


RESEARCH BRIEF

The effect of differential privacy on Medicaid participation among racial and ethnic minority groups

Christoph F. Kurz PhD^{1,2}  | Adriana N. König MSc, MBR^{1,2}  |
Karl M. F. Emmert-Fees MSPH^{2,3}  | Lindsay D. Allen PhD MA⁴ 

¹Munich School of Management and Munich Center of Health Sciences, Ludwig-Maximilians-Universität München, Munich, Germany

²Institute of Health Economics and Health Care Management, Helmholtz Zentrum München, Neuherberg, Germany

³Department of Sport and Health Sciences, Technical University of Munich, Munich, Germany

⁴Department of Emergency Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

Correspondence

Christoph F. Kurz, Munich School of Management and Munich Center of Health Sciences, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 Munich, Germany.
Email: c.kurz@lmu.de

Abstract

Objective: To investigate how county and state-level estimates of Medicaid enrollment among the total, non-Hispanic White, non-Hispanic Black or African American, and Hispanic or Latino/a population are affected by Differential Privacy (DP), where statistical noise is added to the public decennial US census data to protect individual privacy.

Data Sources: We obtained population counts from the final version of the US Census Bureau Differential Privacy Demonstration Products from 2010 and combined them with Medicaid enrollment data.

Study Design: We compared 2010 county and state-level population counts released under the traditional disclosure avoidance techniques and the ones produced with the proposed DP procedures.

Data Collection/Extraction Methods: Not applicable.

Principal Findings: We find the DP method introduces errors up to 10% into counts and proportions of Medicaid participation rate accuracy at the county level, especially for small subpopulations and racial and ethnic minority groups. The effect of DP on Medicaid participation rate accuracy is only small and negligible at the state level.

Conclusions: The implementation of DP in the 2020 census can affect the analyses of health disparities and health care access and use among different subpopulations in the United States. The planned implementation of DP in other census-related surveys such as the American Community Survey can misrepresent Medicaid participation rates for small racial and ethnic minority groups. This can affect Medicaid funding decisions.

KEYWORDS

census, confidentiality/privacy issues, Medicaid, racial and ethnic differences in health and health care

What is known on this topic

- The census is the primary source of statistical information about the US population, and the accuracy of its data is critical to health services research.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Health Services Research* published by Wiley Periodicals LLC on behalf of Health Research and Educational Trust.

- The Census Bureau will infuse additional noise into its 2020 census data products to protect individuals' privacy.
- Such privacy measures are suspected to disproportionately affect underrepresented populations and ethnic and racial subgroups.

What this study adds

- Increased privacy measures in the census can affect Medicaid participation rate accuracy for the non-Hispanic Black or African American and Hispanic or Latino/a population at the county-level with lower population counts by introducing errors up to 10%.
- State-level accuracy of Medicaid participation rates is rarely affected by DP.
- The implementation of DP in census-related surveys can misrepresent Medicaid participation rates for small racial and ethnic minority groups and affect funding decisions.

1 | INTRODUCTION

The constitutionally mandated decennial census survey, which includes the annual American Community Survey (ACS), is the foremost source of demographic and statistical information about the United States' population. It determines each state's representation in congress and guides the allocation of \$1.5 trillion in federal funds per year.¹ The survey has particular importance in the health policy sphere for two reasons. First, census data are used in thousands of publications annually in social science and policy research, evaluating and informing data-driven health policy.² Health services researchers, in particular, rely on accurate census statistics for measuring health disparities and tracking health care access and use among different subpopulations.³⁻⁵ Second, federal, state, tribal, and local governments use ACS statistics to help with decision making and to allocate over \$675 billion each year back to communities.⁶ Concerningly, recent changes to privacy security measures for the census may have jeopardized the quality and utility of these important US data resources.²

Balancing the need for accurate census survey data with an equally essential obligation to protect respondents' privacy is a trade-off that has come to a head in recent years.⁷ Some argue that the confidentiality of census data is susceptible to “database reconstruction,” a process for inferring individual-level responses from tabular data.² To protect the privacy of the 2020 census participants from these techniques, the US Census Bureau includes additional privacy security methods based on the concept of *Differential Privacy* (DP). Although controversial among researchers,² DP is a promising concept that can protect respondents' privacy against a wide range of attacks, such as reidentification and record linkage.^{8,9}

