Race and Ethnicity Data in Electronic Health Records—Striving for Clarity

Lahia Yemane, MD; Camila M. Mateo, MD, MPH; Angel N. Desai, MD, MPH

More than 20 years ago, a report on unequal treatment by the Institute of Medicine\(^1\) made several recommendations to address racial and ethnic health disparities; chief among them was ensuring accurate and complete race and ethnicity data to identify disparities and inform interventions. Since then, electronic health records (EHRs) have become one of the main sources of population-level race and ethnicity data used for research, quality improvement efforts, and resource allocation for equitable interventions. However, despite its importance, numerous studies have demonstrated that EHR race and ethnicity data are incomplete and inaccurate, especially for minoritized populations at the highest risk for health disparities.\(^2\)

Salhi et al\(^3\) add to this growing body of literature with their retrospective cohort study. Using a statewide clinical registry, they measured the frequency of discordant EHR documentation of race and ethnicity for patients with at least 2 emergency department visits to the same hospital or health system between December 1, 2018, and November 30, 2021. Discordance was defined as a change in race and ethnicity categorization across visits. In their large sample of 403,587 adult and 121,839 pediatric patients, they found that the rate of discordance was 1.7% for adults and 7.0% for children.\(^3\) The process of race and ethnicity data collection was unknown, and authors suggested that the discordance observed could be due to variability in data collection. The authors state that this level of discordance may widen disparities or limit the ability to detect them. This may have downstream effects and highlights the need for best practices in race and ethnicity data collection.

The study by Salhi et al\(^3\) is one of only a few studies that has looked specifically at EHR discordance across emergency department visits and stratified for both adult and pediatric data. Most studies on EHR discordance have been conducted among adults and compared EHR data with self-reported surveys; pediatric studies have typically had smaller sample sizes or have been from single clinical sites. Sahil et al\(^2\) found higher rates of discordance for children vs adults, which is not surprising considering there is a lack of guidelines and consistent practices in demographic data collection among pediatric populations.\(^4\) In pediatrics there are unique challenges to address: Should the data be provided by a caregiver or child? What age is appropriate for individuals to self-report? How do we reconcile discordance among caregivers or between caregiver and the child? It may be helpful to have a designated area in the EHR to document who provided the data (eg, self, caregiver, or staff observation). Also, because multiracial children are one of the largest growing racial groups in the US, rendering clear data collection processes that have expansive race and ethnicity categories is very important.

The complexity of collecting race and ethnicity data can be used to understand how discordance may occur. Overall, these complexities can be distilled into the 5Ws (who, what, where, when, and why): who collects and provides the data, what race and ethnicity categories are available to choose from, when are data collected, where are data collected, and making clear to patients why data are collected and what will be done with them. Of note, the Salhi et al\(^3\) study was conducted on data before recent calls to more clearly describe the processes used for race and ethnicity data collection from funding agencies, health systems, and studies published in medical journals. For example, the National Committee on Quality Assurance highlighted the issue of discordance from multiple data sources and developed guiding principles on how to prioritize sources of truth and the need for data management systems to allow documentation of how data are collected to better inform reconciliation processes.\(^5\) In addition, the Centers for Medicare & Medicaid Services, the biggest payer of health care in the US, has made expanding the collection, reporting, and analysis of
standardized data, including race and ethnicity, 1 of their 5 priority areas in their Framework for Health Equity 2022-2032. Because minoritized groups are more likely than majority groups to have incomplete and inaccurate data, Centers for Medicare & Medicaid Services and others use various imputation methods to indirectly approximate data on the population level, but there is a tension between using those analytic methods over focusing on the more time-intensive and person-centered approach of improvement efforts to consistently implement the criterion standard of self-reported race and ethnicity in health care spaces.

Our society’s understanding of race and ethnicity as fluid, socially constructed identities continues to evolve, and we must adapt our data processes to this reality. In a cross-sectional study by Agawu et al., the authors measured discordant EHR documentation of race in a pediatric health system over a 9-year period and found a 6.7% change overall that occurred most commonly in multiracial groups, racial groups with smaller populations, and with increasing number of medical encounters. Similar to the study by Salhi et al., although EHR race categories did not change during that time, the process of data collection was not known in this study so they could not conclude whether change over time was due to collection methods or a true change in self-identity. Aside from the need for clarity in data collection processes, these findings bring up additional questions: What amount of discordance on an individual level may be normal and reflect changing conceptions of self-identity? How often should we ask and verify data? Also, with a person-centered approach, we cannot be ahistorical; racism in the practice of medicine exists. Race and ethnicity have been used to perpetuate false racist stereotypes about biological differences between groups. This has led to mistrust in minoritized communities about why we are collecting the data and for what purpose they will be used. With these considerations, it is critical that agencies and health systems engage minoritized community members in the creation of new practices in collecting race and ethnicity data that will enable more self-reporting. Shapiro et al. shared their enterprise-wide framework that began in 2020 and addressed each of the 5 Ws; their campaign, We Ask Because We Care, has now been adapted by many other large health systems across the country in the last 2 years. A key principle included in their framework was the need for continuous measurement and monitoring with real-time dashboards; periodic assessment of discordance rates may also serve as a quality check on implementation of processes and specific gaps for additional intervention. Kader et al. shared their early work on a community-academic-governmental collaboration, Innovations in Data Equity for All Laboratory, on improving evidence-based data collection on race and ethnicity and using community member perspectives and shared decision-making to determine what is collected and why.

As Sahil et al. illustrate, high-quality data are needed to appropriately measure and address pervasive racial and ethnic health disparities with data-driven policies that advance health equity. Now is the time to work strategically on addressing all 5 Ws in race and ethnicity EHR data collection and to collaborate with minoritized communities that historically have been left out or misclassified in our data collection and reporting processes.
REFERENCES


