Journal of the American Medical Informatics Association, 26(8-9), 2019, 730–736
doi: 10.1093/jamia/ocz113
Advance Access Publication Date: 31 July 2019
Research and Applications

Research and Applications

Challenges with quality of race and ethnicity data in observational databases

Fernanda C. G. Polubriaginof,[1,2] Patrick Ryan,[2,3] Hojjat Salmasian,[4,5] Andrea Wells Shapiro,[1] Adler Perotte,[2] Monika M. Safford,[6] George Hripcsak,[2,7] Shaun Smith,[8] Nicholas P. Tatonetti,[2] and David K. Vawdrey[1,2]

[1]Value Institute, NewYork-Presbyterian Hospital, New York, New York, USA, [2]Steele Institute for Health Innovation, Geisinger, Danville, Pennsylvania and Department of Biomedical Informatics, Columbia University, New York, New York, [3]Epidemiology Analytics, Janssen Research & Development, LLC, Titusville, New Jersey, USA, [4]Division of General Internal Medicine, Harvard Medical School, Boston, Massachusetts, USA, [5]Department of Quality and Safety, Brigham and Women's Hospital Boston, Massachusetts, USA, [6]Division of General Internal Medicine, Department of Medicine, Weill Cornell Medical College, New York, New York, USA, [7]Medical Informatics Services, Human Resources, NewYork-Presbyterian Hospital, New York, New York, USA, and [8]NewYork-Presbyterian Hospital, New York, New York, USA

Corresponding Author: David K Vawdrey, PhD, Steele Institute for Health Innovation, Geisinger, Danville, Pennsylvania and Department of Biomedical Informatics, Columbia University, New York, New York; dvawdrey@geisinger.edu

Received 3 January 2019; Revised 14 May 2019; Editorial Decision 20 May 2019; Accepted 14 June 2019

ABSTRACT

Objective: We sought to assess the quality of race and ethnicity information in observational health databases, including electronic health records (EHRs), and to propose patient self-recording as an improvement strategy.

Materials and Methods: We assessed completeness of race and ethnicity information in large observational health databases in the United States (Healthcare Cost and Utilization Project and Optum Labs), and at a single healthcare system in New York City serving a racially and ethnically diverse population. We compared race and ethnicity data collected via administrative processes with data recorded directly by respondents via paper surveys (National Health and Nutrition Examination Survey and Hospital Consumer Assessment of Healthcare Providers and Systems). Respondent-recorded data were considered the gold standard for the collection of race and ethnicity information.

Results: Among the 160 million patients from the Healthcare Cost and Utilization Project and Optum Labs datasets, race or ethnicity was unknown for 25%. Among the 2.4 million patients in the single New York City healthcare system's EHR, race or ethnicity was unknown for 57%. However, when patients directly recorded their race and ethnicity, 86% provided clinically meaningful information, and 66% of patients reported information that was discrepant with the EHR.

Discussion: Race and ethnicity data are critical to support precision medicine initiatives and to determine healthcare disparities; however, the quality of this information in observational databases is concerning. Patient self-recording through the use of patient-facing tools can substantially increase the quality of the information while engaging patients in their health.

Conclusions: Patient self-recording may improve the completeness of race and ethnicity information.

Key words: electronic health records, ethnic groups, data quality, patient-facing tools

INTRODUCTION

Race and ethnicity information has long been collected by hospitals in the United States.[1,2] These data are collected for many reasons, including for clinical, administrative, and research purposes. Clinically, race and ethnicity information are commonly used for estimating disease risk[3–5] and assessing racial and ethnic health disparities.[6–9] From an administrative standpoint, the Centers for Medicare and Medicaid Services, through the Meaningful Use incentive program, requires standardized collection of patients' race and ethnicity. This is because race and ethnicity are a part of a patient's socioeconomic status, which has been considered as a method for risk adjusting in payment reform.[10] From a research perspective, studies frequently report patients' demographic information, including race and ethnicity.

