Risks and Rewards of Demographic Data Collection: How Effective Data Privacy Can Promote Health Equity

This four-part report series is a joint project of the National Health Law Program, Disability Rights Education and Defense Fund, Justice in Aging, Movement Advancement Project, and Race Forward

Striking the Balance: Approaches to Racial Equitable Data Collection that Protect Privacy in Health

Patrick L. Mason, PhD.
Race Forward

Striking the Balance: Approaches to Racial Equitable Data Collection that Protect Privacy in Health

By Patrick L. Mason

On his first day in office, President Biden signed the historic Executive Order on Advancing Racial Equity and Support for Underserved Communities Through the Federal Government, recognizing that:

> *Many Federal datasets are not disaggregated by race, ethnicity, gender, disability, income, veteran status, or other key demographic variables. This lack of data has cascading effects and impedes efforts to measure and advance equity. A first step to promoting equity in Government action is to gather the data necessary to inform that effort.* [1]

This order was reinforced with the additional Executive Order, Further Advancing Racial Equity and Support for Underserved Communities Through the Federal Government which solidifies the Administration's commitment to racially equitable data practices across the federal government.[2]

The focus on collecting demographic data is vital for racial equity work in health care entities, including the agencies that administer health care programs such as CMS, which administers Medicaid, CHIP, the healthcare.gov Marketplace, and Medicare. Agencies' data collection policies and processes drive a multitude of decisions that affect people's lives and well-being every day. The lens brought to those practices often determine the outcome. In the simplest terms, you cannot get the right answers if you do not ask the right questions. As the New York City Health Department has noted:

> *[D]ata are a social construct, made for and by people, and the ways that data are collected and used in public health practice is an act of power with profound consequences.*[3]

When these policies and processes ignore existing racial inequities and the systemic factors that drive it, even very rigorous evidence-based decisions can perpetuate harm and reinforce structural racism. Indeed, how data is collected and used can serve to deepen existing racial disparities and inequities. With the right insights and approaches,

racially equitable data strategies can also begin to close long standing and harmful gaps and serve as a key driver to advance racial equity.

To improve racial equity outcomes using data requires change at all levels, including data privacy approaches. Methods that balance access to data and preserve individual privacy have become an increasing concern as they generally do not explicitly consider racial equity. Without a racial equity lens, researchers can create unintended consequences by either needlessly hampering the utility of the data in significant ways or distributing privacy risks inequitably.

For example, according to the Migration Policy Institute, in 2019 only 4.7% of residents of Vermont were foreign born**.** [4] And out of that 4.7 % only 14.3% were Black, leaving the population of Black foreign born residents – many of them refugees – a very small population. Some data privacy and confidentiality methods would remove references to race or obscure more specific ethnicities or location to protect privacy. However, obscuring or removing reference to this population makes seeing and addressing the needs of that community far more difficult. Given communities of color especially those with lower incomes, are both underserved and more susceptible to privacy attacks, agencies, programs, and health care entities need strategies that both protect privacy and ensure that important racial equity outcomes can be achieved.

This paper discusses the current practices and policies of the Centers for Medicare & Medicaid Services (CMS) to protect individual privacy and prevent re-identification of individuals in its datasets. CMS needs administrative data, and in particular, demographic data, to monitor and address health equity in its programs. However, with advancing technology and access to information comes increased risks of re-identification. Particularly for those interested in targeting individuals—for fraud, discrimination, exclusion, or other negative purposes—combination of publicly available data with private datasets poses a threat to individual privacy. CMS must pay increased attention to the users of its data as well as the methods of protecting its data in order to uphold the integrity of its programs.

I. Data Utility vs. Risk of Disclosure

A tradeoff exists between data utility, that is, the usefulness of administrative data for functions such as service delivery, program management, research, and analysis, and the risk of re-identification, that is, the riskiness association with collecting, distributing, and analyzing data containing sensitive personal information.[5] Administrative data can contain sensitive information on individuals, families, and organizations. Medicaid

applicants, for example, supply their full name, date of birth, sex, Social Security number, home address (street, city, county, state, zip code), mailing address, phone numbers (work and cell), email address, preferred spoken and written (other than English), plans to file federal income tax, tax status and name of spouse, names of tax dependents, whether claimed as a dependent on someone else's taxes and the name of that person, pregnancy status, need for health care coverage, health condition and limitations on activities in daily living or residence in a nursing, citizenship status, veteran status, and immigration document numbers.[6]

Demographic data, a subset of administrative data, can include collection of an individual's race, ethnicity, language, sex, sexual orientation, gender identity, disability and age. Agencies use these data for program operations and monitoring. Health care providers, such as hospitals or clinics, need detailed information to maintain accurate records, measure quality of care, and provide proper treatment. Additionally, users of federal administrative data include researchers employed by for-profit businesses, public policy organizations, academic institutions, and federal, state, and local governmental agencies. These researchers seldom use the bulk of identifying information included in federal administrative data. However, some persons, organizations, and governments may wish to use this personally identifying information for illegitimate purposes.

Concerns about the collection of demographic data include the potential for re-identification, a privacy violation. Re-identification occurs when a data snooper (one who receives or uses data through inappropriate means) has successfully identified individual respondents in a dataset. Scholars Julia Lane and Claudia Schur posit: "[t]he harm associated with re-identification can be financial (disclosure might lead to denial of insurance coverage, job loss, or lack of job offer) and psychosocial (*e.g.*, revelation of personal information leading to stigma in a social or work circle, or loss of reputation resulting in isolation or difficulty obtaining employment."[7]

A. Conceptualization of Privacy Changes

Consider the upward sloping lines in Figure 1. The line labeled "current method" represent the current set of policies and procedures for collecting, distributing, and analyzing administrative data. As data utility increases, the risk of disclosing personal identity increases. The point $r_0 > 0$ indicates that there is an inherit risk in collecting sensitive data, even if the data are not distributed, analyzed, or intensely used for program management and oversight. The risk arises from the possibility that the data