The DP approach is based on adding a controlled amount of statistical noise to the data in a way that protects the information of any single respondent against identification.¹⁰ While prior work has found that census estimates generated via DP data may be accurate for aggregate population statistics, they are subject to considerable error for mortality rate estimates of subgroups and ethnic and racial minority groups.¹¹

If these errors for members of racial and ethnic minority groups persist in other areas of the survey, there may be important implications for disparities. For example, one of the key questions asked in the ACS is about respondents' health insurance status.¹² The Census

Bureau uses these data to create statistics about the percentage of people covered by health insurance and the sources of health insurance, which are then used to plan government programs, determine eligibility criteria, and encourage people to participate in health insurance programs.¹² For members of racial and ethnic minority groups, who face longstanding disparities in health insurance coverage and health, these programs are essential for accessing health care and protection from high health care costs.¹³

Differential privacy is very likely to be applied to the ACS as well.¹⁴ The purpose of this paper is to estimate the extent to which DP may influence ACS statistical accuracy for topics that have implications for health disparities. We used the example of Medicaid participation rates.

Medicaid provides public health insurance coverage to millions of Americans and plays a substantial role in improving access to care, health outcomes, financial security, and other factors essential for reducing disparities. Without accurate data on who is participating in Medicaid, funds and programs may be misallocated away from areas most in need.

Using the Census Bureau's demonstration dataset containing 2010 decennial counts produced with proposed DP and traditional techniques, we evaluated how the implementation of DP affects estimates of Medicaid participation rates at the county and state levels for the non-Hispanic White, non-Hispanic Black or African American, and Hispanic or Latino/a population. We anticipate that results from this study will inspire more work on the relative drawbacks versus benefits of incorporating DP into the ACS.

2 | METHODS

2.1 | Data

We used two data sources to construct Medicaid participation rates, which we defined using the number of Medicaid enrollees as the numerator and population counts as the denominator. For the numerator, we acquired 2010 Medicaid enrollment data directly from the Centers for Medicare & Medicaid Services (CMS) via the Medicaid Statistical Information System and the Kaiser Family Foundation.¹⁵

Enrollment is defined as the number of individuals who are enrolled in Medicaid at any time during the federal fiscal year. We included all 1735 counties for which complete information on race and ethnicity (i.e., non-Hispanic White, non-Hispanic Black or African American, Hispanic or Latino/a, and Other populations) was available (out of 3195). The group “Other” included the following races and ethnicities: American Indian and Alaska Native alone; Asian alone; Native Hawaiian and Other Pacific Islander alone; Some Other Race alone; and Two or More Races. If data for the year 2010 were not reported, we used data from subsequent years up to 2013. For the state of Idaho, data for the non-Hispanic Black or African American category were missing for all years and could not be included.

We chose to construct these measures at the county and state levels because these are two geographical units of analysis that play a large role in both health services research and health policy. Medicaid policies are typically determined at the state level, while counties operate hospitals, nursing homes, behavioral and mental health care, testing services, and other public health services.¹⁶

For the denominator, we obtained county and state-level data from the 2010 US census that were released using traditional disclosure avoidance techniques, as well as those produced using the DP procedures.¹⁷ Traditional disclosure techniques applied to the 2010 census include household-swapping, rounding, and top- and bottom-swapping.¹⁸ The 2010 Demonstration Data Products introduces DP to the census data. Using the 2010 census data thus provides the opportunity to compare traditional disclosure techniques with DP. Since the 2020 census data only contains DP, this comparison is not possible for the more recent data set.

The DP concept was developed by¹⁹ and leverages random noise—random variation around the true value—to obscure individual-level data. This conserves the statistical properties of the data but makes individual information hard to identify.²⁰

An essential component of DP is the privacy loss parameter, usually denoted by ϵ . This parameter determines the amount of noise added to the computation, and therefore defines what can be learned about an individual as a result of their private information being included in a differentially private analysis. Lower ϵ means more noise is added, increasing privacy but decreasing statistical accuracy. The census DP counts were produced under a global ϵ of 19.61, which includes an ϵ of 17.14 for persons and an ϵ of 2.47 for housing units.²¹ We used the final version of the US Census Bureau Differential Privacy Demonstration Products, released in August 2021. Data were accessed via IPUMS-NHGIS.²² Population counts were available for the same race and ethnicity categories used by CMS.