Race and ethnicity can be collected from patients in a variety of formats and by a variety of personnel. This information is often collected either verbally and documented in the EHR, or through patient-facing tools, such as intake forms completed during a clinical encounter, and then transcribed from intake forms into the EHR.[1] However, there are many challenges to the collection of race and ethnicity information that may degrade the quality of this data in the EHR.[11–16] One reason may be that cultural insensitivity and lack of understanding of the importance of race and ethnicity information are major challenges to collecting race and ethnicity information in the hospital setting. Verbally asking patients their race and ethnicity can be an uncomfortable situation for both healthcare workers and patients.[17] A previous study conducted in 2014 reported that registration personnel felt inadequately trained to ask patients' race and ethnicity.[18] Second, there is a lack of understanding of why this information is collected and how it will be used. A study conducted in 2005 showed that information that is not known to be used by others is not accurately collected.[19] Registration personnel are often unaware of the importance of race and ethnicity and also do not know who uses the information. This lack of awareness presents a barrier to registration personnel asking patients their race and ethnicity. From a patient's perspective, the question is often unexpected and may not come with an explanation of how the information will be used and why it is important.

In the United States, there has been efforts to standardize the data structures for storing race and ethnicity information. The Meaningful Use program specified that race and ethnicity data collection should follow the standard developed by the Office of Management and Budget (OMB).[20] According to this standard, race and ethnicity information can be collected in either a single-question or in a 2-question format. Race includes "American Indian or Alaska Native," "Asian," "Black or African American," "Native Hawaiian or Other Pacific Islander," or "White," and ethnicity is described as "Hispanic or Latino" or "Not Hispanic or Latino." Furthermore, the OMB also established that patient-provided information is considered the gold standard for the collection of race and ethnicity data. Given these standards and expectations, there has been little effort to compare race and ethnicity information directly recorded by patients with the corresponding data stored in the EHR.

With approval of our Institutional Review Board, we undertook a study to evaluate the completeness of race and ethnicity data nationally as well as in a large healthcare system in New York. We also analyzed how the quality of this information changed over time, as well as the impact of patients' participation in reporting their race and ethnicity directly via paper survey.

MATERIALS AND METHODS

Data

U.S. national databases

We analyzed data from 2 large observational health databases: the Healthcare Cost and Utilization Project (HCUP) and the Optum Labs Data Warehouse. The HCUP database is a hospital transactional database created by Agency for Healthcare Research and Quality that includes over 90 million inpatient, emergency visits, and ambulatory surgery encounters from multiple hospitals in the United States.[21] The Optum Labs Data Warehouse is an administrative claims database of more than 70 million commercially insured and Medicare Advantage enrollees, with greatest representation in the Midwest and South U.S. census regions.[22,23]

We examined HCUP data from 2000 to 2011 and Optum data from 2000 to 2016. The 2 databases stored race and ethnicity information using slightly different categories. Both included "White," "Black or African American," "Hispanic or Latino" and "Unknown" as categories, so we only included only these 4 options and reported the remaining groups collectively as "Other." These race and ethnicity categories were collected throughout the study period for all databases. Both databases include data from multiple institutions and each institution might collect race and ethnicity data using different approaches. Some might ask patients for their race and ethnicity information verbally while others might use registration forms; however, this information is not available as part of these datasets. A detailed description of these datasets, including the categories used to collect race and ethnicity information, sample size and timeframes, is shown in Table 1.

In addition to the HCUP and Optum databases, we examined the dataset generated from the National Health and Nutrition Examination Survey (NHANES). NHANES collects data from 5000 U.S. adults and children per year.[24] Among other information, it collects race and ethnicity in a single-question format, with the response coded as "White," "Black or African American," "Hispanic or Latino," "Not Hispanic or Latino," or "Unknown." We used NHANES data from 1999 to 2011 as a source of respondent-recorded race and ethnicity data.

Academic healthcare system in New York city

We conducted a retrospective analysis of race and ethnicity data recorded for patients that had at least 1 inpatient, outpatient, or emergency department visit from January 2014 to December 2015 at an academic health system that serves a racially and ethnically diverse population in 10 hospital campuses in and around New York City, including a quaternary care hospital. The Ambulatory Care Network consists of 14 primary care practice sites and more than 50 specialty care clinics. The academic health system provides millions of visits annually, including 2.2 million outpatient visits, 286 000 emergency department visits, and 126 000 inpatient discharges.