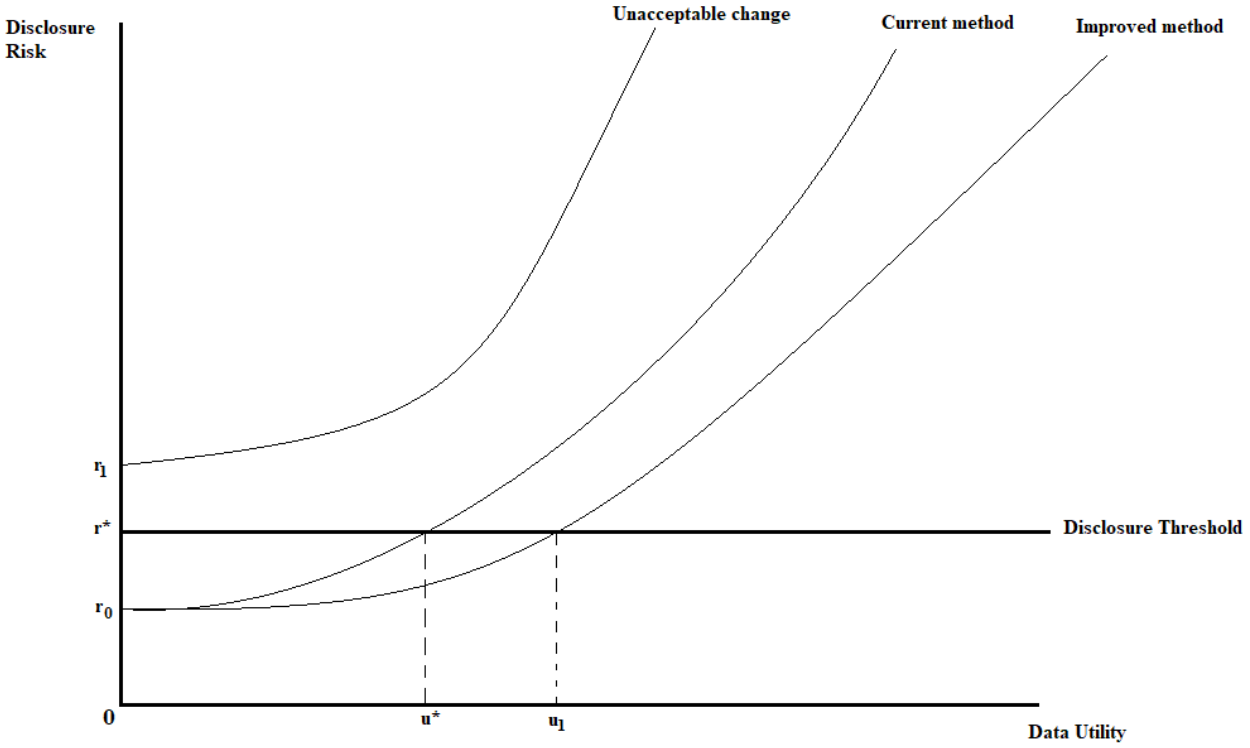
may become the property of a data snooper or the persons with legitimate access and good intentions may accidentally use the data in an inappropriate manner.

The upward slope of "current method" occurs because as data become more useful, the risk of re-identification increases. Factors that increase data utility include:

- distributing the data across broad set of localities;
- distributing the data to a broad set of users;
- making it easy to identify and individuals within the data (especially when the individuals may occur multiple times);
- making it easy to identify the census block and other locational aspects of the data;
- making it easy to identify specific institutions within the data (especially when those institutions occur multiple times); and
- making it easy to merge the data with other datasets, including a wide variety of individual characteristics and socioeconomic outcomes, including a wide variety of characteristics of organizations and their personnel, etc.

But, increases in the quality and quantity of the components of data utility are also the same items that increase the probability of re-identification.

Figure 1. Hypothetical application of Duncan-Roehrig risk-utility confidentially map



The disclosure threshold is r*, the maximally acceptable level of risk of re-identification. At all risk levels about r*, the benefit of data utility is less than the cost of re-identification: it is useful at this point to make changes in data distribution and usage. Suppose the current data utility - disclosure risk combination is (u*,r*). Some data policies represent "unacceptable change," they raise the level of disclosure risk for all levels of data utility – pushing the risk-utility line to the left. For example, merging Medicaid application data with census data, such as the American Community Survey, without making offsetting changes in data distribution and use is an unacceptable change. The merger of administrative data with survey data would establish a richer dataset, but it would also make it easier to re-identify individuals, families, and organizations.

There are other data policies representing an "improved method," they lower or do not raise the level of disclosure risk for all levels of data utility – pushing the risk-utility line to the right. For example, merging Medicaid application data with census data, such as the American Community Survey, but removing Social Security numbers or other unique identifiers, or increasing the oversight and management of the distribution of and access to the combined data. This merger of administrative data with survey data establishes a richer dataset that also does not make it easier to re-identify individuals, families, and organizations. The remainder of this report discusses policies that facilitate improved methods of use of administrative data, that is, these are policies that increase data utility without increasing the risk of re-identification, moving from (r*,u*) to (r^*,u_1).

II. Methods that Balance Improved Administrative Data Utility with Decreased Risk of Disclosure

A. De-identification of Data

The Office for Civil Rights (OCR), U. S. Department of Health and Human Services, endorses de-identification of administrative data as a procedure to reduce the risk of disclosure without impairing utility.[8] The "safe harbor method" (SHM) is one approach for de-identifying data. As described in the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule, SHM removes identifiers of individuals, as well as their relatives, employers, or household members.[9] Also, per the Privacy Rule for use and disclosure of protected health information, covered entities "must have no actual knowledge that the remaining information could be used to identify the individual."[10]

"Expert determination" is a second approach for de-identifying administrative data. Expert determination requires,

> *"A person with appropriate knowledge of and experience with generally acceptable statistical and scientific principles and methods for rendering individual information not individually identifiable: 1) applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and, 2) documents the methods and results of the analysis that justify such a determination."*[11]

There is no specific credential required for determining an expert, though OCR reviews and evaluates credentials. There is no precise statistical measure for determining when the risk is very small. This determination is based on the opinion of experts with experience in data analysis and security.

B. Data with Small Numbers

Small numbers within a particular category of data substantially increase the risk of re-identification. It is imperative then to obfuscate the count number while preserving data utility. For example, a Medicaid claims dataset could list one individual of a particular race who received services at a particular provider. Were that information disclosed, it would be relatively easy to identify that individual based on the characteristics described. For populations, current CMS policy is to reveal data counts only when the number in the dataset is less than 11. When the count is between 1 and 10, CMS will obfuscate the data.[12] When CMS does so, it also changes all rates and proportions that depend on that count.

Some states go further than CMS policy to protect privacy and prevent re-identification in Medicaid data. For example, the state of Washington further advises when the count is "unknown," do not change that designation.[13] It also instructs not to obfuscate the dataset when the number in it is zero, unless a zero in one cell reveals information about every other cell. For example, suppose a women's health clinic is queried on the number of people utilizing the clinic during 2021 who did not have an abortion. In this case, reporting that zero patients did not have an abortion would indicate that 100% of the people using the clinic had an abortion. Anyone who knows or who had observed a person using the clinic would know her medical history on this procedure.

Appendix A discusses several further methods for obfuscating numbers in datasets to protect the privacy of individuals, including suppressing the count of individuals in a particular category, "bottom-coding" and "top-coding," and aggregating data.