2.2 | Analysis

We defined Medicaid participation counts for county c and race and ethnicity r as $A(c,r)$. We defined population counts separately for the traditional census approach $P(c,r,trad)$ and the novel DP approach, denoted as $P(c,r,DP)$.

This allowed calculation of the Medicaid participation rate for county c and race and ethnicity r , $m(c,r,j)$, as the numerical.

Medicaid participants count divided by the population,

$$m(c,r,j) = \frac{A(c,r)}{P(c,r,j)},$$

where j refers to either *trad* or *DP*. The absolute difference in the participation rate is then defined as

$$d_a = |m(c,r,trad) - m(c,r,DP)|.$$

The relative difference in the participation rate between traditional and DP Census data is

$$d_r = \frac{|m(c,r,trad) - m(c,r,DP)|}{\left(\frac{|m(c,r,trad) + m(c,r,DP)|}{2}\right)}$$

We multiplied participation rates by 100 to get percentages, then calculated the absolute and relative difference between the two. The absolute difference is a simple measure of the absolute deviation of two values, independent of size. The relative difference, which compares two quantities while considering the sizes of the two measures being compared, is the preferred quantitative indicator where the two outcomes are expected to be the same. For example, if Medicaid participation among Hispanic or Latino/a individuals is 5% using traditional population counts and 6% using DP population counts, the absolute difference is $|5\% - 6\%| = 1\%$. The relative difference is $|5\% - 6\%| / ((5\% + 6\%) / 2) = 18\%$. We applied the same analytical procedure to the state-level data.

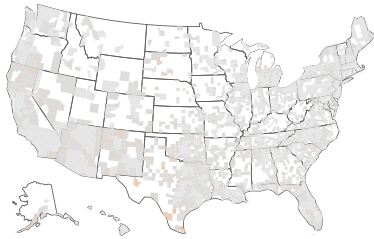
3 | RESULTS

Figure 1 illustrates the differences in traditional- versus DP-defined race and ethnicity-specific Medicaid participation rates at the county level, expressed as absolute and relative differences. We find that DP introduces higher absolute and relative errors into racial and ethnic minority populations compared to larger populations. Among these minority groups, the largest absolute differences appear among the non-Hispanic Black or African American population. Counties in the eastern and southern United States—where the average county population size is much smaller than other areas—were most affected. For example, in the 2010 census, Gilmer County, Georgia, has a non-Hispanic Black or African American population count of 98, but the DP count is 55. For the non-Hispanic White population, the difference is usually small and negligible (see Table 1 for more examples). The large discrepancy in differences between counties is further illustrated by the coefficient of variation (defined as [standard deviation]/mean). It is 242% for non-Hispanic White, 310% for non-Hispanic

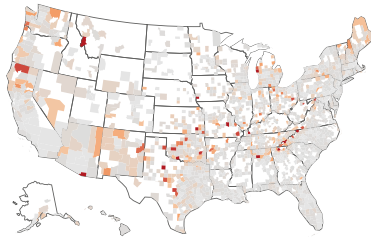
Medicaid Participation Rates (DP/Trad.)

Absolute Difference in Percent

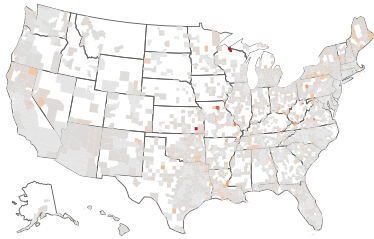
Non-Hispanic White Pop.



Non-Hispanic Black and AA Pop.



Hispanic and Latino/a Pop.



Other Pop.

