VIEWPOINT

Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research

Vence L. Bonham, JD
National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland.

Eric D. Green, MD, PhD
National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland.

Eliseo J. Pérez-Stable, MD
National Institute on Minority Health and Health Disparities, National Institutes of Health, Bethesda, Maryland.

+
Viewpoint and Editorial

Corresponding Author: Vence L. Bonham, JD, National Human Genome Research Institute, National Institutes of Health, 31 Center Dr, Room B1-B37-G, Bethesda, MD 20892-2070 (bonhamv@nih.gov).

The complexities of social identity and genetic ancestry have led to confusion and consternation related to the use and interpretation of race, ethnicity, and ancestry data in biomedical research. These discussions and overt debates have intensified with advances in genomics and knowledge about how social factors interact with biology. As more information about genomic diversity becomes available, the limitations of assigning social, political, and geographic labels to individuals become clearer; these limitations have led to growing challenges for researchers to communicate information about human genomic variation.

Imprecise use of race and ethnicity data as population descriptors in genomics research has the potential to miscommunicate the complex relationships among an individual's social identity, ancestry, socioeconomic status, and health, while also perpetuating misguided notions that discrete genetic groups exist. Self-identified race and ethnicity commonly correlate with geographical ancestry and, in turn, geographical ancestry is a contributing factor to human genomic variation. While self-identified race and ethnicity correlate with the frequency of particular genomic variants at a population level, they cannot be used exclusively to predict a patient's genotype or drug response.[1]

A recent analysis found significant heterogeneity among US clinical laboratories in the way race, ethnicity, and ancestry are ascertained; specifically, no 2 clinical laboratories used the same descriptive categories to designate a group or population on their requisition forms (C. Bustamante and A. Popejoy, written communication, August 2018). In light of the current realities, the complexity of ancestral populations requires a new approach for discussing genomics, disease risk, race and ethnicity, and social determinants of health.

In 2016, the National Human Genome Research Institute (NHGRI) and the National Institute on Minority Health and Health Disparities (NIMHD) of the US National Institutes of Health convened a workshop to discuss the use of self-identified race and ethnicity data in genomics, biomedical, and clinical research, and the implications of this use for minority health and health disparities.[2] Several major themes emerged from that workshop. For example, while the current use of the US Office of Management and Budget's (OMB's) racial and ethnic categories in research is important, there was a call for researchers to increase the scientific rigor in collecting such data, especially in clinical settings. Specifically, researchers should ensure the collected data reflect the multidimensional nature of a person's identity, especially within the context of race, ethnicity, socioeconomic status, and geographic ancestry. Further, as these data sets are curated and refined, there should be parallel efforts to standardize data collection methods. A positive step forward would involve capturing self-identified race and ethnicity data, social and cultural identity, family background, and ancestry data derived from genomic analyses. In addition, other dimensions of race should be recognized, including perceived race or ethnicity (what others believe a person to be), reflected race (the race a person believes others assume her or him to be), and the cumulative burden of discrimination. New approaches are required to minimize survey burden in the collection of such additional information because it would be a challenge to collect detailed information about each of these variables.

Another theme from the workshop was to expand beyond the traditional categories used to explain population differences.[2] Race and ethnicity are operationalized inappropriately when they serve as proxies for other demographic variables, such as an individual's socioeconomic status. One study examined the role of African ancestry and education in association with hypertension among black patients and found that having education beyond high school was significantly associated with lower systolic blood pressure, but proportion of African ancestry was not.[3] Understanding how social, demographic, and biological factors interact and affect health will require analyses that include these variables. To avoid undermining the scientific integrity of conclusions drawn from research studies, other types of data providing more nuanced insights should be collected in addition to race, ethnicity, and genetic ancestry, such as a person's educational attainment, income, and geographic residence.

The NHGRI and the NIMHD have supported work exploring how physicians and researchers collect and report race and ethnicity data as well how such data should be used for biomedical research. The NHGRI supports implementation research in the use of ancestral data in clinical genomic reports; studies have demonstrated the need to report such ancestry data to assist clinical laboratories and health professionals in interpreting the medical relevance of genomic variants. It is time for the broader scientific community to develop and adopt consensus practices for the use of race, ethnicity, social determinants of health, and ancestry data in study design, interpretation of results, publications, and medical care.

How Are Race and Ethnicity Data Currently Used in Genomics Research?

Today, racial and ethnic categories are used commonly as population descriptors in the study of genomic variation; they are also used as surrogates for ancestral background. Some researchers have identified ancestral informative markers as a tool for inferring disease risk. Similarly, some studies have inferred race and ethnicity from ancestral informative markers, which is a technique used to estimate admixture and continental genetic ancestral proportions. However, there are significant limitations to such approaches.

