EQUITY-CENTERED DATA GOVERNANCE, ANALYSIS, AND REPORTING OF DISAGGREGATED RACE/ETHNICITY

A guide for data managers, systems administrators, and analysts

NYU CENTER FOR THE STUDY OF ASIAN AMERICAN HEALTH

CACF
Coalition For Asian American Children+Families

NYU Grossman School of Medicine at NYU Langone Health

BACKGROUND

Health disparities on the basis of race and ethnicity are well-known. However, there are still many gaps in our collective knowledge due to the limitations of the current federal reporting standards. These standards were determined by the Office of Management and Budget (OMB) in 1997 and consist of five minimum race categories (White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander), and one ethnic group (Hispanic/Latino).

Studies find that race/ethnicity data can be more accurate and complete when patients self-report racial or ethnic identity. It is important to note that the OMB recommends expanding, or disaggregating, the minimum categories into more specific groups based on the local population, but this practice is rarely implemented. All the while, an increasing number of individuals select "Other" as their race, or decline to self-identify at all. Further, the standard categories homogenize groups that diverse and populous, often masking critical within-group health disparities (see FAQ below).

To overcome these issues, health researchers, patient advocates, and physicians have called for disaggregation practices to be widely implemented. However, there is pushback against changing categories without an official mandate, often motivated by uncertainty around the complexity of disaggregated race/ethnicity data governance, analysis, and interpretation. This toolkit aims to address some of those issues and provide relevant resources to data managers, systems administrators, statisticians, and other individuals involved in the configuration and maintenance of data governance systems.

USING THIS TOOLKIT

We offer high-level responses to frequently asked questions related to the data analysis and governance of disaggregated race/ethnicity. Existing guides offer detailed explanation of recommended methods, potential challenges, and other considerations related to data disaggregation. These resources are provided at the end of this document and are organized by issue.

There is no one-size-fits all approach to disaggregating race/ethnicity information. Specific terms and collection methods must take into consideration community perspectives and local contexts (detailed further in other parts of this toolkit); these steps will ultimately guide data analysis and use. Therefore, we do not prescribe specific protocols, but present commonly recommended methodologies and practices that can be adapted to local population needs, organizational capacity, and other factors.

WHO WE ARE

The Innovations in Data Equity for All Laboratory (IDEAL) initiative is led by the NYU Center for the Study of Asian American Health and Coalition for Asian American Children and Families with support by colleagues from the New York Academy of Medicine, NYU Langone Health, and the New York State Department of Health. Our goal is to reduce racial/ethnic health disparities using data collection and analysis methods that are inclusive, equitable, and patient-centered.

FAQS

OPEN-ENDED RESPONSES TO RACE/ETHNICITY INQUIRIES ARE OFTEN CITED AS A BEST PRACTICE, BUT THIS KIND OF DATA IS DIFFICULT TO MANAGE AND ANALYZE. HOW DO WE HANDLE THESE TEXT-BASED RESPONSES?

Free-text responses are the best way to collect individuals' accurate race/ethnicity information in a given population. In 2018, the Robert Wood Johnson Foundation commissioned a comprehensive report that included key data disaggregation recommendations, one of which is for the OMB to develop protocols for enumerating open-ended responses (PolicyLink). However, there are already a number of practices in places for such analysis. Depending on institutional resources, size, and other factors, some methods may be preferred over others.

For example, the Institute of Medicine's Subcommittee on Standardized Collection of Race/Ethnicity Data for Healthcare Quality Improvement notes that free text "may improve self-identification but can impose additional administrative burdens if labor-intensive manual coding must be undertaken in the absence of automated systems or optical scanning technology." The subcommittee recommends expanding race/ethnicity categories on questionnaires based on local data, such as the American Community Survey, to offer survey respondents with more options to identify with a granular ethnicity and only providing the option to write in a response if their preferred identity is not listed (e.g. "Race/ethnicity not listed, please specify_____"). (IOM)

As part of a series on disaggregated race/ethnicity, experts from UCLA recommended manually coding free text race/ethnicity responses for small sample sizes. For larger datasets, automated coding procedures (see below: "Coding and Machine-Learning Strategies for Disaggregated Racial/Ethnic Data") include matching responses to those in a preexisting dictionary and algorithms based on machine learning, such as natural language processing. The presenters recommend a range of programming languages and software depending on data complexity and technical expertise.