C. Public Use v. Restricted Use Data

All CMS data are not equally accessible to the public.[14] The most accessible files are Public Use Files (PUFs), that is, Non-Identifiable Data Files. PUFs are edited and stripped of all information that can re-identify individuals, such as names, demographic identifiers, phone numbers, income, Social Security numbers and addresses. PUFs are free and easy to download.[15]

CMS offers users two methods for obtaining access to restricted data. Researchers may use the Virtual Research Data Center (VRDC) for timely, efficient, and cost-effective access to Medicaid and Medicaid administrative data.[16] Academic researchers, innovators, and for-profit organizations are provided access to the VRDC through a virtual desktop. Further, researchers must be residents of the U. S., connect from a U. S. registered internet protocol address, and have a broadband internet connection. Reports and results of statistical analysis may be downloaded to a local computer. The VRDC satisfies all of CMS's privacy and security requirements, and it provides access to Research Identifiable Files (RIF). Researchers may upload external data to their workplace to analyze with CMS data.

For-profit organizations and innovators[17] are limited to using the VRDC, where all research is on-site, but academic users have a physical data option – they can access a Limited Data Set (LDS) if they complete a data use agreement (DUA).[18] LDS files contain beneficiary level protected health information, information that a data snooper may wish to use to re-identify beneficiaries. (See Appendix B for a comparison of identifying information contained in each dataset).

Research Identifiable Files (RIFs) are the most restricted data. RIF contain beneficiary level protected health information (PHI).[19] Researchers seeking to use RIFs must complete a DUA and have their request for data reviewed by CMS's Privacy Board. The organization sponsoring the researcher or innovator must submit a detailed self-attestation questionnaire that is based on the CMS Acceptable Risk Safeguards.[20] Furthermore, "the RIF DUA is created for a distinct project, and only the work described in the RIF request materials is allowed under that DUA."[21]

Table 1. Overview of file difference by privacy level

	PUF	LDS	RIF
Requires Privacy Board Review?	No	No	Yes
Requires a Data Use Agreement?	No	Yes	Yes
Files include beneficiary-level data?	No	Yes	Yes
Researchers may request customized cohorts (e.g. Breast cancer patients residing in MA)?	No	No	Yes
Data can be linked at beneficiary level to non-CMS data using a beneficiary identifier?	No	No	Yes
Claim run off period	NA	Annual file: 6-month run off	Annual file: 12-month run off
		Quarterly file: 3-month run off	Quarterly file: 3-month run off

Source: Siedelman, (2016), Table 1. National Provider Identifier (NPI)/Unique Physician Identification Number (UPIN).

CMS data can be combined with other public health data to produce more useful information about access, service delivery, enrollment, outcomes, and more. The National Center for Health Statistics (NCHS) offers public use files that may be downloaded from the Centers for Disease Control and Prevention downloadable public-use data files through the Centers for Disease Control and Prevention's (CDC) FTP server.[22] Additionally, 70 public use files are available for free from the Inter-university Consortium for Political and Social Research, Institute for Social Research, University of Michigan. NCHS public use files have general data use restrictions, but users are not required to file a data use agreement.

NCHS survey data have been linked to the research-oriented Medicaid files that consist of demographic and eligibility information as well as information about individuals' inpatient care, institutional long-term care, pharmacy services, and other services.[23] These linked data files are accessible only through the NCHS Research Data Center (RDC). All interested researchers must submit a research proposal to the RDC to obtain access.[24]

D. Mergers with Private Data: Genealogy, DNA Testing, and Federal Data

In addition to administrative data, such as CMS data discussed above, the confluence of many proprietary datasets with sensitive financial, health, and other information on millions of Americans has elevated the risk of disastrous re-identification. Consider several data breaches that have occurred during 2017 – 2022. There are no reported Medicare or Medicaid data breaches; yet, there was a significant breach of Health Insurance Marketplace data on October 16, 2018.[25] Data snoopers obtained the names, dates of birth, addresses, last four digits of the Social Security number, expected income, citizenship status, employer name, and other highly sensitive information for 93,689 applications from Healthcare.gov.[26]

Information Collected on the Healthcare.gov Marketplace Application

- Name, date of birth, address, sex, and the last four digits of the Social Security number (SSN), if SSN was provided on the application;
- Other information provided on the application, including expected income, tax filing status, family relationships, whether the applicant is a citizen or an immigrant, immigration document types and numbers, employer name, whether the applicant was pregnant, and whether the applicant already had health insurance;
- Information provided by other federal agencies and data sources to confirm the information provided on the application, and whether the Marketplace asked the applicant for documents or explanations;
- The results of the application, including whether the applicant was eligible to enroll in a qualified health plan (QHP), and if eligible, the tax credit amount; and
- If the applicant enrolled, the name of the insurance plan, the premium, and dates of coverage.

Many private companies now hold large swaths of consumers' health information, and that information is no more or less susceptible to breach. For example, several recent high-profile breaches of DNA testing databases resulted in the disclosure of personally identifying health information, financial information, and demographic information.[27]

The damage caused by data leaks and data breaches is only amplified by the combination of public and private datasets. Malevolent merging of health care data with other datasets is a nightmare scenario that would result in the disclosure of the most personal and sensitive details of an individual's life. Particularly where demographic data is involved, piecing together information from various datasets has the potential to "out" someone who is LGBTQI+, disclose immigration status, or reveal details of an individual's functional status. Sophisticated data snoopers, have the capacity to execute this nightmare scenario: for example, Equifax is among the three largest credit

reporting agencies in the U.S. Between May and July 2017, the company experienced a data breach impacting 147 million people.[28] Data snoopers obtained "names, home addresses, phone numbers, dates of birth, Social Security numbers,…driver's license numbers…[and] credit card numbers of approximately 209,000 consumers."[29] The combination of the Equifax dataset with even publically available datasets, like public Medicaid data or data from the American Community Survey, could result in malevolent re-identification and targeting of individuals.

III. Striking the Balance: Promoting Racial Equity and Privacy Practices in CMS Datasets

Thus far, we have discussed only "how" CMS and other agencies use various policies and procedures to protect the privacy of individual persons. It is self-evident that administrative data should have such procedures. Health care data contain considerable confidential information: doctor summaries of a patient's mental and physical state, prescriptions, specific conditions, genetic disposition for certain diseases, and more. Researchers and program administrators need demographic information such as race, gender identity, and so forth to study health care access, outcomes, and experiences. In many cases, such as high-level analysis of care quality, cost, or outcomes, this information can be preserved while protecting the identity of the particular persons associated with each set of conditions, outcomes, and analyses.

A. Demystifying De-identification: The Racial Equity Analysis

Publicly exposing the identity of individuals and their specific health care data to individuals beyond the patient's insurer, provider, or network has several negative consequences: humiliation and invasion of privacy for individuals with sensitive health care issues;[30] predation by insurance companies; legal and political harassment for persons seeking care that is at variance with the religious views of others (such as abortion, birth control, or blood transfusions); discrimination by employers; and harassment of individuals who have or have not taken up a particular therapy, for example, COVID vaccines.

