Population Descriptors in Big Data Research

The importance of curation and choice

Curating and analyzing data are not neutral tasks. They involve making choices about how groups are *lumped* together, *split* apart, or excluded entirely. These choices affect how people like researchers, clinicians, policy makers, and patients make sense of and use the research.

Standardization Efforts

Because data curation and analysis choices can have such large effects, many scholars have pushed for more standardization for these tasks. Examples of resources to help with this include:

- The National Academies of Sciences, Engineering, and Medicine (NASEM)'s 2023 report, "Using Population Descriptors in Genetics and Genomics Research" has useful advice even for researchers outside of genetics and genomics.[1] It includes a short set of recommendations. Examples include:
    - Researchers should avoid using the term "Caucasian" because it is rooted in white supremacy.[2,3]
    - Researchers should document and share how they chose population descriptors they use. They should reflect on, document, and share their reasons for lumping (combining) population descriptors together.

- The PhenX Toolkit is a collection of approved methods for measuring peoples' traits (phenotypes) and the things they are exposed to (exposures).[4] For example, it includes a protocol for measuring social vulnerability using variables in existing data.[5] This helps keep methods consistent across studies, making it easier to compare and replicate results. It may be useful as you explore and use population descriptors in your work. It was created by scholars from many fields, including CHIRON academic workgroup member Maile Tauali'i.

Community Advocacy

Advice on population descriptors often comes from communities speaking up for themselves when the usual grouping schemas don't meet their needs. For example,

- Many Pacific Islanders reject the common grouping "Asian American and Pacific Islander" (AAPI). When researchers lump them together with the much larger "Asian American" group, their data is often not visible in the results. When governments then use this research to decide how to allocate resources, they overlook the needs of Pacific Islanders.[6]
- Some groups also critique the label "Asian American" itself because it lumps many diverse groups into one.[7] Since 1997, the U.S. Census has listed "Asian" and "Native Hawaiian or other Pacific Islander" as separate options. People filling out the Census can also choose a specific subgroup.[8]

Changes to the U.S. Census

As in the example above, the U.S. Census has often been a site for debate about race and ethnicity categories. Two changes set to take place in 2030 include:

- Adding a "Middle Eastern North African" (MENA) category for race and ethnicity.[9] Before this point, these groups have had no option to identify themselves other than "White". Similar to Pacific Islanders in the grouping AAPI, this made their data invisible. This change comes after extensive advocacy from Arab Americans.[10]

- Race and ethnicity will be asked in a single multiple-choice question instead of two separate ones. This change was made so that people who identify as Hispanic or Latino can select that option on its own. Previously, many had to choose “Some Other Race” in the separate race question.[11] However, some worry this change will make it harder to see the data of Latinos who choose more than one response, like Afro-Latinos.[12,13]

Sex and Gender Categories

Sometimes changes that are meant to provide more accuracy do the opposite instead. In 2022, NASEM released guidance on asking about sex and gender in surveys. They recommended this two-question approach (direct excerpt):[14]

> Q1: What sex were you assigned at birth, on your
> original birth certificate?
>
> - Female
> - Male
> - (Don’t know)
> - (Prefer not to answer)
>
> Q2: What is your current gender? [Mark only one]
>
> - Female
> - Male
> - Transgender
> - [If respondent is American Indian or Alaska Native] Two-Spirit
> - I use a different term: [free text]
> - (Don’t know)
> - (Prefer not to answer)

While these questions are meant to help researchers collect data on transgender participants, they fall short. Critics point out that, for example, a trans woman would be forced to choose between "transgender" and "woman." They also note that including "transgender" as an option and not "cisgender" presents the cisgender experience as "normal."[15]

The Sexual and Gender Minority Interest Group at the National Cancer Institute (NCI) published a set of revisions to NASEM's guidance. Their advice includes:[16]

- Use "man" and "woman" when talking about gender, not "male" and "female."
- Include "cisgender" as a response option when "transgender" is a given option.

Other critics of these NASEM guidelines highlight other suggestions in their paper "Queering genomics: How cisnormativity undermines genomic science":[17]

- Understand that "transgender" and "cisgender" are not genders. They are another category called *gender modalities*,[18] meaning how someone's gender relates to their gender assigned at birth.
- Never assume that "sex assigned at birth" tells you about peoples' karyotypes, or chromosomes. For example, people assigned female at birth do not always have XX chromosomes.

Even though this advice was written for people who run surveys, other types of researchers can also learn from it. Repository researchers also make choices about categorizing sex and gender data that can impact the accuracy of research.