In addition, genome-wide association studies are designed to investigate the relationships between common genomic variants and complex disease. Such studies often use "population ancestry" data as part of the analyses using racial, ethnic, and geographical categories, among others. However, the majority of genome-wide association studies published to date have only included "European ancestral populations."[4] This lack of ancestral diversity severely restricts how the study findings can be applied clinically, for example, by incorrectly assigning genomic variants as pathogenic when their risk-conferring role may be more common in certain ancestral populations. Today, major US National Institutes of Health efforts are in place to enhance the representation of diverse ancestral populations in genomic studies by including admixed populations (eg, the Population Architecture using Genomics and Epidemiology Consortium, the Clinical Sequencing Evidence-Generating Research Program, the Trans-Omics for Precision Medicine Program, and the Implementing Genomics in Practice Program).[5]

Genetic Ancestry, Populations, and Health

Numerous studies have found that the frequency of genomic variants differs among people with different biogeographic ancestral backgrounds. This knowledge is important in understanding disease risk at the population level based on observed epidemiology. At the patient level, population studies may not correlate with disease risk and thus fail to guide the appropriate treatment for an individual patient. Patient care requires individualized treatment that moves beyond the constructs of race, ethnicity, and ancestry, and instead looks at an individual's social, behavioral, and environmental context as well as their relevant genomic features to help determine their disease risk and identify appropriate therapies.

The use of racial and ethnic categories as a surrogate for global genomic variation has significant limitations for medical care. Some major population groups, such as Latino or Hispanic, already represent an admixture of ancestry and defy the categorization of genomic variation by race. In addition, the proportion of persons who self-identify as mixed race will increase over time. Furthermore, some individuals argue that the use of these categories may reify race as a series of biological groups. There is extensive heterogeneity at the individual patient level and within specific OMB racial and ethnic categories. A new approach for examining and reporting global genomic variation, disease, and populations is needed. This landscape is far from simple. It is critical to avoid creating fictitious, discrete genomic groups while recognizing that self-identified race and ethnicity are highly associated with genetic ancestry at the continental and population level.

Differential Participation in Genomics Research

Since the completion of the Human Genome Project, there has been an attempt to improve the inclusion of individuals from diverse geographic and ancestral backgrounds in genomic studies. For example, the 1000 Genomes Project greatly expanded knowledge about the presence and distribution of common and rare genomic variants among the world's population groups that are important in health and disease.[6] To fully understand the diversity of genomic variation, the participation of people from all ancestral backgrounds is needed. Gaining complete insights about the relative roles of genomic variation, social context, and physical environment in human traits, health, and disease requires greater participation of individuals with ancestors from all regions of the world. Today, the opportunity exists to support research in diverse populations that promises to offer a scientific basis for challenging the extrapolations made by the current uses of race and ethnicity.

What Can Be Done?

Shortly after the completion of the Human Genome Project in 2003, Kaplan and Bennett proposed guidelines to follow when race and ethnicity are addressed in biomedical publications.[7] These recommendations included "When race/ethnicity is used as a study variable, the reason for its use should be specified" and "In the interpretation of racial and ethnic differences, all conceptually relevant factors should be considered."[7] Review of these guidelines might reveal a renewed importance for such guidelines, considering the fast pace of genomic advances and the new tools available to shed detailed light on the history shaping the diversity of the population. Genomic knowledge has not changed the need to move beyond the misuse of social categories of race and ethnicity as a proxy for genomic variation. The challenge that scientists and medical journal editors must address is how to report human genomic variation without inappropriately describing racial and ethnic groups as discrete population groups. It will be necessary to build consensus about how race and ethnicity data should and should not be used in biomedical research and publications. It is time to bring together diverse stakeholders with expertise to identify common ground and to help the public understand the rich diversity and common history of humankind.

ARTICLE INFORMATION

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Published Online: September 24, 2018. doi:10.1001/jama.2018.13609

Additional Contributions: We thank Lawrence C. Brody, PhD, Lucia A. Hindorff, PhD, and Teri Manolio, MD, PhD, of the National Human Genome Research Institute, National Institutes of Health for their assistance in providing feedback about this article. None received compensation for the contributions.

REFERENCES

1. Bonham VL, Callier SL, Royal CD. Will precision medicine move us beyond race? *N Engl J Med*. 2016;374(21):2003-2005. doi:10.1056/NEJMp1511294

2. National Human Genome Research Institute. *Workshop on the Use of Race and Ethnicity in Genomics and Biomedical Research*. Bethesda, MD: National Human Genome Research Institute; 2016. https://www.genome.gov/pages/about/irminorities/2016_oct_workshop_summary_and_themes.pdf.

3. Non AL, Gravlee CC, Mulligan CJ. Education, genetic ancestry, and blood pressure in African Americans and Whites. *Am J Public Health*. 2012;102(8):1559-1565. doi:10.2105/AJPH.2011.300448

4. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538(7624):161-164. doi:10.1038/538161a

5. Hindorff LA, Bonham VL, Brody LC, et al. Prioritizing diversity in human genomics research. *Nat Rev Genet*. 2018;19(3):175-185. doi:10.1038/nrg.2017.89

6. Auton A, Brooks LD, Durbin RM, et al; 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

7. Kaplan JB, Bennett T. Use of race and ethnicity in biomedical publication. *JAMA*. 2003;289(20):2709-2716. doi:10.1001/jama.289.20.2709