All these scenarios are compounded by systemic racism. Race has historically been a factor in excluding people from access to coverage and care.[31] The effects of that exclusion persist in policies that do not explicitly take race into account when designing equity interventions.

Privacy controls can protect against these negative effects with virtually no impact on research or program administration. Where researchers seek to understand differences and disparities experienced by different racial and ethnic groups, this analysis can generally be completed with fully de-identified data. As discussed in Part II.A, de-identifying data includes removing personally identifying details such as name, birthdate, and other non-relevant details to prevent individual subjects from being identified. De-identification does not restrict program administrators' or researchers' ability to use administrative and demographic data for examination of program quality, even for evaluation that requires access to demographic details, such as the examination of racial equity issues. De-identification does not remove race or other important demographic information, but simply makes it difficult to connect that information with the identity of a specific individual. In the event that analysis of differences by race, preferred language, disabilities, or other demographic categories requires information that may be used to re-identify individuals (and that will be very rare), researchers must use restrictive use data instead of public use files (See Part II.B, above).

In part, racial equity is advanced by improving visibility of demographic information in data; the risks of collecting and using that information in program research and evaluation lie in the mishandling of that data. Failure of administrators to protect privacy by preventing data breaches and data leaks sets back research advancements by fueling data skepticism and reluctance to share information. This in turn may result in limitation of information available for analysis to much less useful, older data.[32] On the contrary, proper de-identification measures allow programs, researchers, advocates, and the public to safely access and analyze data, including demographic differences and inequities, and use that data to advance racial equity.

B. Race, Racism, and Data Privacy

Should CMS collect and report data on the racial identity of individuals? Arguably, including the racial identity of persons in CMS data increases the risk of re-identification. Yet, including racial identity increases the utility of the data for public policy analysis, program administration and service, and health equity analysis; addressing issues of racial equity in health and the socioeconomic determinants of health becomes extremely difficulty without racial identifiers for persons within health datasets.

The relative costs, benefits, and risks of including race are being discussed by various groups within the US and other countries. For example, France has a "color-blind" policy

of data collection and public policies. Scarred by the racism of Vichy France during War World II, France in 1978 outlawed "the collection and computerized storage of race-based data without the express consent of the interviewees or a waiver by a state committee. France therefore collects no census or other data on the race (or ethnicity) of its citizens."[33] Instead, France uses "geographic or class criteria to address issues of social inequalities," while also aggressively targeting hate speech.[34] Compared to the US, France is less aggressive on racial discrimination in "jobs, housing, and in provision of goods and services."[35] The Colombian national census omitted references to race during 1918 – 1993.[36] Until 2020, Mexico omitted racial identity from its national census. Both Mexico and Colombia emphasized "*mestizaje*," the notion that the nation has a unified mixed-race population (chiefly, Spanish and Indigenous) of many colors and that racial distinctions were divisive and unnecessary.[37] The Mexican and Colombian approaches to collecting data on race de-emphasize structural racism, while the French approach suggests that structural inequalities flowing from class and geographic location are suitable proxies for addressing racism without collecting information on race. Yet in all those countries, just as in the United States, there are pervasive and damaging racial inequities.

The U.S. approach towards collecting racially disaggregated data has the potential to confront race more honestly and address disparities. It often (but not always) emphasizes collecting and reporting the racial identity of persons, acknowledging that race is a factor in almost every life outcome. This has the advantage of allowing data users to document the nature and extent of racial inequality in health outcomes and access to alternative dimensions of heath care. It allows users to examine intertemporal trends and spatial variations in health care quality across and within racial groups. Assessing the extent of progress in addressing racial equity requires quantitative assessment, using representative large sample microdata with direct measures of racial identity.

Both identifying race and de-identification strategies for privacy overall are essential to equitable outcomes. The U.S. has a long and well documented history of racial discrimination, racial injustice, and unethical racial experiments. Cognizant of this history, Professor Anita L. Allen synthesizes three sets of issues where data snoopers might use information on the racial identity of individuals to violate individual privacy: first, discriminatory oversurveillance – targeting Black individuals for "fraud, waste, and abuse;" second, discriminatory exclusion – excluding services and institutions used by Black individuals; and, third, discriminatory predation – including Black individuals in wasteful and harmful programs.[38]

Discriminatory oversurveillance in health programs is analogous to racial profiling by police, wherein Black citizens are excessively stopped, searched, and charged relative to otherwise identical white drivers. Program oversight requires programs like Medicare, Medicaid, and CHIP to prevent waste, fraud, and abuse. However, an analysis of Medicare fraud investigations demonstrates disproportionate auditing of providers that serve communities of color, resulting in a poorer or reduced level of service.[39] Discriminatory exclusion means that communities of color are unfairly excluded from Medicaid and CHIP by way of limited access to services and programs to which they are entitled.[40] Finally, discriminatory predation means that data snoopers use racial identification information to target particular groups of Medicaid and Medicare providers and patients for fraudulent activities.[41]

Including racial identity information within CMS data has great benefits to advance racial equity. But there are potential costs without adequate protections in place—excessive surveillance, racial exclusion, and racial predation—and these costs should be included in evaluating systems of data collection, maintenance, and distribution.

C. Nightmare Scenario: Racial Predation using Public and Private Datasets

Current CMS policies for use of administrative data are well designed to protect privacy concerns when the data are used by traditional users, *i.e.*, academic and policy researchers, program administrators, and covered agencies. Moreover, in addition to the specific policies and procedures some traditional data users have strong incentives to avoid re-identification of persons in the data. For example, a university faculty member may be subject to loss of tenure for serious abuses of privacy procedures associated with the use of administrative data.

Notwithstanding these policies, current privacy procedures may be lagging behind the potential threats arising in data mergers of large private datasets with CMS administrative data. Private datasets often include detailed personal and family income that may be merged with the demographic and other information in CMS data, resulting in re-identification and use of individual information for nefarious purposes. This is a "nightmare scenario" for persons concerned with protecting individual privacy.

Consider Ancestry Corporate, a multinational proprietary family research corporation, consisting of a dozen brands. Its primary brand, Ancestry.com, is an online database for genealogical research, allowing millions of customers' access to billions of historical documents. Customers are able to create and store extensive family trees linked to historical documents and other family trees. Trees contain complete names,

contemporary and past residential locations, dates of birth and death, gender, place of birth and death, and whether the person is currently living or deceased for each person listed in the tree. AncestryDNA is a separate but interrelated brand. This is a DNA testing service that maintains a database of DNA results. Users of this service have the option of allowing Ancestry to locate relatives by matching their DNA results against others in their enormous database. This is very helpful for locating previously unknown relatives and verifying relatives whose familial relationship was questioned.