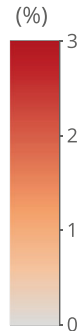
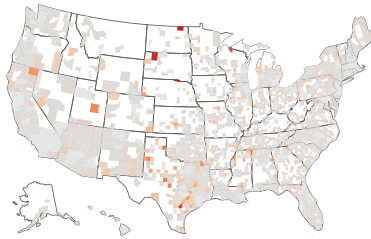
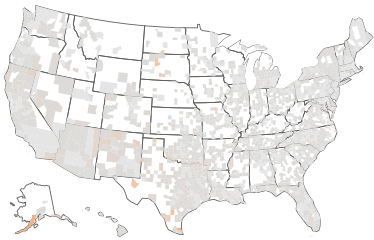


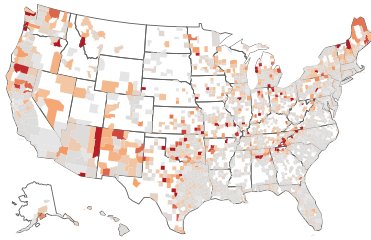
FIGURE 1 Effect of Differential Privacy on Medicaid participation rates at the county level. Darker red color indicates larger differences. White areas on the map indicate missing data. AA, African American; DP, differential privacy; Pop., population; Trad., traditional disclosure methods [Color figure can be viewed at wileyonlinelibrary.com]

Relative Difference in Percent

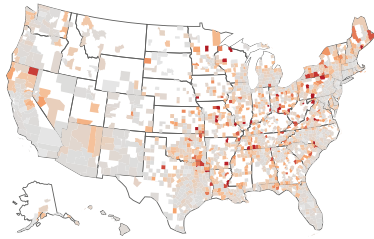
Non-Hispanic White Pop.



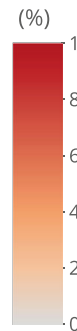
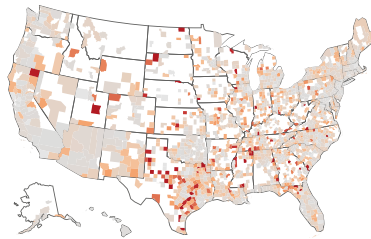
Non-Hispanic Black and AA Pop.



Hispanic and Latino/a Pop.



Other Pop.



Black or African American, 441% for Hispanic or Latino/a, and 267% for other populations.

In counties with very homogeneous racial and ethnic populations that are coupled with low total population size, DP errors are more evident. For example, Iron County, Wisconsin, has a total population of 5916, where almost 98% identify as non-Hispanic White; the non-Hispanic Black or African American population contains only three individuals, and the Hispanic or Latino/a population counts 35. However, the DP census lists 6 non-Hispanic Black or African Americans and 46 Hispanic or Latino/a individuals, respectively. Combined with

high Medicaid participation, this results in large differences. Such shifts make Medicaid participation rates for certain combinations of county and race and ethnicity hard to interpret.

As the population denominator increases, the errors introduced by the DP diminish. This is illustrated in Figure 2, which shows calculations at the state level (please note the different color ranges compared with Figure 1). At the state level, the maximum absolute difference is 0.1 percentage points, while the maximum relative difference is less than 0.5%. Even for states with little racial and ethnic diversity, such as Montana or Wyoming, the effect of DP has a minor

TABLE 1 2010 Census population counts using traditional and Differential Privacy disclosure methods for selected counties

County	Total Pop.	Total Pop. DP	NH White	NH White DP	NH Black or AA	NH Black or AA DP	Hispanic or Latino/a	Hispanic or Latino/a DP	Other	Other DP
Gilmer County, GA	28,292	28,292	25,078	25,109	98	55	2677	2680	439	448
Crawford County, MO	24,696	24,694	23,804	23,832	64	48	365	347	463	467
Marshall County, KY	31,448	31,450	30,669	30,680	48	41	350	354	381	375
Wilson County, KS	9409	9408	8866	8920	30	34	217	160	296	294
Mills County, IA	15,059	15,061	14,390	14,386	57	49	359	396	253	230
Ashe County, NC	27,281	27,276	25,420	25,411	148	121	1311	1349	402	395
Stonewall County, TX	1490	1491	1206	1216	38	34	209	194	37	47
Swain County, NC	13,981	13,982	9168	9155	75	95	540	510	4198	4222
Winston County, AL	24,484	24,483	23,237	23,245	115	97	639	655	493	486
Iron County, WI	5916	5919	5772	5770	3	6	35	46	106	97
Wilson County, KS	9409	9408	8866	8920	30	34	217	160	296	294
Macon County, MO	15,566	15,570	14,735	14,789	349	349	150	122	332	310
Taylor County, WV	16,895	16,896	16,377	16,382	125	110	143	122	250	282
Ste. Genevieve County, MO	18,145	18,146	17,607	17,590	117	106	149	181	272	269
Appomattox County, VA	14,973	14,975	11,483	11,498	2998	3014	167	142	325	321
Boyd County, NE	2099	2100	2024	2033	1	0	33	35	41	32
Cavalier County, ND	3993	3992	3890	3893	4	1	24	30	75	68
Perkins County, SD	2982	2984	2883	2886	2	2	20	28	77	68
Karnes County, TX	14,824	14,827	5956	5931	1351	1344	7376	7429	141	123

