turn to outdated population estimates released prior to DP in their COVID-19 calculations.

The decisions the Census Bureau makes now will have long-term repercussions for what we can learn about COVID-19. Scientists, policymakers, and journalists turn toward the last major global pandemic – the 1918 Spanish Flu – to draw important parallels from the historical clues left behind in pictures, 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 during the next major global pandemic, as we demonstrate here, any statistical measures arising from the United States will be far less meaningful due to the injection of noise in the very building blocks of COVID-19 rates. The US Census Bureau should consider alterative datasets, alternative disclosure avoidance systems, or a larger privacy budget during this historical pandemic. It is entirely possible that future scientists of 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.

When, and not if, the Census Bureau releases DP data, the breadth of data users analyzing COVID-19 need to be aware of these limitations in using DP data for COVID-19 analyses. Based on our findings, we offer three recommendations to scientists and policy makers. First, we suggest a minimum cell size of 1000 persons for the calculation of any COVID-19 rates (fatality, incidence, and prevalence). COVID-19 rates rapidly approach acceptable error rates as population sizes get larger than 1000 persons. Second, scientists and policymakers can combine areas to create larger cell sizes via regions, sacrificing geographic detail for specificity. The Census Bureau uses this approach for their public use microdata samples (PUMS), and we recommend a similar approach for COVID-19 analyses. Third, scientists can pool data together in either wider age intervals (ie 20-year age intervals rather than 10-year age intervals) or wider race classifications (ie using OMB’s 2, 4, or 5 race classifications rather than the fully detailed 9 race classification). These strategies, either in isolation or in combination, will minimize the uncertainty in COVID-19 rate calculations.

As the pandemic continues, scientists, policy makers, and journalists should embrace minimum standards for COVID-19 analyses using Census 2020 data products. Recent visualizations by the New York Times and the CDC demonstrate the intense hunger for detailed COVID-19 analysis. Future analyses should be, at minimum, informed of the issues of using noise-infused population counts and should employ strategies outlined above to ensure analyses are of the highest possible fidelity.

**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/fp52x/?view\_only=754d9a72a2ea4f6b8e0c193dc9a590d1.

**References**

1. Wadhera, R. K. *et al.* Variation in COVID-19 hospitalizations and deaths across New York City boroughs. *Jama* (2020).

2. Dowd, J. B. *et al.* Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9696–9698 (2020).

3. Remuzzi, A. & Remuzzi, G. COVID-19 and Italy: what next? *The Lancet* **395**, 1225–1228 (2020).

4. Banerjee, A. *et al.* 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**, 1715–1725 (2020).

5. Mervis, J. Can a set of equations keep U.S. census data private? *Science* (2019) doi:10.1126/science.aaw5470.

6. Zayatz, L. Disclosure avoidance practices and research at the US Census Bureau: An update. *J. Off. Stat.* **23**, 253 (2007).

7. Ruggles, S., Fitch, C., Magnuson, D. & Schroeder, J. Differential privacy and census data: Implications for social and economic research. in *AEA papers and proceedings* vol. 109 403–08 (2019).

8. National Historical GIS. Differentially Private 2010 Census Data. https://www.nhgis.org/differentially-private-2010-census-data (2019).

9. Santos-Lozada, A. R., Howard, J. T. & Verdery, A. M. How differential privacy will affect our understanding of health disparities in the United States. *Proc. Natl. Acad. Sci.* (2020) doi:10.1073/pnas.2003714117.

10. Hooper, M. W., Nápoles, A. M. & Pérez-Stable, E. J. COVID-19 and racial/ethnic disparities. *Jama* (2020).

11. Abowd, J. M. & Schmutte, I. M. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *Am. Econ. Rev.* **109**, 171–202 (2019).

**Figure 1** Changes in population pyramids due to the implementation of differential privacy in 2010 U.S. Census data for six US counties.

A picture containing looking, bunch, filled, many

Description automatically generated

**Figure 2** The distortion of COVID-19 age-sex specific mortality rates for US counties. We show only those county age-sex groups with less than 500% error. Red dots correspond to county age-sex groups with mortality rates that impossibly exceed 1.0.

A picture containing text, map, table, photo

Description automatically generated

**Figure 3** The distortion of COVID-19 race-specific mortality rates for US counties. We show only those county race groups with less than 500% error.

A close up of a map

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age-Sex** | | | | |
| **Pop. Size** | **Median Abs. % Err** | **Mean Abs. % Err** | **n** | **% of county-age-sex groups** |
| < 1,000 | 13.40% | 40.90% | 17,950 | 42.90% |
| < 2,500 | 8.30% | 27.60% | 28,836 | 69.00% |
| < 5,000 | 6.50% | 23.60% | 34,379 | 82.20% |
| < 10,000 | 5.50% | 21.60% | 37,711 | 90.20% |
| < 20,000 | 5.00% | 20.50% | 39,735 | 95.00% |
| All | 4.60% | 19.50% | 41,812 | 100.00% |
| **Race/Ethnicity** | | | | |
| **Pop. Size** | **Median Abs. % Err** | **Mean Abs. % Err** | **n** | **% of county-age-sex groups** |
| < 1,000 | 15.66% | 47.21% | 16,085 | 60.40% |
| < 2,500 | 11.62% | 41.46% | 18,460 | 69.32% |
| < 5,000 | 9.46% | 38.00% | 20,202 | 75.86% |
| < 10,000 | 7.66% | 35.03% | 21,950 | 82.43% |
| < 20,000 | 6.28% | 32.70% | 23,533 | 88.37% |
| All | 4.17% | 28.91% | 26,629 | 100.00% |