The demographic, biographical, and locational information contained with the Ancestry data, that is, Ancestry.com and AncestryDNA, are sufficient to obtain high quality re-identification matches with Medicaid applicants without the use of Social Security numbers or similar individual identifiers. In particular, for each person in a family tree, the Ancestry data has the person's complete name, gender, whether alive or dead, birth and death dates, and places of birth and death, along with similar information on their spouse, children, siblings, parents, and many other relatives. Additionally, Ancestry trees often contain information on race, citizenship status, military service, country of origin for immigrants, phone numbers, current and past residential addresses, marriage dates, school attended, and pictures. Re-identification could occur if the Ancestry data are matched with Medicaid applicant information: full name, date of birth, gender, location (street, city, county, state, zip code), mailing address, phone numbers, name of spouse, names of tax dependents, citizenship status, veteran status, and immigration records.

Merger of Ancestry and Medicaid data would create a file of millions of persons with substantial biographical, demographic, locational, genetic, and health information. There is a long list of persons, organizations, and agencies that might be interested in such information: insurance companies, police and intelligence agencies, criminals, foreign governments, advertising companies, data snoopers, and more. There is a high probability of such a merger given the relative openness of Ancestry data and access to Medicaid application and other administrative files.

IV. Summary and Discussion

Medicaid administrative data are used for program management, service, analysis, research, and other useful activities. A tradeoff exists, however, between data utility and the risk of re-identification, the former needed for racial equity solutions and the latter to prevent harm. When re-identification occurs, previously anonymous individuals and families are identified, potentially along with their personal information, medical history, financial information, and identifiers that link them to many other survey and

administrative datasets. Successful data snoopers may cause a variety of injuries: denial of insurance coverage, job loss, or lack of job offer; revelation of personal information leading to stigma in a social or work circle, or loss of reputation resulting in isolation or difficulty obtaining employment.[42] Given stigma, joblessness, insurance and access and more are core features of systemic racism, data privacy is a core racial equity concern.

Researchers, policy makers, and program managers may use administrative data to improve the health and wellbeing of citizens. Optimizing the utility of administrative data requires increasing the usefulness of the data for positive purposes without increasing the risk of re-identification. De-identification of administrative data is one policy for reaching this goal. De-identification removes identifiers of individuals, as well as their relatives, employers, or household members. A second policy is to obfuscate and aggregate observational counts within a cell when the count is so small that it substantially increases the risk of re-identification. CMS must take care to balance use of these protective measures so as not to remove or make hidden underserved populations, reducing the utility of the data for racial equity. Instead, CMS should apply a racial equity lens to collecting demographic data, choosing tactics that allow visibility for racial differences while ensuring appropriate restrictions on use and de-identification measures are taken.

CMS and other custodians of administrative and survey data recognize that data users have diverse interest in accessing data. Accordingly, there are public use files and restricted access files. Public use files have the fewest restrictions on use. Further, the data in public use files are strongly de-identified, small cell counts have been concealed, and other measures have been taken to substantially reduce the risk of re-identification. Restricted access files carry a greater risk of re-identification and, therefore, have more limited access than public use files. For example, restricted access files may require: a data use agreement for the individual user and the affiliated organization; utilization of the data only at secure physical or online sites; licensing of users; various measures to secure the data if it is being used outside of a secure site; data inspectors; and, even sub-dividing restricted access data to more or less restricted files, for example, the distinction between Limited Data Sets (very restrictive) and Research Identifiable Files (the most restrictive files).

Finally, it is important to consider that data breaches by sophisticated snoopers can produce a nightmare scenario – merging CMS data with other types of sensitive data such as credit bureau reports, DNA data, and genealogical data. Data breaches of greater or lesser severity have occurred for each of these data sources. The possibility of preventing the nightmare scenario should be a high priority for CMS.

V. Conclusion

Erasing race from data by aggregating or completely removing racial data severely compromises its utility to advance racial equity. Traditionally some federal agencies limit the risk of breach of privacy by suppressing or not releasing data and distorting the representation on communities of color, but such data restrictions cause unintended harm. This can lead to a disconnect between stated goals and the impact on communities.

Curators of data, including CMS, should embrace innovation on safely expanding access to confidential data. For example, the Urban Institute and the Brookings Institute have developed a privacy-preserving methodology with the IRS Statistics of Income Division that can be explored for other agencies.[43] Agencies need support to innovate and capacity to develop racial equity strategies that maintain a deep commitment to data privacy. Both of the Biden-Harris Administration's Executive Orders on Advancing Equity identified data strategies and consultation with impacted communities as high priorities. These two priorities are inextricably linked. In order to develop responsive data strategies that both meet needs and protect privacy, ongoing consultation with communities of color is essential to gauge the actual experiences and impacts that the current gaps in data still obscure. Agencies have a historic opportunity to reshape one of the most important governing tools for justice in our arsenal: racially equitable data strategies that close gaps and keep communities safe.

Appendix A

Examples

Consider Table 2, which provides the hypothetical age distribution for a small town where N = 1,801. There are 6 infants and toddlers (0 – 3 years of age) and 9 pre-school age children (4 – 5 years of age), representing 0.33% and 0.50% of the population, respectively. Similarly, there are 6 persons 95 – 98 years of age (0.33% of population) and 1 person 99 years of age (0.06% of the population).

Aggregate population data (if possible) to limit the need to suppress cell counts. Use secondary suppression as needed to prevent recalculation of cells through subtraction. Use top-coding and bottom-coding to protect the identity of individuals with uniquely high or uniquely low ages. With bottom-coding, all infants, toddlers, and pre-schoolers are assigned

an age of 5 and the cell numbers and percentages are aggregated for persons 0 – 5 years of age. With top-coding, all persons 86 years of age and above are assigned an age of 86 and the cell numbers and percentages are aggregated for persons 86-94, 95-98, and 99 years old.

Table 2. Distribution of ages for small town

Original			Top- and bottom-coding			Collapse cells		
Age	N	Perc.	Age	N	Perc.	Age	N	Perc.
0 – 3	6	0.33%						
4-5	9	0.50%	5	15	0.83%	0-5	15	0.83%
6 – 12	125	6.94%	6 - 12	125	6.94%	6 – 12	125	6.94%
13 – 17	200	11.10%	13 - 17	200	11.10%	13 – 17	200	11.10%
18 – 25	290	16.10%	18 - 25	290	16.10%	18 – 25	290	16.10%
26 – 35	324	17.99%	26 - 35	324	17.99%	26 – 35	324	17.99%
36 – 55	350	19.43%	36 - 55	350	19.43%	36 – 55	350	19.43%
56 – 75	277	15.38%	56 - 75	277	15.38%	56 – 75	277	15.38%
76-85	125	6.94%	76-85	125	6.94%	76-85	125	6.94%
86 – 94	88	4.89%	86	95	5.27%	86 – 99	95	5.27%
95 – 98	6	0.33%			0.00%			0.00%
99	1	0.06%			0.00%			0.00%
Total	1801	100%	Total	1801	100%	Total	1801	100%

Collapsing of cells is another strategy for preventing re-identification because of small numbers. In this case, the two lowest age groups are combined to produce a 0 – 5 age group and the three highest age groups are combined to produce an 86 – 99 age group.

