National Rural Health Association Policy Brief

The negative effect of differential privacy and synthetic data from the U.S. Census Bureau on rural health and communities

Author: Whitney Zahnd, PhD

Introduction

The U.S. Census Bureau collects demographic, socioeconomic, and business data on the U.S. population through its decennial census, ongoing American Community Survey (ACS), Current Population Survey, and other surveys and programs.[i] These data are made publicly available through the U.S. Census Bureau, state data centers, the Integrated Public Use Microdata Series (IPUMS) hosted at the University of Minnesota, and other platforms. Some examples include county-level estimates of median household income, racial/ethnic composition, or aggregated estimates of the number of registered nurses within a specific state.

Data from the U.S. Census Bureau are extensively used by rural communities, businesses, hospitals, clinics, nonprofit organizations, researchers, and others for describing rural communities and assessing health disparities.[ii] For example, a rural, nonprofit hospital may use county-level ACS data to describe the sociodemographic characteristics of their catchment area as part of a community health needs assessment. ACS data may also be presented in a nonprofit organization's grant application to estimate how many people may be served by their initiatives. Rural health workforce researchers may analyze Public Use Microdata Sample (PUMS) data to identify and quantify health care workforce shortages. This data is critical as it is used for the distribution of more than $675 billion in federal funding and to determine eligibility for social programming.[iii]

Analysis

With any survey data collection and public dissemination, participant privacy is critically important. Since 2000, the U.S. Census Bureau has performed data swapping at the census block level, the geographically smallest unit at which the Bureau provides data. Individual or household data are swapped across blocks within a census tract to protect privacy.[iv] With the increasing use of "big data" (large datasets with large volumes and variety of data used to identify trends and patterns), the U.S. Census Bureau recognized that the data could be linked with other sources, making it easy for census and survey participants to potentially be identified. In 2010, they began to develop and test new privacy techniques to better protect participant privacy in preparation for the 2020 census, including a new disclosure avoidance system.[v] Specifically, they developed an approach called differential privacy, which adds "noise" to the data to help protect privacy. In utilizing this approach, the U.S. Census Bureau aimed to balance privacy and accuracy.

Additionally, the U.S. Census Bureau has announced that by 2025, they will make its ACS data fully synthetic, meaning the data are statistically modeled to resemble an unreleased, confidential dataset. However, they have yet to announce a process for evaluating this change. This change will, for all intents and purposes, make ACS data analysis impossible to use for research purposes as unanticipated relationships between variables would be exceedingly difficult to assess.[vi]

There is an increasing concern, including from IPUMS, that the implementation of these differential privacy algorithms for the dissemination of census and other data will have a significant impact on the understanding of population dynamics and health.[vi] This impact will have a disproportionate effect on the

accurate description and assessment of rural, racial/ethnic minority populations. Studies show greater discrepancies among counties with increasing rurality and proportions of minoritized populations.[vii] This method has implications for measuring health in rural communities as well. For example, applying the differential privacy algorithms to historical data on infant mortality led to greater variation in rates in non-metropolitan counties, meaning data were likely to be more inaccurate.[viii]

Another study demonstrates increasing variation in mortality rates across the rural-urban continuum for rates calculated using data produced through differential privacy algorithms.[ix] These variations were even more pronounced when examining the intersection of racial/ethnic minority status and rurality. Beyond explicit health data concerns with this privacy algorithm, additional data accuracy concerns have been expressed in examining population migration. A recent study showed that more than half of counties would have inaccurate estimates for 5-year age groups, with inaccuracies in counties with fewer than 50,000 people.[x] This may be of particular concern and importance for rural governments and economic development professionals trying to assess population dynamics.

