Categories of COVID

By Anne L. Washington, Joshua Arrayales, Nicole Contaxis, Rachel S Kuo, David C Morar, Hope Muller & Molly Nystrom

Who has COVID19? This deceptively simple question relies on multiple assumptions about the production of data sources, specifically how we categorize human bodies[1]. It assumes there are clear classifications and that the classification is uniformly applied. The COVID19 pandemic is the first pandemic occurring in a data-driven era. The public has become accustomed to having access to open data sets released by governments and scientists. There is an expectation that the available open data is ready to answer questions.

In this position paper we consider how the COVID19 open data fell short of expectations based on our investigation of available data sources in the United States. Our investigation found surprising inconsistencies that challenge our understanding of the pandemic in the U.S. We suggest that without clear standards, open data efforts will have little impact on comparisons across jurisdictions. We argue that outdated categories can render advanced data technology useless in a crisis.

Whose lives matter?

As early as April 2020, scholars and opinion leaders called for better demographic data about the pandemic's impact across populations. Once collected, deaths and infection rates indicated dramatic differences across Black, Indigenous, and other people of color (BIPOC) populations. Other research found differences on survival rates between men and women. These disparities motivated us to examine the available evidence.

Our investigation started with a Data 4 Black Lives[2] spreadsheet that for a few weeks in April 2020 tracked open government data sets that contained demographic information. We started with the list as a pilot sample to learn how states and other jurisdictions organized and released demographic information. We expanded our investigation to all 50 states and the District of Columbia.

---

[1] In this paper we consider the local open data that tends to be more in-depth and detailed. These data sets are vital for communities to explore or understand local risk factors. There are four major nationwide projects that track comparative racial demographic data. • CDC Weekly Updates - https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm • Johns Hopkins University Racial Data Transparency -https://coronavirus.jhu.edu/data/racial-data-transparency • APM Research Lab: The Color of Coronavirus https://www.apmresearchlab.org/covid/deaths-by-race • COVID Racial Data Tracker - https://covidtracking.com/race

[2] Data4BlackLives.org released their list as a Google spreadsheet at https://docs.google.com/spreadsheets/d/1NFViedF47p-P0MKKl8_O0mKAhba0Yqn200EfUR4GlcQ/htmlview D4BL COVID-19 Disparities Tracker . For more about race and data see: Dixon-Román, E. (2017). Toward a hauntology on data: On the sociopolitical forces of data Assemblages. Research in Education, 98(1), 44-58. ; McGlotten, S. (2016). Black data. In E. P. Johnson (Ed.), No tea, no shade: New writings in black queer studies (pp. 262-286). Durham, NC: Duke University Press.

Who are you?

Consistent categories are essential for understanding data trends. In a public health emergency such as the current pandemic, critical decisions that could save lives are at stake. All categories are political.[3] The groupings of humans can be deeply fraught depending on who is creating groups and for what purpose. To avoid category language assumptions and to heighten awareness of the disputes of these groupings, we refer to all of these categories as body identities. In our research, we investigated common descriptions that fall into three broad groupings:

- male/female
- black/white
- age groups

-- Male or Female?

The male / female option is typically labeled sex or gender. Most states labeled this option as gender which is the more recent term. Only four states used three categories instead of the typical binary. Originally attempting to classify reproductive bodies, this binary category[4] no longer has the same salience when social media sites allow individuals to choose from over 40 categories.

-- Black or White?

The Black / White option is commonly labeled as race with Hispanic and Asian, if available, labeled as ethnicity. Descriptions of this category had race only, ethnicity only, or a combination of both. Across all states the items available in this category ranged from 2-9 choices with the exception of North Dakota which did not include any BIPOC category. The common categories in open data sets were : Black 98%, White 98%, Latino/Hispanic 92%, Asian 90%. This limits the choices of people from the Philippines[5], who defy the simple dark skin / Asian / Hispanic distinction common across the United States. Furthermore, people with Native American heritage or parents from two categories did not have as much opportunity to express their identity.

[3] For more on this idea see Geoff Bowker and Susan Leigh Starr (2000) Sorting Things Out. MIT Press; Alex Hanna, Emily Denton, Andrew Smart, Jamila Smith-Loud (2020) Towards a critical race methodology in algorithmic fairness. FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency January 2020 Pages 501–512 https://doi.org/10.1145/3351095.3372826; Deborah Thompson (2016) The Schematic State Race, Transnationalism, and the Politics of the Census. Cambridge University Press, https://doi.org/10.1017/CBO9781316442951

[4] Bivens and Haimson, 2016, Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers, https://doi.org/10.1177/2056305116672486 There are many debates about how to describe people outside this binary along questions of chromosomes, phenotypes, active hormones, internal and external reproductive organs. See one debate here: Fausto-Sterling, Anne (March–April 1993). "The five sexes: why male and female are not enough". The Science. doi:10.1002/j.2326-1951.1993.tb03081.x. How common is intersex? a response to Anne Fausto-Sterling Leonard Sax https://pubmed.ncbi.nlm.nih.gov/12476264/

[5] The Latinos of Asia : How Filipino Americans Break the Rules of Race" https://www.sup.org/books/title/?id=23819 Also Melissa Adler, Classification Along the Color Line: Excavating Racism in the Stacks; https://doi.org/10.24242/jclis.v1i1.17

-- Age?

Age ranges were inconsistent sometimes grouping by decade, some by two decades, and few starting with the same year. If the ability to represent a quantitative value is problematic, it is easy to understand why politically fraught categories are hard. Hawaii squished everyone over 60 in a single category while South Carolina had a category for the over 109 crowd. In Arizona, everyone 0-20 was in one bin while in Utah the first age group was 0-1.

Why does this matter?

Our analysis reveals that the state-level open data sets rely on categories that are mostly outdated and oversimplify many life experiences. Everyone outside of a 20-40 age range would have difficulties finding and comparing their cohort across these data sets. Parents have little comparative data about their child's cohort in neighboring states. Anyone interested in how each state cares for the elderly may be thwarted to make comparisons.

Small populations within a state, taken together across the nation, may represent a critical mass. For instance people with American Indian or Alaskan Native heritage represent 5.2 million people[6] according to the 2010 census yet in our analysis 45% of the states did not publish data with a category for this group. People who do not present within the male/female binary are a growing voice, especially amongst younger populations, and serve to represent those both born and identifying with new designations. White skinned non-reproductive male bodies were consistently classified. An early estimate suggests that 53% of mixed-race people have to either choose one category or be classified as "Other" in these data. Anyone with non-white skin or from an un-named descent has to shoe horn themselves to fit.

Pandemic data science needs better sources to achieve the promises we have come to expect from large scale analysis of open data. Inconsistent categories limit what we can understand about the pandemic across jurisdictions. Solutions are not immediately obvious for this conundrum. One approach is to support open data standards to create meaningful comparisons. We argue that this is a practical route if people represented in those data are part of the conversation. Another approach is to support crosswalk standards that translate interpretations of categories. We suggest that public health professionals plan information infrastructure that supports meaningful comparisons. Multi-stakeholder groups will provide an opportunity to recognize bias, consider embedded tradeoffs, and deliberate alternatives together.

Equity and differential impact is the hallmark of the pandemic. Without good data, disparate outcomes will be more difficult to track.

[6] National Congress of American Indians Demographics http://www.ncai.org/about-tribes/demographics