Race and ethnicity data were collected by the health system in 1 of 2 ways: (1) patients completed paper forms as part of the registration process, which were then transcribed into the EHR by a registration clerk or (2) registration clerks verbally asked patients about their race and ethnicity. To collect race and ethnicity, the health system used a 2-question format, the first field capturing the patient's race ("American Indian or Alaska Native," "Asian," "Black or African American," "Native Hawaiian or Other Pacific Islander," "White," "Unknown," "Other," or "Declined to Answer") and the second field capturing the patient's ethnicity ("Hispanic or Latino,"

Table 1. Description of the data sources, including time frames and race and ethnicity categories

Dataset	Description	Time frame	Patients	Black or African American (%)	White (%)	Hispanic or Latino (%)	Other race (%)	Unknown race (%)
Observational health databases								
HCUP	HCUP is a hospital transactional database that includes inpatient and emergency visits	2000-2011	91 983 358	10.75	52.13	9.40	2.42	25.31
OPTUM	Health claims database for members of United Healthcare, which includes patients enrolled in commercial plans, Medicaid, and Legacy Medicare Choice	2000-2016	73 992 364	7.45	53.21	9.69	3.64	26.00
EHR[a]	Data from the EHR of an academic healthcare system in New York City including inpatient, outpatient, and emergency department visits	2014-2015	2 338 421	6.09	23.55	9.51	10.82	59.54
Patient-provided databases								
NHANES	NHANES is a database from a national survey that includes both adult and children information	1999-2011 (biennially)	71 916	23.82	38.30	24.24	7.24	6.40
HCAHPS[a]	HCAHPS is a database from a survey regarding patient satisfaction sent to patients via U.S. mail after a hospitalization	2014-2015	25 308	9.02	63.85	5.16	12.00	13.71

[a]Numbers do not sum to 100% because race and ethnicity were collected separately.

EHR: electronic health record; HCAHPS: Hospital Consumer Assessment of Healthcare Providers and Systems; HCUP: Healthcare Cost and Utilization Project; NHANES: National Health and Nutrition Examination Survey.

"Not Hispanic or Latino," "Declined to Answer," or "Unknown"). Race and ethnicity information were collected at every encounter and stored in a centralized location in the EHR.

From the same academic health system, we examined data from the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) Survey administered to patients who had a hospital stay from January 2014 to December 2015. The HCAHPS Survey was sent via U.S. mail after hospital discharge and patients directly recorded demographics information. To collect race and ethnicity, the survey used a 2-question format, with 1 field capturing race ("White," "Black or African American," "Asian," "Native Hawaiian or other Pacific Islander," "American Indian or Alaska Native") and the second field capturing ethnicity (Hispanic or Latino origin or Not Hispanic).

Analysis

As each data source collected race and ethnicity using different categories, we described groups that were available in all data sources and reported the remaining groups collectively as "Other racial and ethnic groups." We performed descriptive statistics for "White," "Black or African American," "Hispanic or Latino," and "Other racial and ethnic groups." Respondents classified as "Unknown" or "Declined to Answer" were considered to have clinically uninformative data; we combined these categories into a larger group designated as "uninformative" for further analysis. Completeness was assessed based on the percentage of "uninformative" race and ethnicity in the database. Using EHR data, we also calculated descriptive statistics on the frequency of race and ethnicity pairs, as these 2 fields are highly correlated.

Changes in race and ethnicity information in the EHR

We analyzed changes to race and ethnicity recorded for the same patient over multiple visits, using system logs from the EHR. Patients with 2 or more visits during the study period were included in this analysis. We reported descriptive statistics on changes of race and ethnicity pairs over time.

A race and ethnicity pair was recorded during each clinical encounter. To quantify the frequency of changes recorded for a patient's race and ethnicity, each race and ethnicity pair was scored based on the amount of information it contained. Race and ethnicity pairs were broken down into concept pairs, with one concept for race and another for ethnicity, and each concept was scored individually. Each informative concept received a score of 1 and each uninformative concept received a score of 0. For example, "White," "Hispanic" would receive a score of 2 because both the race and ethnicity concepts are informative. Likewise, "White" with "Unknown" ethnicity would receive a score of 1 and "Unknown" race, "Unknown" ethnicity received a score of 0. The scores were compared for each pair chronologically.