HOW DO YOU REDUCE THE RISK OF RE-IDENTIFICATION FOR RACIAL AND ETHNIC GROUPS WITH A SMALL POPULATION SIZE?

Although granular race and ethnicity information is useful in many ways, it can risk privacy and confidentiality if not handled with care. The following strategies may include some or all of the following:

1. Restrict access to data to a small set of "trusted researchers"
2. Suppress and aggregate data from small populations
3. Releasing only limited data

It is important to keep in mind that methods to ensure privacy may also inadvertently distort or misrepresent data in published reports, limiting its usefulness. The resources below provide more detailed explanations of the above recommendations and how to implement them without compromising the accuracy and usefulness of reported data.

Because "data privacy restrictions can significantly reduce data usefulness and accuracy," the Urban Institute is developing a methodology with the IRS that generates synthetic data and a validation server to expand access to confidential administrative tax data while also protecting privacy (see 'Recommended resources' for link to report). It is important to continue validating and reproducing such technologies and methods so they may be implemented across fields.

SHOULDN'T PROVIDERS WAIT FOR THE FEDERAL GOVERNMENT TO OFFICIALLY AMEND THE STANDARD CATEGORIES BEFORE CHANGING PATIENT DEMOGRAPHIC QUESTIONNAIRES?

Federal changes to the race/ethnicity standard are likely coming soon. In May 2022, the US Census Bureau announced in that it will consider deviating from OMB standards for the Census questionnaire. The proposed changes include adding a Middle Eastern and North African race category and combining the 5 current race categories and the Hispanic/Latino ethnicity category into a single question.3 One month later, the White House announced that the OMB would also examine these recommendations. A federal interagency working group has subsequently begun evaluating research and public comments to inform its recommended revisions to the federal minimum reporting requirements.

However, the lack of a formal change has not precluded state and local health departments from disaggregating data among their constituents. In fact, one of the Robert Wood Johnson Foundation's current key initiatives is "Advancing state and local policy change to promote data disaggregation."

To illustrate, the State of Hawaii, the County of Santa Clara, and the State of Michigan are just a few jurisdictions that have collected COVID-19-related data from disaggregated racial/ethnic groups that are particularly populous in the region. This helped the health departments address potential disparities that wouldn't have otherwise been detected in data findings. In all cases, the disaggregated race/ethnicities could be "rolled up" or combined into the OMB standards, allowing for data harmonization.

New York State is following suit; in December 2021, the state governor signed New York State Law S.6639-A/A.6896-A. Following a two-year implementation period, the bill will require all state agencies' race/ethnicity questionnaires to include disaggregated response options for Asian Americans (AA) and Native Hawaiian or Other Pacific Islanders (NH/PI), allowing constituents to identify with a more specific group. Thus, it is in the best interest of both providers and patients alike to begin planning for a more flexible, dynamic demographic data collection scheme even in the absence of a federal mandate.

WHAT ARE RECOMMENDED PRACTICES FOR IMPUTING MISSING RACE/ETHNICITY DATA?

Self-reporting is the "gold standard" of collecting race/ethnicity information. However, self-reporting is not always available, and can have "high, non-random missingness" (Elliot, 2022). Imputing missing race/ethnicity information on an existing data set can be a helpful first step for institutions to begin investigating population demographics alongside disaggregation efforts.

