Differential privacy in the 2020 Census will distort COVID-19 rates

Abstract

Keywords: Census 2020, Differential Privacy, COVID-19

Background: Scientists and policy makers 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 due to 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 via a disclosure avoidance system based on differential privacy.

Methods: We combine the US Census Bureau's DAS demonstration products segmented by age and sex with empirical COVID-19 age and sex mortality curves from the CDC and a hypothetical 70% infection rate, constituting the theoretical herd immunity for the US. This allows for a simulation of the difference between hypothetical mortality rate calculations using counts produced with DAS from population counts produced using current methods.

Results: We show that differential privacy will introduce substantial distortion in COVID-19 mortality rates – sometimes causing mortality rates to exceed 100% – hindering our ability to understand the pandemic. This distortion is particularly large for population groupings with fewer than 1000 persons – 40% of all county-level age-sex groupings and 60% of race groupings.

Conclusions: The US Census should consider alternative datasets specifically tailored for COVID-19 analyses, alternative disclosure avoidance systems, or a larger privacy budget during this historical pandemic. Data users should consider pooling data for minimum cell sizes above 1000 persons to increase population sizes to minimize distortion.

As the coronavirus disease 2019 (COVID-19) grips the world, scientists, policy makers, and journalists use population data to calculate various population-level COVID-19 rates (incidence or the new case rate, prevalence or the total case rate, and mortality) to better understand, communicate, address, and inform mitigation efforts of the COVID-19 pandemic (1, 2). Because of these rate calculations, we know that the elderly are more susceptible to COVID-19 related mortality (3) and that racial minorities are presently affected at higher rates (4). 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, scientists have largely focused on properly counting COVID-19 deaths (5) with a focus on the numeric amount of cases and deaths. However, scientists and policy makers in the United States must be mindful of population counts in the denominator of COVID-19 rate calculations due to the implementation of differential privacy (DP) in the publication of Census 2020 counts.

The Disclosure Avoidance System (DAS) to be implemented with the 2020 Census tabulations (6) relies on DP, where population counts will be subject to noise infusion in an effort to protect respondent privacy. 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 (7). Starting with Census 2020, the US Census Bureau plans to "modernize" its disclosure avoidance practices using DP (8). 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 (9).

Under the Census Bureau's proposed DAS using DP, population counts will be subject to noise infusion where 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 post-processes the data to eliminate fractional and negative populations created during the DP process. The differences between the underlying, "true" population counts in the Census Summary File and the noise infused DAS counts could lead to substantial over/under estimation of COVID-19 rates, dependent on the divergence between the two. The Census Bureau has yet to finalize their DAS algorithm, though they are continually trying to improve it, and it is unclear how similar the demonstration products are to final product. Importantly, the Census Bureau could implement less privacy in exchange for less noise and more utility.

Scientists are only beginning to study DAS, 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 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 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.

To demonstrate the extent to which differential privacy could distort COVID-19 rates by age-sex and by race, we combine the most recent US Census Bureau's DAS demonstration products segmented by age and sex (10) with empirical COVID-19 age and sex mortality curves from the CDC (3) and a hypothetical 70% infection rate, constituting the theoretical herd immunity for the US (11). This allows for a simulation of the difference between hypothetical mortality rate calculations using counts produced with DAS from population counts produced using current methods. Though we use mortality rates, COVID-19 incidence and prevalence would be identical in both bias and in their rate calculation.

MATERIALS AND METHODS We utilize 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 differential privacy 2010 demonstration product (10) 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 Census 2010 Summary File 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/sex, we apply empirical age/sex mortality rates from the CDC (3) to the 2010 Census Bureau Summary File 1 data (SF) that are not produced using DP and assume a 70% infection rate before herd immunity halts the spread (11). This allows us to estimate the anticipated mortality for the underlying, "true" population ($D_{i,a,s,SF}$) by county i, age group a, and sex group s. COVID-19 mortality rates are simply calculated as the numeric deaths divided by the population. We calculate the mortality rate under an SF and a DP denominator such that $m_{i,a,s} = D_{i,a,s,SF}/P_{i,a,s,c}$ 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 CDC to each race group r in each county i and a 70% infection rate to estimate the COVID-19 mortality rates under SF and DP ($m_{i,r} = D_{i,r,SF}/P_{i,r,}$). We then calculate a mortality rate ratio (MRR), expressed as the ratio of the DP to SF mortality rates ($M_{DP}/M_{SF} - 1$), where values above 1.0 represent DP mortality rate which exceeds the SF mortality rate.

All data and code necessary to reproduce the reported results are publicly available here.