Abbreviations: AA, African American; DP, differential privacy; NH, non-Hispanic; Pop., population.

influence on Medicaid participation rate accuracy. According to the 2010 census, Montana has a non-Hispanic Black or African American population count of 3743; this changes to 3731 in the DP census demonstration file, resulting in only a small relative difference, even with high Medicaid participation among non-Hispanic Black or African American individuals in this state.

4 | DISCUSSION

In this paper, we have shown that applying the DP algorithm to the 2010 US Decennial Census results in a misrepresentation of Medicaid participation rates among already-marginalized racial and ethnic groups. Distortions from the DP approach were most evident when examining data at the county level and in states with less racial and ethnic diversity. It has to be noted that county-level data were only available for 1735 out of 3195 counties. Medicaid participation rates among certain combinations of county and race and ethnicity differed between DP and non-DP counts on an absolute scale by more than three percentage points, while relative difference sometimes exceeded 10%. Importantly, non-Hispanic White individuals are the only ethnic and racial subgroup for which the DP algorithm accurately captured Medicaid participation rates.

This finding may have important implications for health policy. Although our study used data from the 2010 Decennial Census (the

only data set for which DP-adjusted data are available), health insurance coverage information was not collected as part of this survey. Rather, this information is collected via the annual ACS, to which DP will likely be applied in the future.¹⁴ The ACS insurance data play in planning government programs, resource allocation, and policy evaluation and tracking. Therefore, policy makers and researchers alike need to be aware of DP's potential pitfalls, especially when it comes to racial and ethnic groups most affected by insurance and related health disparities.

Our results are in line with previous findings showing the distorting effect of DP census data on COVID rates²³ and mortality.¹¹ In contrast,²⁴ find that census tract estimates of mortality rates and inequities are not affected by DP. Their analysis, however, is limited to the state of Massachusetts, a populous state with more diverse demographics. We also did not find an effect for Massachusetts state and counties.

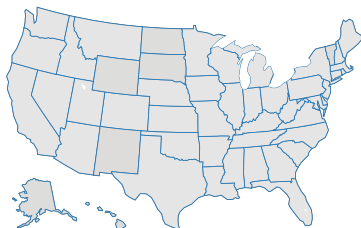
Beyond insurance, the application of DP to the 2020 census data may compound other equity concerns about that year's survey, including a decision to cut the reporting period short at a time when 60 million households still needed to be recorded. Notably, those households were hard-to-count populations, disproportionately people of color.²⁵

Even setting aside these potential equity implications, the use of DP in the 2020 census remains highly controversial. On the one hand, protecting respondent privacy is a chief concern. NPR details

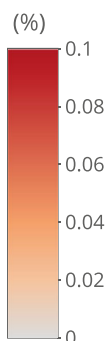
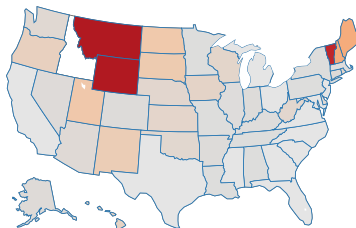
Medicaid Participation Rates (DP/Trad.)

Absolute Difference in Percent

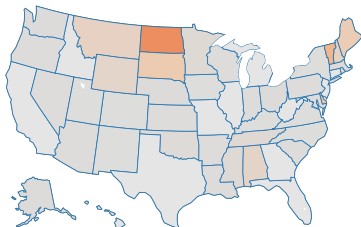
Non-Hispanic White Pop.