The changes in the content of the race and ethnicity pairs were classified as information loss, neutral, or information gain. If the patient had the same score in the previous and current visit (ie, the difference between the previous and current race and ethnicity score was 0), it was considered to be neutral. If the score from the second visit was greater than the previous visit, it was considered information gain. Finally, if the score from the second visit was less than the previous visit, it was considered information loss. We reported descriptive statistics of the aggregated scores.

Comparison with respondent-recorded data

We assumed respondent-recorded data to be the reference standard for race and ethnicity data collection. To assess differences between respondent-recorded race and ethnicity information and data from observational databases, we evaluated race and ethnicity reported in the NHANES survey and the HCAHPS survey. As with the national observational databases, we reported the percentage of respondents with uninformative race and ethnicity data. Because we had patient-level data from the New York academic healthcare system from both the EHR and the HCAHPS survey, we also reported the concordance between the patient's race and ethnicity information in the EHR and the self-recorded from HCAHPS.

Comparison to census data

To assess how well the data from the EHR and the HCAHPS survey represented the population of the community in which the academic health system was located, we compared the EHR and HCAHPS race and ethnicity distribution by zip code of the patient's home address to the race and ethnicity distribution for that zip code as reported by the U.S. Census from the American Community Survey 5-Year Demographic and Housing Estimates. For each zip code, we calculated the percent difference between the U.S. Census data and the EHR data for each race and ethnicity category. For this analysis, we included zip codes that had at least 50 patients in the EHR and HCAHPS data.

RESULTS

U.S. national databases

There were 165 975 722 combined patient records in the HCUP and Optum databases. Of these, 25.3% and 26.0%, respectively, had uninformative race and ethnicity (Table 1). There were 71 916 records in the NHANES survey, and only 6.4% contained uninformative race and ethnicity.

Local healthcare system in New York city

In the New York academic health system, 2 338 421 patients had at least 1 visit during the 2-year study period. As shown in Table 1, 57.9% of patients did not have race or ethnicity identified in the EHR. The distribution of all race and ethnicity pairs is described in Table 2.

Table 2. Frequency of race and ethnicity pairs in the academic health system electronic health record

	Ethnicity		
Race	Hispanic or Latino (%)	Not Hispanic or Latino (%)	Uninformative (%)
Asian	0.05	1.28	1.78
Black or African American	0.70	3.00	2.39
White	3.27	9.51	10.77
Native Hawaiian or Pacific Islander	0.10	0.07	0.06
American Indian or Alaska Native	<0.01	0.04	0.08
Other	2.46	0.90	4.00
Uninformative	2.92	2.74	53.88

Changes in race and ethnicity information in the EHR

We identified 1 205 796 patients who had more than 1 visit to the academic health system.

There were 161 114 modifications made to race or ethnicity fields in the EHR for 147 061 distinct patients (12.0% of total population). There were 0.13 changes to race and ethnicity fields made per patient, on average, over the 2-year study period (maximum = 18).

Modifications to race or ethnicity often improved completeness (ie, a change was made from an "uninformative" concept to a specific race or ethnicity category), but this was not always the case. Overall, we observed that 47.0% of the changes made in race and ethnicity improved completeness (information gain), 35.8% of the changes resulted in information loss, and 17.2% of the changes were information neutral.

The most frequent change resulting in information gain was an update of previously documented race "Unknown" and ethnicity "Hispanic" to race "Other race" and ethnicity "Hispanic"; the most frequent change resulting in information loss was a modification from race "White" and ethnicity "Not Hispanic" to race "Unknown" and ethnicity "Unknown"; and last, the most frequent change that did not affect the amount of race and ethnicity information collected was an update from race "White" and ethnicity "Hispanic" to race "Other" and ethnicity "Hispanic."

Comparison to patient-recorded data

During the study period, there were 27 108 HCAHPS surveys completed. A small number of these patients (n = 1255, 4.9%) completed the survey more than once, and 356 patients had conflicting self-recorded race and ethnicity information. This led to 25 664 unique patients who responded to the HCAHPS survey. Excluding patients who had conflicting self-reported information, 86.3% of the remaining 25 308 patients provided informative race or ethnicity data.