RESULTS

Figure 1a shows the distortion of COVID-19 age-sex specific mortality rates by population size for US 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 or a single age group, rather these errors are present in all age groups. Additionally, using DAS as the denominator causes some age-specific mortality rates to rarely, but impossibly exceed 100% (red dots). For example, Census 2010's Kent County Texas contained 58 women aged 85+ but the DP count is 2. If the

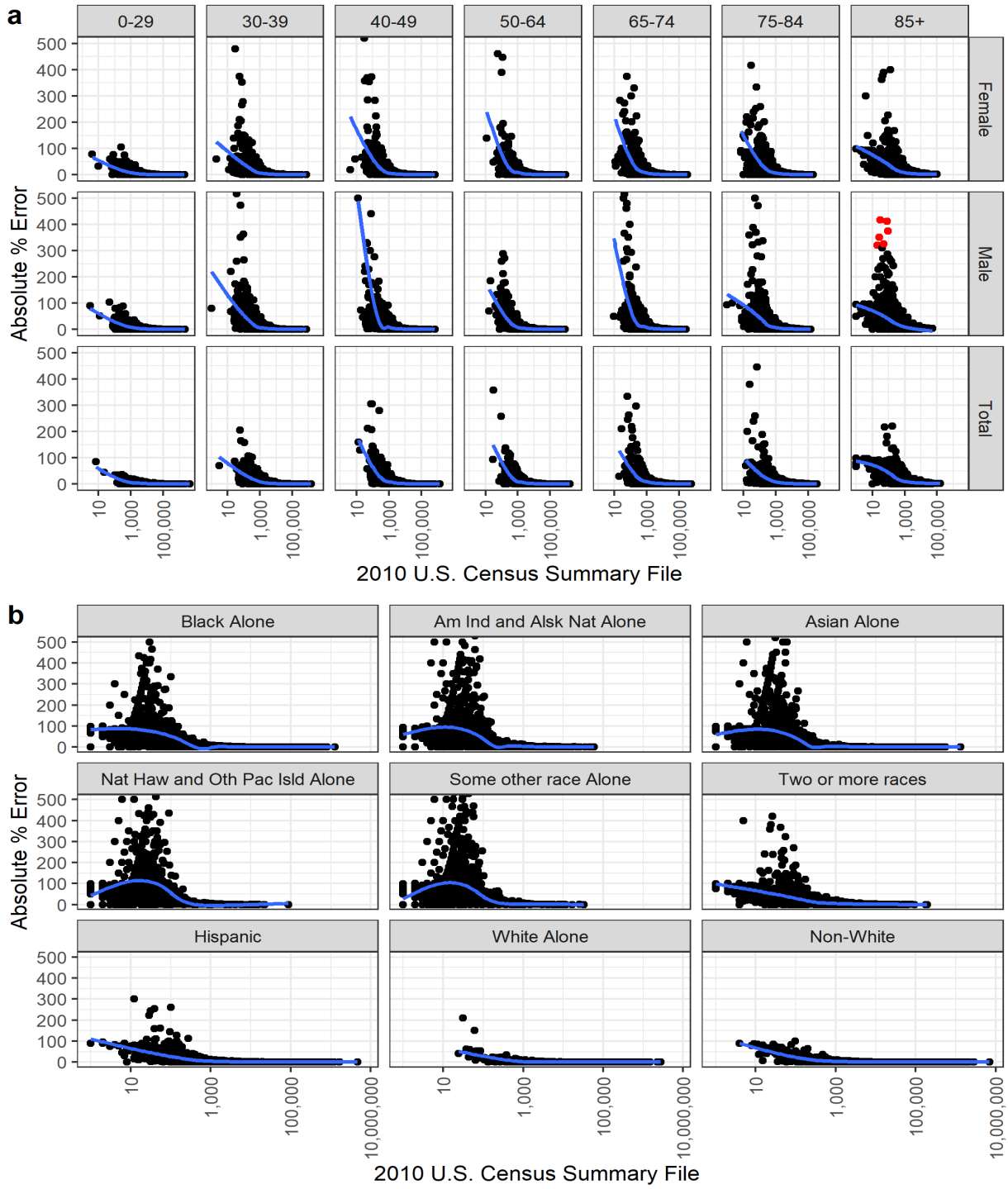

Figure 1: **The distortion of COVID-19 age-sex specific mortality rates for US counties.** We show only those county age-sex groups with less than 500% error. Red dots correspond to county age-sex groups with mortality rates that impossibly exceed 1.0. (a) shows age-sex specific mortality rates and (b) shows race-specific mortality rates. Errors drop precipitously with at least 1000 persons.

Table 1: Absolute percent errors by population size for age-sex groups and for race-ethnic groups. Pop refers to populations less than or equal to a given value.