Why we can't just give the "right answers"

While it would be nice if we could simply share the best practices for using population descriptors, we cannot. This is because:

1. The "best" way will be different depending on the project. What works for one project might not work for another.
2. As shown by these examples, this is an ongoing conversation. A simple guide would likely become outdated quickly.

Because of this, researchers must think carefully about what is right for their projects.

CHIRON Exercise 1: Representing Groups Thoughtfully helps researchers make mindful decisions about these aspects of their research.

Sources and Further Reading

1. Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, Board on Health Sciences Policy, Committee on Population, Health and Medicine Division, Division of Behavioral and Social Sciences and Education, National Academies of Sciences, Engineering, and Medicine. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. National Academies Press; 2023:26902. doi:10.17226/26902
2. Too many scientists still say Caucasian. Accessed November 11, 2025. https://www.nature.com/articles/d41586-021-02288-x
3. Sapiens. Why Do We Keep Using the Word “Caucasian”? SAPIENS. February 1, 2017. Accessed November 11, 2025. https://www.sapiens.org/culture/caucasian-terminology-origin/
4. PhenX Toolkit: Accessed November 11, 2025. https://www.phenxtoolkit.org/index.php
5. PhenX Toolkit: Protocol - Social Vulnerability. Accessed November 11, 2025. https://www.phenxtoolkit.org/protocols/view/290701?origin=search
6. Why it’s time to retire the term ‘Asian Pacific Islander’ | The Seattle Times. Accessed November 11, 2025. https://www.seattletimes.com/seattle-news/why-its-time-to-retire-the-term-asian-pacific-islander/
7. Nguyen VT. Opinion | The Beautiful, Flawed Fiction of ‘Asian American.’ *The New York Times*. https://www.nytimes.com/2021/05/31/opinion/culture/asian-american-AAPI-decolonization.html. May 31, 2021. Accessed November 11, 2025.
8. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. The White House. Accessed November 11, 2025. https://obamawhitehouse.archives.gov/node/15626
9. Revisions to OMB’s Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. Federal Register. March 29, 2024. Accessed November 11, 2025. https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and

10. Stepansky J. ‘Transformative’: US Census to add Middle Eastern, North African category. Al Jazeera. Accessed November 11, 2025. https://www.aljazeera.com/news/2024/3/28/transformative-us-census-to-add-middle-eastern-north-african-category

11. Bureau UC. What Updates to OMB&#8217;s Race/Ethnicity Standards Mean for the Census Bureau. Census.gov. Accessed November 11, 2025. https://www.census.gov/newsroom/blogs/random-samplings/2024/04/updates-race-ethnicity-standards.html

12. The new census racial categories ‘erase’ Afro Latinos. Accessed November 11, 2025. https://thehill.com/opinion/4572410-the-new-census-racial-categories-erase-afro-latinos/

13. CHC Chair Barragán Response on Update to OMB Federal Race & Ethnicity Standards | Congressional Hispanic Caucus. March 28, 2024. Accessed November 11, 2025. http://chc.house.gov/media-center/press-releases/chc-chair-barragan-response-update-omb-federal-race-ethnicity-standards

14. National Academies of Sciences, Engineering, and Medicine; Division of Behavioral and Social Sciences and Education; Committee on National Statistics; Committee on Measuring Sex, Gender Identity, and Sexual Orientation. *Measuring Sex, Gender Identity, and Sexual Orientation*. (Becker T, Chin M, Bates N, eds.). National Academies Press (US); 2022. Accessed November 11, 2025. http://www.ncbi.nlm.nih.gov/books/NBK578625/

15. Pratt-Chapman ML, Tredway K, Wheldon CW, et al. Strategies for Advancing Sexual Orientation and Gender Identity Data Collection in Cancer Research. *JCO Oncol Pract*. Published online July 2024. doi:10.1200/OP.23.00629

16. Strategies for Advancing Sexual Orientation and Gender Identity Data Collection in Cancer Research | JCO Oncology Practice. Accessed November 11, 2025. https://ascopubs.org/doi/10.1200/OP.23.00629

17. Jamal L, Zayhowski K, Berro T, Baker K. Queering genomics: How cisnormativity undermines genomic science. *Hum Genet Genomics Adv*. 2024;5(3). doi:10.1016/j.xhgg.2024.100297

18. Ashley F. ‘Trans’ is My Gender Modality: A Modest Terminological Proposal. https://www.florenceashley.com/uploads/1/2/4/4/124439164/florence_ashley_trans_is_my_gender_modality.pdf