Table 3. Distribution of ages for small town: minimal suppression

Age	N	Perc.
0 – 3	*	*
4-5	*	*
6 – 12	125	6.94%
13 – 17	200	11.10%
18 – 25	290	16.10%
26 – 35	324	17.99%
36 – 55	350	19.43%
56 – 75	277	15.38%
76-85	125	6.94%
86 – 94	88	4.89%
95 – 98	*	*
99	*	*
Total	1801	100%

Table 3 maintains all age ranges. But, for cells with small numbers replace both the numbers and the percentages with "*." Failure to replace the percentages with a "*" will allow data snoopers to re-establish the age groups for each of the cells with small counts.

Table 4. Distribution of ages by race for a small town

				Original				
	White		Black		Other		Total	
Age	Count	Percent	Count	Percent	Count	Percent	Count	Percent
0 – 3	3	0.28%	2	0.50%	1	0.32%	6	0.33%
4 – 5	3	0.28%	3	0.74%	3	0.96%	9	0.50%
6 – 12	25	2.30%	48	11.88%	52	16.67%	125	6.94%
13 – 17	100	9.22%	50	12.38%	50	16.03%	200	11.10%
18 – 25	130	11.98%	90	22.28%	70	22.44%	290	16.10%
26 – 35	210	19.35%	60	14.85%	54	17.31%	324	17.99%
36 – 55	270	24.88%	50	12.38%	30	9.62%	350	19.43%
56 – 75	177	16.31%	70	17.33%	30	9.62%	277	15.38%
76-85	90	8.29%	20	4.95%	15	4.81%	125	6.94%
86 – 94	70	6.45%	11	2.72%	7	2.24%	88	4.89%
95 – 98	6	0.55%	0	0.00%	0	0.00%	6	0.33%
99	1	0.09%	0	0.00%	0	0.00%	1	0.06%
Total	1085	100.00%	404	100%	312	100%	1801	100%
				Collapse of Cells				
0 - 12	31	2.86%	53	13.12%	56	17.95%	140	7.77%
13 - 17	100	9.22%	50	12.38%	50	16.03%	200	11.10%
18 - 25	130	11.98%	90	22.28%	70	22.44%	290	16.10%
26 - 35	210	19.35%	60	14.85%	54	17.31%	324	17.99%
36 - 55	270	24.88%	50	12.38%	30	9.62%	350	19.43%
56 - 75	177	16.31%	70	17.33%	30	9.62%	277	15.38%
76-85	90	8.29%	20	4.95%	15	4.81%	125	6.94%
86 - 99	*77*	*6.45%*	*11*	*2.72%*	*	*	*95*	*4.89%*
76-99	167	15.39%	31	7.67%	22	7.05%	220	12.22%
Total	1085	100.00%	404	100%	312	100%	1801	100%

Table 4 shows why it is sometimes necessary to have secondary suppression. For each racial group, there are small counts for toddlers, infants, and pre-school children. There are 15 children in these age groups, but a minimal age group of 0 – 5 would have an unacceptable risk of re-identification; there would be $1 \leq n \leq 10$ for each racial group, even though the total count is 15. Collapse of the three lowest age groups to create a 0 – 12 age group with 140 total observations and $n \geq 11$ for each racial group.

Similarly, there are 7 persons (all white) in the two highest age groups and 95 persons in the three highest age groups. One strategy is to collapse the two highest age groups to a single 86 – 99 age group and to replace the Other count and percent with "*." This strategy does not sufficiently reduce the risk of re-identification for Other persons. There are multiple ways

to figure out that there are 7 persons in this cell, for example, total – White cell count – Black cell count. Although it reduces information, the best strategy is to create a 76 – 99 maximal age group; each cell will have n ≥ 11 and thereby sufficiently reduce the risk of re-identification.

Appendix B

Table 5. Variable differences between RIF and LDS files

Variable	File	LDS	RIF
Unique Beneficiary Identifier	Claims & Enrollment files	Encrypted identifier	Encrypted identifier
	MedPAR	No identifier	Encrypted identifier
Health Insurance Claim or Social Security Number	Claims & Enrollment files	Not included in file	Included as add-on with special permission only
Dates (MM/DD/YYYY)	Claims files	Included as of CY2010	Included
	MedPAR	Quarter and year only	Included
Claim from date	Claims files	Not included	Included
Claim through date	Claims files	Included	Included
Beneficiary Zip Code	Claims & Enrollment files	County and state	Included
	MedPAR	State only	Included
Beneficiary Date of Birth	Claims, MedPAR & Enrollment files	Not included. Age year or age range	Included
Date of Death	Enrollment files	Included, for validated dates of death only	Included
NPI/UPIN for person level provider	Claims files	As of 2013, the real NPI is included	Included
	MedPAR	Not included	Not included
Facility provider number	Claims files & MedPAR	Included	Included

Variable	File	LDS	RIF
NPI of the facility	Claims files & MedPAR	Included	Included

Source: Siedelman, (2016), Table 2. For 2005-present, "MEDPAR files contain information for 100% of Medicare beneficiaries using hospital inpatient services. The following fields are furnished: total charges, covered charges, Medicare reimbursement, total days, number of discharges and average total days."