	Pop	Median Error %	Mean Error %	n	% of Total
Age	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%
	10,000,000	4.2%	12.1%	45,062	100.0%
Race	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%
	10,000,000	4.5%	28.6%	26,819	100.0%

COVID-19 incidence, prevalence, or fatality, exceeds 2 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-level age-sex groupings in the US – exhibit particularly large errors making any meaningful COVID-19 rate calculation difficult to interpret for large segments of the country.

DAS distorts general mortality rates for racial/ethnic minorities (12) and **Figure 1b** 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, 2020 U.S. Census data under DP will significantly alter our understanding of COVID-19 via noise-infused population counts. Using empirical age-sex specific COVID-19 mortality curves from the CDC, we show that DAS will introduce substantial 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 the COVID-19 global pandemic expected to last well into 2021 or beyond.

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 of DP distortion – namely the old and minority populations – are the very groups COVID-19 harms the most and 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 their DAS algorithm and has previously expressed concern about the tradeoff between privacy and utility (13). A second run of the DAS algorithm dealt with numerous concerns of the data user community (14), yet its utility still needs assessment. Census data are foundational to many kinds of analyses – some analyses the Census Bureau probably never envisioned – and unfortunately the COVID-19 pandemic arose in the midst the Census Bureau's privacy changes. Because the Census Bureau DAS demonstration products are so new, deep analysis of the impact these changes will have on the utility of public health data are yet to be determined. 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 their DP algorithm or improve the privacy budget to allow more stable estimates in more population groups.

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 very likely reliant on Census 2020 noise-infused population counts rather than population counts produced using traditional methods. If the DAS algorithm continues to produce distorted COVID-19 rates, data users might turn to outdated population estimates released prior to DAS in their COVID-19 calculations.

The US Census Bureau should consider alternative datasets specifically tailored for COVID-19 analyses, 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. 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-19 during the next major global pandemic, any statistical measures arising from the United States will be far less meaningful due to the infusion of noise in the very building blocks of COVID-19 rates.

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 (incidence, prevalence, and mortality). The distortion in COVID-19 rates rapidly shrinks as population sizes increase, especially in sizes larger than 1000 persons. Second, scientists and policymakers can combine areas to create larger cell sizes via regions, sacrificing geographic detail for population 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 the Office of Management and Budget's minimum race classifications rather than the fully detailed 9 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 (15). The potential errors in COVID-19 mortality by age and sex are already significantly large and we believe analyzing COVID-19 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.

As the pandemic continues, scientists, policy makers, and journalists should embrace minimum standards for COVID-19 analyses using Census 2020 and subsequent data products. 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 accurately reflect their chosen measurement and the social phenomenon of interest.

References

[1] Wadhera, R. K., Wadhera, P., Gaba, P., Figueroa, J. F., Maddox, K. E. J., Yeh, R. W., and Shen, C. *Jama* (2020).

[2] Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., Liu, Y., and Mills, M. C. *Proceedings of the National Academy of Sciences* **117**(18), 9696–9698 (2020).

[3] CDC. (2020). Website: https://covid.cdc.gov/covid-data-tracker.

[4] Price-Haywood, E. G., Burton, J., Fort, D., and Seoane, L. *New England Journal of Medicine* (2020).

[5] Banerjee, A., Pasea, L., Harris, S., Gonzalez-Izquierdo, A., Torralbo, A., Shallcross, L., Noursadeghi, M., Pillay, D., Sebire, N., Holmes, C., et al. *The Lancet* (2020).

[6] Mervis, J. *Science* **10** (2019).

[7] Zayatz, L. *Journal of Official Statistics* **23**(2), 253 (2007).

[8] Ruggles, S., Fitch, C., Magnuson, D., and Schroeder, J. In *AEA papers and proceedings*, volume 109, 403–08, (2019).

[9] Garfinkel, S. L., Abowd, J. M., and Powazek, S. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 133–137, (2018).

[10] Van Riper, D., Kugler, T., and Schroeder, J. (2020). Website: https://www.nhgis.org/privacy-protected-demonstration-data.

[11] Kwok, K. O., Lai, F., Wei, W. I., Wong, S. Y. S., and Tang, J. W. *Journal of Infection* **80**(6), e32–e33 (2020).

[12] Santos-Lozada, A. R., Howard, J. T., and Verdery, A. M. *Proceedings of the National Academy of Sciences* (2020).

[13] Abowd, J. M. and Schmutte, I. M. *American Economic Review* **109**(1), 171–202 (2019).

[14] US Census Bureau. (2020). Library Catalog: www.census.gov Section: Government.

[15] Hooper, M. W., Nápoles, A. M., and Pérez-Stable, E. J. *Jama* (2020).