There were 25 014 patients had race and ethnicity information available from HCAHPS surveys and in the EHR. Among these patients, 16 625 (66.5%) patients recorded race or ethnicity information that was discordant with data recorded in the EHR. Table 3 provides a list of the most common discrepancies between EHR and self-reported race and ethnicity data. While 6540 had both race and ethnicity as "uninformative" in the EHR, self-reported data provided meaningful race or ethnicity information for 5533 of these patients, 84.6% of patients that did not otherwise have meaningful information recorded.

Comparison with U.S. Census data

There were 44 zip codes with more than 100 patients in the EHR and HCAHPS datasets. When comparing the distribution of race and ethnicity categories among the EHR, HCAHPS, and U.S. Census datasets, we observed that, on average, the EHR data contained a higher proportion of uninformative race than the U.S. Census (63% vs 14%) (Figure 1). However, when performing the same comparison using patient-reported information, we observed that the rate of uninformative race in the HCAHPS dataset was similar to the U.S. Census dataset (18.1% vs 14%) (Figure 1). Supplementary Table 1 contains the distribution of race and ethnicity categories for each borough from zip codes included in the analysis.

DISCUSSION

Accurate collection of race and ethnicity information is key to recognizing disparities that affect racial and ethnic minority populations.[6–9] Furthermore, this information can be used to perform disease risk assessment both for individuals and populations.[3–5] Despite its importance, previous studies have reported challenges in

Table 3. The 10 most common discrepancies between EHR and self-reported data

Self-recorded race	EHR race	Self-recorded ethnicity	EHR ethnicity	Frequency (%)
White	Uninformative	Not Hispanic	Uninformative	20.54
White	White	Not Hispanic	Uninformative	19.92
White	Uninformative	Not Hispanic	Not Hispanic	6.88
Uninformative	Uninformative	Uninformative	Hispanic	4.00
Uninformative	White	Uninformative	Not Hispanic	3.06
White	Uninformative	Uninformative	Uninformative	2.98
Asian	Asian	Not Hispanic	Uninformative	2.53
Asian	Uninformative	Not Hispanic	Uninformative	2.45
Black	Uninformative	Not Hispanic	Uninformative	2.33
Uninformative	White	Uninformative	Hispanic	2.26

EHR: electronic health record.

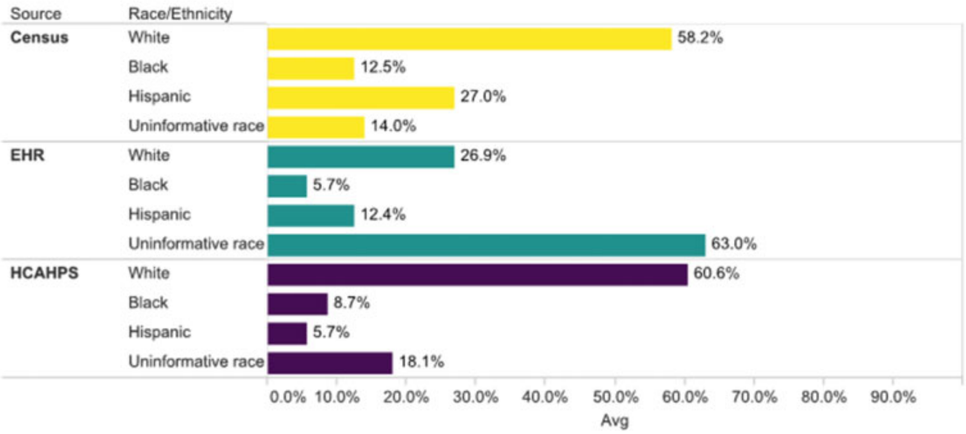

Figure 1. Comparison of the average U.S. Census, electronic health record (EHR), and Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) racial and ethnic distribution among 44 zip codes that contained at least 100 patients in the EHR and HCAHPS datasets.

collecting race and ethnicity data.[11–16] For example, a study conducted in 2015 reported data quality issues by comparing patients' race and ethnicity information from different data sources within the same institution.[15]

In our study, a large proportion of patients did not have informative documentation regarding their race and ethnicity, either in the national observational databases or in the urban academic health system. Our findings suggest that it is still challenging to capture this information despite the collection of race and ethnicity data being required as part of the U.S. Meaningful Use program. When analyzing race and ethnicity data in the EHR from a single institution, changes over time did not always improve the data quality of race and ethnicity. Indeed, information loss occurred in 35.8% of updates.