For datasets with high rates of race/ethnicity nonresponse, we recommend using a RAND Corporation-developed technique called Bayesian Improved Surname Geocoding (BISG) to make group-level inferences about the racial and ethnic composition of the population. We have provided resources below that further detail the various applications and restrictions of BISG, which relies on surname and geographic information. BISG additionally only imputes race/ethnicity for six aggregate categories (American Indian/Alaska Native, Asian Pacific Islander, Black, Hispanic, White, and Multiracial). Although BISG has not yet been developed for disaggregated racial/ethnic categories, similar methodologies may be implemented for estimating more granular racial/ethnic groups. Because this method has been constructed for population-level analysis, it is not appropriate for imputing racial/ethnic classification of an individual (Elliot, 2022).

Using previously-validated surname lists are an additional way to impute granular race/ethnicity. This method has been most commonly applied to only certain groups for whom surname can be a predictor of ethnic identification. For example, surnames lists derived from Arabic language speakers and/or specific countries of origin have been used to estimate the proportion of Arab Americans or Middle Eastern/North African individuals broadly in secondary data sets. However, the sensitivity of surname lists can vary, they may not be applicable for all groups, and they may create bias.

Surname analysis may be a starting point for estimating the proportion of the population of interest that may identify with a certain racial/ethnic group where that data is lacking. Surname list applications require appropriate cultural understanding of the population to which the surname list is being applied and, like the BISG method, they are not suitable replacements for self-reported data. This method is not recommended for to make inferences about an individual, but to estimate the proportion of your data set that may identify with a certain race/ethnicity. It is helpful to cross-reference imputation results with other data sources, such as the American Community Survey or census data at the national level, that include race/ethnicity variables.

The Racial Equity Analytics Lab at the Urban Institute also recommends the following "ethics checkpoints" to make before imputing missing race/ethnicity. These checkpoints are: 1) before imputation, audit input data for bias, 2) during imputation, examine where bias could be introduced at each step, and 3) after imputation, assess whether imputed race/ethnicity data are accurate enough to be used ethically for analytic purposes. We recommend reviewing the full report, included in the resource list below.

I WORK FOR AN INSTITUTION THAT USES A DATA COLLECTION AND MANAGEMENT SYSTEM THAT DOESN'T ALLOW FOR EXPANDED CATEGORIES. IS THERE A WAY TO OVERCOME THIS CHALLENGE?

Data management and analytical challenges rooted in an existing system infrastructure, particularly those that are well established in a given organization, may not be easily remedied. This is a commonly-cited issue for healthcare organizations using one or more electronic health records (EHR) systems. For example, some data collection forms require re-programming to allow for additional fields to capture more granular race and ethnicity.

Some systems also prohibit selection of multiple categories, causing overestimations or underestimations some racial and ethnic groups. A report from the Colorado Trust states: "Despite the added administrative burden, many organizations that recognize the importance of understanding their patient demographics and corresponding health outcomes manually code and analyze the data in the absence of adequate automated systems. As health care organizations become more aware of how EMRs and other health IT systems could complement their efforts to understand racial and ethnic health disparities, they can encourage health IT and EMR companies to develop tools that can better capture and analyze patient demographic information."

The Trust further recommends: "the inclusion of additional fields for race and ethnicity that align with new federal standards; use of the HHS Race and Ethnicity Data Standards; listing subpopulations that are relevant... the inclusion of mechanisms to ensure uniformity of race and ethnicity data collection; and more robust data for large commercial payers, Medicaid, Medicare, self-insured and small group plans."

Ultimately, system overhauls may be necessary to fulfill some of the best practices this toolkit describes, which requires significant time and resources. Individuals who work closely with their organization's data should consider the system functions required to ensure data is as accurate, complete, and equity-driven as possible and encourage leaders to invest in new and updated systems.

WHAT ARE BEST PRACTICES FOR REPORTING FINDINGS STRATIFIED BY DISAGGREGATED RACE/ETHNICITY?