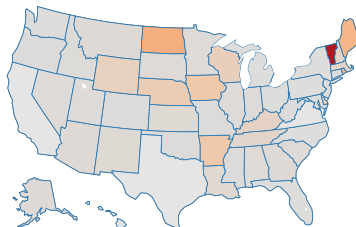
Non-Hispanic Black and AA Pop.



Hispanic and Latino/a Pop.

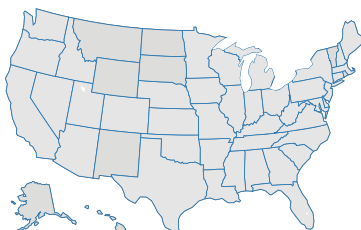


Other Pop.

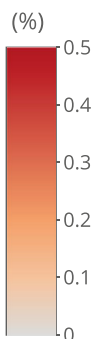
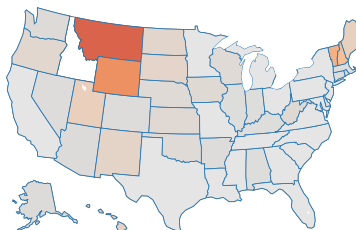


Relative Difference in Percent

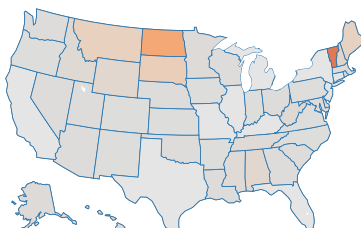
Non-Hispanic White Pop.



Non-Hispanic Black and AA Pop.



Hispanic and Latino/a Pop.



Other Pop.

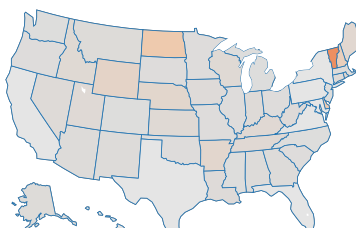


FIGURE 2 Effect of Differential Privacy on Medicaid participation rates at the state level. Darker red color indicates larger differences. White areas on the map indicate missing data. AA, African American; DP, differential privacy; Pop., population; Trad., traditional disclosure methods [Color figure can be viewed at wileyonlinelibrary.com]

a recent security exercise in which Census Bureau researchers were able to reconstruct a complete set of records for every person included in the 2010 census numbers.²⁶ After cross-referencing the reconstructed data with purchased commercial records, they were able to reidentify 52 million people by name. The researchers estimated that more than half of the population included in the 2010 census could be identified if attackers were able to procure even more commercial data records. On the other hand, the reconstruction described represents an extraordinarily resource-intensive process, unlikely to occur outside the Bureau's laboratory, leading some to proclaim that "differential privacy goes far beyond what is necessary to keep data safe under census law and precedent."²

As our study illustrates, the introduction of DP to future census products may introduce data inaccuracies, especially among already-marginalized populations. Researchers and policy makers should be aware of this potential pitfall when designing studies and allocating resources in the future.

ACKNOWLEDGMENTS

None.

