A Data Infrastructure for Clinical Trial Diversity

David Blumenthal, M.D., M.P.P., and Cara V. James, Ph.D.

Federal research and regulatory authorities have long sought to increase the number of people from underrepresented racial and ethnic populations who are included as participants in clinical research. Recently, the Centers for Medicare and Medicaid Services has required, in conditionally approving Aduhelm (aducanumab) for use in clinical trials, that those trials enroll representative numbers of people of color. The Food and Drug Administration has also recently issued draft guidance for clinical trial sponsors on approaches to enrolling participants who are racially and ethnically representative of the populations being studied. A critical rationale for these policies is that well-defined measures of race and ethnicity capture patient attributes — biologic, sociodemographic, cultural, and behavioral — that may help practicing clinicians anticipate how their patients' conditions will respond to tested interventions. This ability could help increase the appropriateness, and thus the equity, of care for historically marginalized groups.

Achieving this form of diversity is particularly relevant for phase 3 clinical trials, which constitute a primary source of guidance for regulators and practicing clinicians. Nevertheless, a variety of recent assessments have shown that such trials currently fall short of including representative numbers of participants from diverse racial and ethnic groups.[1-3]

There are multiple barriers to increasing participation in phase 3 clinical trials by members of underrepresented groups. Obstacles include mistrust of the clinical research enterprise, geographic and economic factors that discourage participation by communities of color, lack of inclusion of diverse communities in the planning and execution of research, and a dearth of clinical investigators from historically marginalized groups.[4]

An additional barrier, and one that requires governmental attention, is the lack of robust and reliable data on race and ethnicity in the electronic patient databases maintained by increasing numbers of clinical providers and public and private insurers. With appropriate patient consent, such databases offer fast and efficient methods of identifying people with clinical attributes of interest. These people can then be contacted to see if they are willing to participate in trials. It would greatly facilitate the development of diverse cohorts if reliable race and ethnicity data were available and searchable in these data warehouses. Often, however, they are not.

Several factors explain this lack of availability.[5] First, there is no consensus about how to define race and ethnicity, and therefore, our recorded data are not consistent. For federal research and regulatory purposes, the Office of Management and Budget (OMB) establishes definitions of races and ethnic groups. These definitions have not been revised since 1997 and are not specific enough to capture the increasing diversity of the U.S. population. The Department of Health and Human Services (HHS) implemented new, more refined definitions in 2011, but these do not apply to all HHS data sets or to data from other federal agencies, state governments, or private actors, unless those entities voluntarily adopt them. Medicare data have a particular problem, because the race and ethnicity of beneficiaries are provided by the Social Security Administration using enrollment data obtained at birth; the race and ethnicity of current seniors were thus recorded using definitions from 65 or more years ago — long before the most recent OMB standards were created. Because of deficiencies in quality and completeness, Medicaid data on race and ethnicity from 22 states are considered highly problematic or unusable.[5]

Second, health professionals, administrators, and patients may be reluctant to ask or answer questions about race and ethnicity because of their personal discomfort with the issue or fear that the data may be insecure or potentially misused.

Third, health care providers and payers that do not currently record race and ethnicity data, or that use definitions that differ from OMB or HHS standards, will incur costs in changing intake and enrollment processes and educating health professionals and administrators about how to sensitively obtain answers to the necessary questions. At a time when many provider organizations

are experiencing financial stress, they are reluctant to undertake this added burden.

Overcoming these obstacles will be difficult, but a number of public policy interventions would promote the creation of data sources from which diverse clinical trial cohorts could be more efficiently and safely assembled.[5] Examples of such interventions include the following.

The OMB, working with researchers and communities of color, could update and standardize race and ethnicity definitions and data-collection requirements for all federal agencies. New standards would not only ensure consistency within important national health care databases but also offer guidance to many private actors — including philanthropies, providers, insurers, and industry research funders — in defining key race and ethnicity terminology. In further consultation with researchers and representatives of communities of color, the OMB could provide guidance on the best way of reconciling existing data on race and ethnicity with any new definitions, so that potentially valuable existing patient information can be optimally employed.

Because even improved definitions of race and ethnicity are likely to be imperfect indicators of clinically relevant patient attributes, the federal government can also support future research to refine descriptors of race and ethnicity for use in electronic health data collection. In addition, the federal government can provide targeted administrative matching funds and technical assistance to help state Medicaid programs improve the quality and completeness of race and ethnicity data for Medicaid enrollees, using revised OMB data standards.

More broadly, federal agencies with an interest in diverse trials can develop educational resources to inform health care providers, payers, and patients about the reasons for collecting race and ethnicity data and include in these materials practical guidance for asking about this potentially sensitive information in person or on patient portals and websites.

Particularly in light of this sensitivity, federal policymakers should address comprehensively the many gaps in current privacy protections under existing laws and regulations. As burdensome as requirements under the Health Insurance Portability and Accountability Act (HIPAA) are often felt to be, this law is maladapted to the electronic age. A revised privacy and security framework could help reassure the public that race and ethnicity data recorded as part of receiving care or obtaining insurance coverage would be safe and secure. Such a framework should include substantial penalties for inadvertent or intentional misuse of patient data of all types.

Ultimately, government may also have to offer financial incentives to private providers and payers to help defray the costs of collecting race and ethnicity data at the point of care — or it may have to create regulatory requirements for collecting these data.

Though it's not a cure for all the challenges confronting phase 3 clinical trials, increasing the racial and ethnic diversity of clinical trial participants constitutes an important step toward improving trials' precision and equitable application in practice. This goal will be more efficiently and easily achieved if the infrastructure for identifying racially and ethnically diverse potential participants is available. Creating such an infrastructure will require efforts by public and private actors with an interest in equity and improving the public's health.

Disclosure forms provided by the authors are available at NEJM.org.

From the Commonwealth Fund, New York (D.B.); and Grantmakers in Health, Washington, DC (C.V.J.).

This article was published on April 27, 2022, at NEJM.org.

1. Lolic M, Araojo R, Okeke M, Woodcock J. Racial and ethnic representation in US clinical trials of new drugs and biologics, 2015–2019. JAMA 2021;326:2201-3.
2. Flores LE, Frontera WR, Andrasik MP, et al. Assessment of the inclusion of racial/ethnic minority, female, and older individuals in vaccine clinical trials. JAMA Netw Open 2021;4(2):e2037640.
3. Loree JM, Anand S, Dasari A, et al. Disparity of race reporting and representation in clinical trials leading to cancer drug approvals from 2008 to 2018. JAMA Oncol 2019;5(10):e191870.
4. Armstrong K, Ritchie C. Research participation in marginalized communities — overcoming barriers. N Engl J Med 2022;386:203-5.
5. Grantmakers in Health; National Committee for Quality Assurance. Improving data on race and ethnicity: a roadmap to measure and advance health equity. December 2021 (https://www.gih.org/publication/improving-data-on-race-and-ethnicity-a-roadmap-to-measure-and-advance-health-equity/).

DOI: 10.1056/NEJMp2201433