ENDNOTES

[1] Exec. Order No. 13,985, 86 Fed. Reg. 7009 (Jan. 20, 2021), *available at* https://www.federalregister.gov/documents/2021/01/25/2021-01753/advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government.
[2] Exec. Order No. 14,901, 88 Fed. Reg. 10825 (Feb. 16, 2023), *available at* https://www.federalregister.gov/documents/2023/02/22/2023-03779/further-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal.
[3] Hannah L. Gould et al., *Data for Equity: Creating an Anti-Racist, Intersectional Approach to Data in a Health Department*, 29 J. PUB. HEALTH MGMT. & PRACTICE 11-20 (2023), *available at* https://journals.lww.com/jphmp/Fulltext/2023/01000/Data_for_Equity__Creating_an_Antiracist,.4.aspx.
[4] *Vermont*, MIGRATION POL'Y INST., *available at* https://www.migrationpolicy.org/data/state-profiles/state/demographics/VT (last visited Mar. 1, 2023).
[5] George T. Duncan, Stephen E. Fienberg, Ramayya Krishnan, Rema Padman, & Stephen F. Roehrig, *Disclosure Limitation Methods and Information Loss for Tabular Data*, *in* CONFIDENTIALITY, DISCLOSURE AND DATA ACCESS: THEORY AND PRAXECCTICAL APPLICATIONS FOR STATISTICAL AGENCIES 135-166 (Doyle, Pat, Julia Ingrid Lane, Jules J. M. Theeuwes, Laura M. Zayatz eds., 2001).
[6] *Application for Health Coverage & Help Paying Costs*, MARKETPLACE.GOV, *available at* https://marketplace.cms.gov/applications-and-forms/marketplace-application-for-family.pdf (last visited Feb. 3, 2023).
[7] Julia Lane & Claudia Schur, *Balancing access to health data and privacy: a review of the issues and approaches for the future*, 45 HEALTH SERV. RES. 1456-67 (2010), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1472736.
[8] *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, U.S. DEP'T OF HEALTH & HUM. SERVS., *available at* https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (last visited Feb. 3, 2023).
[9] *Id.*
[10] Covered entities include health plans, health care providers, health care clearinghouses, and certain business associates of covered entities. *See* U.S. DEP'T OF HEALTH & HUM. SERVS., OCR PRIVACY BRIEF: SUMMARY OF THE HIPAA PRIVACY RULE 4 (Oct. 19, 2022), *available at* https://www.hhs.gov/sites/default/files/privacysummary.pdf.
[11] *Id*. at 10.
[12] *CMS Cell Suppression Policy: Guidance for CMS Cell Suppression Policy Web Page*, U.S. DEP'T OF HEALTH & HUM. SERVS. (Sept. 15, 2020), *available at* https://www.hhs.gov/guidance/document/cms-cell-suppression-policy; *CMS Cell Size Suppression Policy*, RES. DATA ASSISTANCE CTR. (May 8, 2017), *available at* https://resdac.org/articles/cms-cell-size-suppression-policy.
[13] Cathy Wasserman & Eric Ossiander, Washington State Dep't of Health, DEPARTMENT OF HEALTH AGENCY STANDARDS FOR REPORTING DATA WITH SMALL NUMBERS (May 2018), *available at* https://doh.wa.gov/sites/default/files/legacy/Documents/1500//SmallNumbers.pdf.

[14] Lori Siedelman, *Differences between RIF, LDS, and PUF Data Files*, RES. DATA ASSISTANCE CTR. (Aug. 10, 2016), *available at* https://resdac.org/articles/differences-between-rif-lds-and-puf-data-files.
[15] *Id*.
[16] *CCW Virtual Research Data Center (VRDC)*, RES. DATA ASSISTANCE CTR. (2020), *available at* https://resdac.org/cms-virtual-research-data-center-vrdc. For comparison, consider access to restricted use data such as the Medical Expenditures Panel Survey, an Agency for Healthcare Research and Quality dataset. Researchers must examine this data at a Federal Statistical Research Data Center. *See Restricted Data Files Available at the Data Centers*, AGENCY FOR HEALTHCARE RES. & QUALITY *available at* https://meps.ahrq.gov/mepsweb/data_stats/onsite_datacenter.jsp?# (last visited Feb. 3, 2023); and, *Available Data*, CENSUS.GOV *available at* https://www.census.gov/about/adrm/fsrdc/about/available_data.html (last visited Feb. 3, 2023). "All researchers wanting to use a FSRDC will need to become Special Sworn Status (SSS) employees of the Census Bureau—in case of incidental access to confidential Census Bureau or Internal Revenue Service data while in an FSRDCs—and will also be required to become National Center for Health Statistics agents (as AHRQ data is based on the National Health Interview Survey) and to take the appropriate training for both roles." *AHRQ-Census Bureau Agreement on Access to Restricted MEPS Data at Federal Statistical Research Data Centers*, AGENCY FOR HEALTHCARE RES. & QUALITY, *available at* https://meps.ahrq.gov/communication/census_announce.shtml (last visited Feb. 2, 2023). The National Center for Education Statistics offers remote access, but users must obtain a restricted-use data license. *See Restricted-Use Data Procedures Manual*, NAT'L CTR. FOR EDUC. STAT., *available at* https://nces.ed.gov/statprog/rudman/ (last visited Feb. 3, 2023).
[17] "An innovator is a researcher associated with a for-profit organization, or anyone conducting research with the intent to also create a product or tool to be sold. For example, an innovator could use CMS data to develop care management or predictive modeling tools." *Innovator Research FAQs*, RES. DATA ASSISTANCE CTR., *available at* https://resdac.org/innovator-research-faqs (last visited Feb. 2, 2023).
[18] DEP'T OF HEALTH & HUM. SERVS., CTRS. FOR MEDICARE & MEDICAID SERVS., INSTRUCTIONS FOR COMPLETING THE LIMITED DATA SET DATA USE AGREEMENT (DUA) (CMS-R-0235L) (Jul. 7, 2022), *available at* https://www.cms.gov/Medicare/CMS-Forms/CMS-Forms/Downloads/CMS-R-0235L.pdf
[19] *See Identifiable Data Files*, CTRS. FOR MEDICARE & MEDICAID SERVS., *available at* https://www.cms.gov/research-statistics-data-and-systems/files-for-order/identifiabledatafiles (last visited Feb. 3, 2023).
[20] *Data Management Plan Self-Attestation Questionnaire (DMP SAQ)*, RES. DATA ASSISTANCE CTR, *available at* https://resdac.org/request-form/dmp-saq (last visited Feb. 3, 2023); CTRS. FOR MEDICARE & MEDICAID SERVS., OFC. OF INFORMATION TECH., CMS ACCEPTABLE RISK SAFEGUARDS (ARS): FINAL (Nov. 21, 2017), *available at* https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/InformationSecurity/Info-Security-Library-Items/ARS-31-Publication.
[21] Email from Kyra Wicklund, Res. Data Assistance Ctr. (on file with author).

[22] *See Public-Use Data Files and Documentation*, CTRS. FOR DISEASE CONTROL & PREVENTION *available at* https://www.cdc.gov/nchs/data_access/ftp_data.htm (last visited Feb. 3, 2023).

[23] Detailed descriptions of each of these files can be found in the following online report: CTRS. FOR DISEASE CONTROL & PREVENTION, NAT'L CTR. FOR HEALTH STAT., THE LINKAGE OF NATIONAL CENTER FOR HEALTH STATISTICS SURVEY DATA TO CENTERS FOR MEDICARE & MEDICAID SERVICES TRANSFORMED MEDICAID STATISTICAL INFORMATION SYSTEM CLAIMS DATA (2014-2019): MATCHING METHODOLOGY AND ANALYTIC CONSIDERATIONS (Oct. 14, 2022), *available at* https://www.cdc.gov/nchs/data/datalinkage/nchs-cms-tmsis-linkage-methodology.pdf.

[24] *See Research Data Center (RDC)*, CTRS. FOR DISEASE CONTROL & PREVENTION, *available at* https://www.cdc.gov/rdc/index.htm (last visited Feb. 3, 2023).