ORCID

Christoph F. Kurz  <https://orcid.org/0000-0001-9498-8002>

Adriana N. König  <https://orcid.org/0000-0002-8884-3533>

Karl M. F. Emmert-Fees  <https://orcid.org/0000-0003-2330-1079>

Lindsay D. Allen  <https://orcid.org/0000-0003-1723-9248>

REFERENCES

- Brown N. Special report: 2020 US census plagued by hacking threats, cost overruns. 2019. <https://www.reuters.com/article/us-usa-census-technology-specialreport-idUSKBN1Y81H8>. Accessed January 20, 2022.
- Ruggles S, Fitch C, Magnuson D, Schroeder J. Differential privacy and census data: Implications for social and economic research. *AEA papers and proceedings*. vol. 109; 2019. pp. 403-408.
- Thomas SB, Quinn SC, Butler J, Fryer CS, Garza MA. Toward a fourth generation of disparities research to achieve health equity. *Annu Rev Public Health*. 2011;32:399-416.
- Derosé SF, Contreras R, Coleman KJ, Koebnick C, Jacobsen SJ. Race and ethnicity data quality and imputation using US census data in an integrated health system: the Kaiser Permanente Southern California experience. *Med Care Res Rev*. 2013;70(3):330-345.
- Martino SC, Mathews M, Agniel D, et al. National racial/ethnic and geographic disparities in experiences with health care among adult Medicaid beneficiaries. *Health Serv Res*. 2019;54:287-296.
- US Census Bureau. American community survey information guide. 2017. <https://www.census.gov/programs-surveys/acs/library/information-guide.html>
- Lane J, Schur C. Balancing access to health data and privacy: a review of the issues and approaches for the future. *Health Serv Res*. 2010; 45(5p2):1456-1467.
- Dwork C, Smith A, Steinke T, Ullman J. Exposed! A survey of attacks on private data. *Annu Rev Stat Appl*. 2017;4:61-84.
- Rocher L, Hendrickx JM, De Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun*. 2019;10(1):1-9.
- Ficek J, Wang W, Chen H, Dagne G, Daley E. Differential privacy in health research: a scoping review. *J Am Med Inform Assoc*. 2021; 20(10):2269-2276.
- Santos-Lozada A, Howard J, Verdery A. How differential privacy will affect our understanding of health disparities in the United States. *Proc Natl Acad Sci U S A*. 2020;117:13405-13412.
- US Census Bureau. Why We Ask Questions About... Health Insurance Coverage. (n.d.). <https://www.census.gov/acs/www/about/why-we-ask-each-question/health/>. Accessed April 26, 2022.
- Kaiser Family Foundation. Health Coverage by Race and Ethnicity, 2010-2019. 2021. <https://www.kff.org/racial-equity-and-health-policy/issue-brief/health-coverage-by-race-and-ethnicity/>. Accessed April 26, 2022.
- Dajani, AN, Lauger AD, Singer PE, et al. The modernization of statistical disclosure limitation at the U.S. Census Bureau. 2020. <https://www.census.gov/library/working-papers/2020/adrm/modernization-statistical-disclosure-limitation.html>
- Mac Taggart P, Foster A, Markus A. Medicaid Statistical Information System (MSIS): A Data Source for Quality Reporting for Medicaid and the Children's Health Insurance Program (CHIP) Perspectives in Health Information Management 2011;8(Spring).
- National Association of Counties. Counties Matter. 2019. <https://www.naco.org/resources/featured/counties-matter>. Accessed April 26, 2022.
- Van Riper D, Kugler T, Schroeder J. IPUMS NHGIS Privacy-Protected 2010 Census Demonstration Data, Version 20210608 [Database]. Minneapolis, MN: IPUMS. 2021.
- Zayatz L. Disclosure avoidance practices and research at the US Census Bureau: an update. *J Off Stat*. 2007;23(2):253.
- Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*. Springer; 2006:265-284.
- Kurz C. Understanding differential privacy. *Significance*. 2021;18(3): 24-27. doi:10.1111/1740-9713.01528
- US Census Bureau. Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results; 2021. <https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html>
- Manson S, Schroeder J, Van Riper D, Kugler T, Ruggles S. IPUMS National Historical Geographic Information System: Version 15.0 [Dataset]. Minneapolis, MN: IPUMS. 2020.
- Hauer ME, Santos-Lozada AR. Differential privacy in the 2020 census will distort COVID-19 rates. *Socius*. 2021;7:1-6.
- Krieger N, Nethery RC, Chen JT, et al. Impact of differential privacy and census tract data source (decennial census versus American community survey) for monitoring health inequities. *Am J Public Health*. 2021;111(2):265-268.
- Stanton Z. 'A brush with catastrophe': Inside the 2020 Census Meltdown. 2020. <https://www.politico.com/news/magazine/2020/09/10/census-2020-trump-pandemic-problem-miscount-412040>. Accessed January 21, 2022.
- Wang HL. For The US Census, Keeping Your Data Anonymous And Useful Is A Tricky Balance; 2021. Available from: <https://www.npr.org/2021/05/19/993247101/for-the-u-s-census-keeping-your-data-anonymous-and-useful-is-a-tricky-balance>. Accessed January 21, 2022.

How to cite this article: Kurz CF, König AN, Emmert-Fees KMF, Allen LD. The effect of differential privacy on Medicaid participation among racial and ethnic minority groups. *Health Serv Res*. 2022;57(Suppl. 2):207-213. doi:10.1111/1475-6773.14000