In addition to the studies demonstrating the problematic nature of these privacy algorithms, organizations such as the Federal-State Cooperative for Population Estimates, the National Council of State Legislatures, and the Southern Demographic Association have expressed concerns.[xi] [xii] Further, although eventually dropped, the state of Alabama had filed a lawsuit with concerns regarding the data's use for redistricting purposes.[xiii] Civil rights groups also expressed concerns with the fitness of the data for minority populations such as Latinos and Asian Americans, especially for assuring compliance with the Voting Rights Act.[xiv]

Policy recommendations

Concerns regarding data fitness for rural populations suggest these privacy approaches are inappropriate for providing sociodemographic information that is important for rural health. As such, NRHA recommends the following actions for the U.S. Census Bureau:

1. Provide greater transparency on the differential privacy process to enable researchers and others to better understand the extent of variation and lack of fitness of data, particularly for rural areas.
2. Work towards halting the use of differential privacy in U.S. Census data.
3. Reconsider the release of synthetic ACS data and consider other approaches that allow for relevant analysis while maintaining adequate privacy.
4. Provide an avenue for researchers to access data absent these differential privacy and synthetic constraints while still maintaining confidentiality. This may include enabling the use of existing research centers to access data either onsite or remotely.

Conclusion

Data from the U.S. Census Bureau are critical for rural health providers and researchers, as well as rural communities. The implementation of differential privacy approaches following the 2020 census will introduce substantial variation to data that will disproportionately affect its fitness for rural communities. It is imperative that mechanisms be put in place to simultaneously protect the privacy of individuals while ensuring researchers and others have access to accurate data to improve the health and well-being of rural communities.

[i] U.S. Census Bureau (2021a). What We Do. Retrieved July 20, 2021. https://www.census.gov/about/what.html
[ii] Blewett, L.A., Thiede Call K., Turner J., Hest, R. (2018). Data Resources for Conducting Health Services and Policy Research. *Annual Review of Public Health*. 39:437-52.
[iii] U.S. Census Bureau (2022). The Federal-State Cooperative for Population Estimates (FSCPE). Retrieved August 9, 2022. https://www.census.gov/programs-surveys/popest/about/fscpe.html
[iv] National Conference of State Legislatures. Differential Privacy for Census Data Explained. Retrieved July 20, 2021. https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx
[v] U.S. Census Bureau (2021). 2020 Census Data Products: Disclosure Avoidance Modernization. Retrieved July 20, 2021. https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html
[vi] IPUMS. Changes To Census Bureau Data Products. Accessed August 9, 2022. https://www.ipums.org/changes-to-census-bureau-data-products
[vii] Mueller, J.T., Santos-Lozada, A.R. (2022). The 2020 US Census Differential Privacy Method Introduces Disproportionate Discrepancies for Rural and Non-White Populations. *Population Research and Policy Review*. 41, 1417-1430.
[viii] Santos-Lozada, A.R. (2021). Changes in Census Data Will Affect Our Understanding of Infant Health. *Socius: Sociological Research for a Dynamic World.* https://doi.org/10.1177/23780231211023642.
[ix] Santos-Lozada, A.R., Howard, J.T., Verdery, A.M. (2020). How differential privacy will affect our understanding of health disparities in United States. *Proceedings of the National Academy of Sciences*. 117, (24): 13405-13412.
[x] Winkler, R.L., Butler, J.L., Curtis, K.J., Egan-Robertson, D. (2021). Differential Privacy and the Accuracy of County-Level Net Migration Estimates. *Population Research and Policy Review*. https://doi.org/10.1007/s11113-021-09664-5
[xi] Schneider M (2022). Researchers ask Census to stop controversial privacy method. Retrieved August 9, 2022 https://apnews.com/article/census-2020-us-bureau-government-and-politics-20e683c71eeb62ee4b7792d7d8530419
[xii] National Conference of State Legislatures. Differential Privacy for Census Data Explained. Retrieved July 20, 2021. https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx
[xiii] Mayes, B.R (2022). Alabama drops Census lawsuit on new privacy protection method. Retrieved August 9, 2022. https://www.washingtonpost.com/dc-md-va/2021/09/09/alabama-drops-census-lawsuit-privacy/
[xiv] Mexican American Legal Defense and Education Fund (2021). Preliminary Report: Impact of Differential Privacy & the 2020 Census of Latinos, Asian Americans and Redistricting. Retrieved July 20, 2021. https://www.maldef.org/wp-content/uploads/2021/04/FINAL-MALDEF-AAJC-Differential-Privacy-Preliminary-Report-4.5.2021-1.pdf