[25] For detailed studies on the history of healthcare data breaches, *see Healthcare Data Breach Statistics*, HIPAA JOURNAL, *available at* https://www.hipaajournal.com/healthcare-data-breach-statistics/ (last visited Feb. 3, 2023) *and* Adil Hussain Seh, Mohammad Zarour, Mamdouh Alenezi, Amal Krishna Sarkar, Alka Agrawal, Rajeev Kumar, & Raees Ahmad Khan, *Healthcare Data Breaches: Insights and Implications*, 8 HEALTHCARE 133 (May 13, 2020).

[26] Evan Sweeney, *CMS increases Healthcare.gov breach total to 93,600*, FIERCE HEALTHCARE (Nov. 13, 2018), *available at* https://www.fiercehealthcare.com/tech/cms-increases-healthare-gov-breach-totals-to-93-000.

[27] Prominent breaches of DNA databases include: DNA Diagnostics Center, November 29, 2021 and Veritas Genetics, November 7, 2019. Veritas Genetics is a for-profit genetic testing service. The company announced that its "customer-facing portal had been accessed by an unauthorized user," but none of personal health records, DNA test results or genomic data were accessed by the data snooper. Jeff Orr, *Incident Of The Week: DNA-Testing Company Veritas Genetics Discloses Unauthorized Access Of Customer Data*, CYBER SECURITY HUB (Nov. 8, 2019), *available at* https://www.cshub.com/attacks/articles/incident-of-the-week-dna-testing-company-veritas-genetics-discloses-unauthorized-access-of-customer-data. The company reports that "only a handful of customers" were affected by the breach.
On August 6, 2021, a data breach occurred that affected 2,102,436 customers of DNA Diagnostics Center, Inc. The data snooper had obtained personal, identifying, and financial information obtained during 2004 - 2014. Richard Console, *Data Breach Alert: DNA Diagnostics Center, Inc. Security Incident Puts Personal Data at Risk*," JD SUPRA (Mar. 28, 2022), *available at* https://www.jdsupra.com/legalnews/data-breach-alert-dna-diagnostics-1109913/. There has been no independent verification of the extent of the data breaches of Veritas Genetics and DNA Diagnostics Center, the number of customers affected, or steps taken by the company to prevent a re-occurrence of a data breach.
A larger and more serious data breach may occur if companies do not do a better job protecting data. Consider GEDmatch, a genetic genealogy service. More than 1 million customers have uploaded their genetic data to GEDmatch, hoping to find matches in their database. Peter Ney, Luis Ceze, Tadayoshi Kohno, Univ. of Washington, GENETIC EXTRACTION AND FALSE RELATIVE ATTACKS: SECURITY RISKS TO THIRD-PARTY GENETIC GENEALOGY SERVICES BEYOND IDENTITY INFERENCE (2019), *available at* https://dnasec.cs.washington.edu/genetic-genealogy/ney_ndss.pdf; Emily Mullin, *A DNA Database Containing Data From 23andMe and Ancestry Is Vulnerable to Attacks*, ONEZERO (Oct. 30, 2019), https://onezero.medium.com/the-dna-database-containing-data-from-23-me-and-ancestry-is-vulnerable-to-attack-6fe5df2497b3.

Professors Ney, Ceze, and Kohno have shown that it's possible to extract genetic details on

any person in GEDmatch's genetic database. Fortunately, these authors have also offered solutions for increasing security against data snoopers.

[28] *Equifax Data Breach*, EPIC, *available at* https://archive.epic.org/privacy/data-breach/equifax/ (last visited Feb. 3, 2023).

[29] *Id*.

[30] For example, STDs, Alzheimer's diagnosis, mental health diagnoses and documentation, impotence and infertility, genetic information that shows an individual does not match a presumed parent could all be disclosed with access to these datasets.

[31] Christina Linke Young, *There are clear, race-based inequalities in health insurance and health outcomes*, BROOKINGS (Feb. 19, 2020), *available at* https://www.brookings.edu/blog/usc-brookings-schaeffer-on-health-policy/2020/02/19/there-are-clear-race-based-inequalities-in-health-insurance-and-health-outcomes/.

[32] By way of analogy, the federal census does release data with individual names and address, but only after 72 years. So, 1950 census with personal identities is available to researchers but the 1960 census with personal identities is not available. The health care information contained in administrative data is needed to solve contemporary problems, that means protecting privacy.

[33] Erik Bleich, *Race policy in France*, BROOKINGS (May 1, 2001), *available at* https://www.brookings.edu/articles/race-policy-in-france/.

[34] *Id*.

[35] *Id.*

[36] *Race and Ethnicity*, *in* COLOMBIA: A COUNTRY STUDY (Dennis M. Hanratty & Sandra W. Meditz, , eds. 1988), available at https://countrystudies.us/colombia/37.htm. There were racial identity questions on the 1993, 2005, 2018 Colombian national census. The International Service for Human Rights (2022) has questioned the accuracy and reliability of this data. *Id*.

[37] Kiko Martinez, *Mexico's 2020 Census Is the First Time Afro-Mexicans Have Been Acknowledged & Counted as Such*, REMEZCLA (Feb. 11, 2021), *available at* https://remezcla.com/culture/mexico-2020-census-results-historical-addition-of-afro-mexicans/.

[38] Anita L. Allen, *Dismantling the "black opticon": privacy, race equity, and on-line data protection reform*, 131 YALE L. J. FORUM (Feb. 20, 2022), *available at* https://www.yalelawjournal.org/forum/dismantling-the-black-opticon.

[39] Lauren Hersch Nicholas et al., *Medicare Beneficiaries' Exposure to Fraud And Abuse Perpetrators*, HEALTH AFF. (May 2019), *available at* https://www.healthaffairs.org/doi/10.1377/hlthaff.2018.05149.

[40] Jane Perkins & Sarah Somers, *The Ongoing Racial Paradox of the Medicaid Program*, 16 J. HEALTH & LIFE SCI. L. (2022), *available at* https://www.americanhealthlaw.org/content-library/journal-health-law/article/1ace7226-252b-43c8-a52d-960a4dd3df8f/The-Ongoing-Racial-Paradox-of-the-Medicaid-Program

[41] Lauren Hersch Nicholas et al., *Medicare Beneficiaries' Exposure to Fraud And Abuse Perpetrators*, HEALTH AFF. (May 2019), *available at* https://www.healthaffairs.org/doi/10.1377/hlthaff.2018.05149.

[42] Julia Lane & Claudia Schur, *Balancing access to health data and privacy: a review of the issues and approaches for the future*, 45 HEALTH SERV. RES. 1456-67 (2010) *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1472736.

[43] Leonard E. Burman, *TPC Builds a Moog – Or How Synthetic Data Could Transform Policy Research*, TAX POL'Y CTR. (Jul. 13, 2020), *available at*

https://www.taxpolicycenter.org/taxvox/tpc-builds-moog-or-how-synthetic-data-could-transform-policy-research.