Data analysts and researchers using race/ethnicity data may inadvertently cause harm by reporting statistical comparisons that "apply positive values to cultural norms associated with whiteness and negatively measure people of color by those norms."

An example can be a report comparing adverse cardiovascular health outcomes among Black and Hispanic/Latino individuals to the relatively healthier White population (Child Trends). While it is important to understand these health disparities, we advise against this kind of "deficit-framing," which can reinforce existing stigmas and stereotypes, a particular risk when reporting on granular ethnocultural groups.

Equity experts recommend "asset-framing" to foreground communities' strengths and aspirations in data reports in addition to challenges and/or needs. When describing community challenges, it is also recommended to contextualize data outcomes in terms of their root causes (such as systemic racism). While interpreting and reporting data, we encourage closely reviewing the ethics guides in the resource list below, as well as reviewing the document in this toolkit aimed at community leaders and community-based organizations. These stakeholders should help supplement statistics reporting with contextual information and terminology to help avoid any mischaracterizations of people based on the numbers alone.

IS IT NECESSARY TO COLLECT INFORMATION ABOUT MULTIRACIAL POPULATIONS? HOW SHOULD THIS INFORMATION BE ANALYZED ALONGSIDE OTHER RACIAL/ETHNIC GROUPS?

Individuals who are considered multiracial are among the most rapidly growing populations in the United States, along with Asian Americans. Currently, nearly 90% of multiracial adults are biracial, with approximately 10% reporting three or more races, and less than 1% reporting four or more races (Pew Research Center). Surveys find that individuals with multiracial identity can have a shifting perception of race over time. Self-identifying as multiracial or with only one racial identity can depend on individuals' stage in life or personal experiences, such as with discrimination.

This is why it is important to encourage selection of multiple races on data collection forms, which may improve understanding of issues affecting multiracial people and any within-group differences for that population. Doing so requires careful deliberations of survey format and wording, tabulation methods, and public education.

At a U.S.-based regional-level (city, county, state, etc.), multiracial data collection instruments can vary widely based on population demographics and institutional resources. A research review of international heterogeneity/granularity in ethnicity classifications notes two distinct national methods from the United Kingdom: "Scotland's mixed category included a write-in option, offering more freedom for responses, but potentially creating difficulties for analysis and interpretation; compared to England and Wales which included disaggregated 'tick boxes' within the Mixed/multiple ethnic group category" (Villarroel et al). Thus, the review authors recommend allowing free text responses as well as multiple response in the interest of "accommodating the increasing population of people identifying themselves as mixed-origin."

RECOMMENDED RESOURCES FOR DATA GOVERNANCE AND STATISTICS

COMPREHENSIVE BEST PRACTICE GUIDES

- Counting a Diverse Nation: Disaggregating Data on Race and Ethnicity to Advance a Culture of Health (PolicyLink)
- Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals (Journal of the American Medical Association)
- Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. (Institute of Medicine)
- How to Embed a Racial and Ethnic Equity Perspective in Research (Child Trends)
- Health Equity and Race and Ethnicity Data (Colorado Trust)

PRIVACY AND SMALL SAMPLE SIZES

- Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. Statistics Surveys, 5, 1-29. https://doi.org/10.1214/11-SS074
- Federal Committee on Statistical Methodology (FCSM), 2005. Statistical policy working group 22 - report on statistical disclosure limitation methodology. U.S. Census Bureau. Available from: https://www.hhs.gov/sites/default/files/spwp22.pdf
- Singpurwalla, DJ. (2021 April 30). Strategies in Mitigating Disclosure Risk in Disaggregated Racial/Ethnic Data [Presentation]. Center for Health Policy Research, UCLA. Available from: https://youtube/gXvZClm162M
- Kags A, Madhub D, Houghton I, Chinganya O, Falconer TA, Lucas S. (23 October 2020). "Balancing data use and data protection – Learning from African Experiences." [Presentation]. 202 United Nations Data Forum. Available from: https://youtube/Uw1gMHY7oxw
- Hadley, Emily. "5 steps to take as an antiracist scientist." (2020). RTI International. Available from: https://www.rti.org/insights/5-steps-take-antiracist-data-scientist
- Becker, T. (2021 Feb 1). Advanced Weighting Strategies for Disaggregated Racial/Ethnic Data [Presentation]. Center for Health Policy Research, UCLA. Available from: https://youtube/xUPKorNMpp0

DATA ETHICS

- Lucas, Sarah. "Data Equity: Six steps beyond data disaggregation." (1 April 2022) William and Flora Hewlett Foundation. Available from: https://hewlett.org/data-equity-six-steps-beyond-data-disaggregation/
- New York City Department of Health and Mental Hygiene. (30 October 2020). Question Wording & Response Sets for Disaggregated Racial/Ethnic Data. [Presentation]. Center for Health Policy Research, UCLA. Available from: https://www.youtube.com/watch?v=gYV4lte3TSU

GROUP-SPECIFIC DATA GUIDES (COMMISSIONED BY ROBERT WOOD JOHNSON FOUNDATION):

AMERICAN INDIAN AND ALASKA NATIVE:

- Report: Villegas, M., Ebarb, A., Pytalski, S., & Roubideaux, Y. (2016). Disaggregating American Indian & Alaska native data: A review of literature. Available from: https://www.ncai.org/DataDisaggregationAIAN-report_5_2018.pdf
- Presentation: http://www.policylink.org/sites/default/files/AmberEbarb_BeyondtheAsterisk.pdf

BLACK/AFRICAN AMERICAN:

- Report: Jackson, J. S., Hamilton, T. G., Ifatunji, M. A., Lacey, K. K., Lee, H. E., & Rafferty, J. A. (2017). Using analytic domains within the Black population to understand disparities in population health. Available from: https://www.policylink.org/sites/default/files/Black-report.pdf
- Presentation: http://www.policylink.org/sites/default/files/mosi-ifatunji-black-population-project-overview-may-24-2017.pdf

HISPANIC OR LATINO/A:

- Report: Alcántara, C., Cabassa, L. J., Suglia, S., Ibarra, I. P., Falzon, A. L., McCullough, E., & Alvi, T. (2017). Disaggregating Latina/o surveillance health data across the lifecourse: barriers, facilitators, and exemplars. Available from: https://www.policylink.org/sites/default/files/Latino-report.pdf
- Presentation: http://www.policylink.org/sites/default/files/carmela-alcantara-alc-cab_rwj-convening_05_24_17_v2.pdf

ASIAN AMERICAN, NATIVE HAWAIIAN, AND PACIFIC ISLANDER:

- Report: Ponce, N., Scheitler, A. J., & Shimkhada, R. (2018). Understanding the culture of health for Asian American, Native Hawaiian and Pacific Islanders (AANHPIs): what do population-based health surveys across the nation tell us about the state of data disaggregation for AANHPIs? Available from: https://www.policylink.org/sites/default/files/AANHPI-report-final.pdf
- Presentation: http://www.policylink.org/sites/default/files/ninez-ponce-ucla-aanhpi-rwj.pdf

NON-HISPANIC WHITE:

- Report: Read, J. N. G. (2017). Challenges and Prospects for Disaggregating Health Data among Non-Hispanic Whites. Duke University. Available from: https://www.policylink.org/sites/default/files/White-report.pdf
- Presentation: http://www.policylink.org/sites/default/files/jennan_rwjf_05-24-17.pdf

WORKING WITH MULTIRACIAL POPULATION DATA:

- Lucas, J and Ruiz, N. (2021, Feb 6). Collection and Reporting of Data on the Multiracial Population [Presentation]. Center for Health Policy Research, UCLA. Available from: https://youtu.be/wxLxrDa0abo
- Villarroel, et al. (2016). Heterogeneity/Granularity in Ethnicity Classifications outside the United States (HGEC project). Available from: https://www.policylink.org/sites/default/files/International-report.pdf
- Lopez, et al. "Half Measures: California's journey toward counting multiracial people by 2022." (2020). Multiracial Americans of Southern California. Available from: http://www.mixedracestudies.org/wp/wp-content/uploads/2022/01/Half-Measures-FINAL.pdf

MISSING DATA

- Elliot, MN. (2021, Oct 14). Imputation Strategies for Racial/Ethnic Data Disaggregation [Presentation]. Center for Health Policy Research, UCLA. Available from: https://youtu.be/Eiw92vvr43s
- Stern A, et al. (2021). Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity: A Case Study Imputing Credit Bureau Data . Urban Institute. Available from: https://www.urban.org/sites/default/files/publication/104582/ethics-and-empathy-in-using-imputation-to-disaggregate-data-for-racial-equity-a-case-study-imputing-credit-bureau-data.pdf
- Brown, K.S., et al. (2021). Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity: Recommendations and Standards Guide. Urban Institute. Available from: https://www.urban.org/sites/default/files/publication/104512/ethics-and-empathy-in-using-imputation-to-disaggregate-data-for-racial-equity_1.pdf
- Adjaye-Gbewonyo, D., Bednarczyk, R. A., Davis, R. L., & Omer, S. B. (2014). Using the Bayesian Improved Surname Geocoding Method (BISG) to create a working classification of race and ethnicity in a diverse managed care population: a validation study. Health services research, 49(1), 268-283. https://doi.org/10.1111/1475-6773.12089
- Randall M, Stern A, Su Y. "Five Ethical Risks to Consider Before Filling Missing Race and Ethnicity Data." (March 2021). Urban Institute. Available from: https://www.urban.org/sites/default/files/publication/103830/five-ethical-risks-to-consider-before-filling-missing-race-and-ethnicity-data-workshop-findings_0.pdf

ANALYZING MULTIPLE-SELECTION AND FREE TEXT RESPONSES

- Comulada, WS. (2020 Dec 15). Coding and Machine-Learning Strategies for Disaggregated Racial/Ethnic Data [Presentation]. Center for Health Policy Research, UCLA. Available from: https://youtube/uUmmYpWINv4
- Liebler CA, Halpern-Manners A. A Practical Approach to Using Multiple-Race Response Data: A Bridging Method for Public-Use Microdata Demography. 2008 Feb; 45(1): 143–155. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2831381/?report=classic
- Gaines, T. Statistical Methods for Analyzing Multiple Race Response Data. Federal Committee on Statistical Methodology. 2007. Available from: https://www.fcsm.gov/assets/files/docs/2007FCSM_Gaines-VII-C.pdf

- Northwestern University. Analysis Considerations for Race, Ethnicity, and Gender Variables. 2022 January 21. Department of Preventive Medicine. Available from: https://www.feinberg.northwestern.edu/sites/firstdailylife/docs/Analysis%2Considerations%20for%20Race,%20Ethnicity,%20%20Gender%20Variables.pdf

MODEL SURVEYS AND METHODS

- Race and Ethnicity using the California Health Interview Survey (CHIS)
- National Health Interview Survey Data, Questionnaires and Related Documentation
- Overview of standards for data disaggregation. - United Nations Statistics Division
- Oregan Health Authority – Race, Ethnicity, Language, Disability (REALD) standards

This toolkit was developed by:

NYU CENTER FOR THE STUDY OF ASIAN AMERICAN HEALTH

CACF Coalition For Asian American Children+Families

NYU Grossman School of Medicine at NYU Langone Health

Artwork by Isabel Lu

Supporting partners:

THE NEW YORK ACADEMY OF MEDICINE

NEW YORK STATE Department of Health