Previous studies have illuminated some of the challenges of obtaining race and ethnicity from patients in the healthcare delivery setting. First, verbally asking patients their race and ethnicity may be perceived as a sensitive topic by both hospital personnel and patients.[17,18] Second, there is a general lack of understanding of why this information is collected and how it will be used.[19] This lack of understanding poses a barrier to registration personnel asking patients their race and ethnicity. From the patients' perspective, the question is often unexpected and may not be framed with an explanation of how the information will be used and why it is important. However, when comparing the quality of patients' race and ethnicity data from the EHR and matched HCAHPS survey responses, this study found that the surveys dataset had fewer uninformative responses. This suggests that patients may be more willing to provide their information when they are able to directly record this information.

Some have argued that collecting race and ethnicity in the healthcare setting is increasingly unnecessary in the context of inexpensive genetic testing.[25] Race and ethnicity have been used in medicine as a proxy to genetics. However, it is well established that traits occur in gradients rather than in predetermined race categories. Currently, with the increased number of mixed populations, heritage can be more informative than the racial category itself. Heritage provides information about how an individual is related to others in their genealogical history, which is directly related to genomic heritage.[26] Interestingly, many Hispanic patients did not seem to consider themselves to belong to any of the OMB-defined race categories, as the majority identified their race as "Other" and their ethnicity as "Hispanic or Latino" when self-recording. Such phenomena have been previously described,[18,27–29] and this behavior raises questions about the efficacy of the 2-question format (ie, collecting race and ethnicity as separate fields) that is now widely used, as well as the meaning of the constructs of "race" and "ethnicity" for patients.

Additionally, with the improvements in genetics and the decreased cost of genetic testing, in the foreseeable future, we could rely on genetic testing instead genetic proxies for determination of disease risk. However, genetic testing availability will not facilitate the elimination of health disparities that have social determinants.

Therefore, until health equity is achieved, collection of race and ethnicity data along with other social determinants of health may still be necessary.

Until we transition to an era of ubiquitous genetic testing and health equity, one way to improve the quality and completeness of patient demographics in electronic health records is to allow patients to review and request updates to their information. In our study, both sources of respondent-recorded information, NHANES and HCAHPS, had high rates of completeness for race and ethnicity, with only 6.4% and 7.1% of the records documented as "Unknown," respectively. This finding suggests that respondents are willing to provide and directly record their race and ethnicity information when they have the opportunity to do so. Additionally, providers may document race and ethnicity information as part of the clinical note instead of using the structured fields, which might partially explain the difference in completeness between survey and observational datasets.[30]

A study conducted at one Veterans Affairs Medical Center compared patient-recorded race and ethnicity information to the data available in the EHR. Investigators mailed 300 surveys to select patients that received care primarily at the Veterans Affairs clinic. Of the completed surveys, 15.7% contained race and ethnicity information discordant from the EHR.[14] We compared race and ethnicity information available in the EHR to data from HCAHPS survey. Among patients with survey data, 86.3% provided informative race and ethnicity information and 66.5% of the answers were discordant with the EHR data. More than 84% of patients with uninformative race and ethnicity in the EHR provided meaningful information in the survey.

Patient-facing tools give patients the opportunity to fill out or review their information directly, removing some of the cultural sensitivity of having someone verbally asking for this information. A previous study demonstrated improvement in race and ethnicity data quality after using a custom patient portal application on a tablet computer to allow patients to review their demographic information.[31]

Our findings suggest that patient-facing tools that allow patients to record race and ethnicity information before, during, or after their healthcare encounters could markedly improve data quality. This could be accomplished in many ways, but one useful method is to use patient portals. When using technology to collect patient's data, it is critical to consider health and eHealth literacy of the patients. Further, healthcare institutions should be mindful of affordability and equitable access to the digital services or tools offered, preventing a potential "digital divide," which has been previously described in the context of patient portal use.[30,32,33] In addition to investigating methods for patients to self-record race and ethnicity information, future work should investigate how patients self-identify to determine how to best capture this information, and potentially revise the currently used race and ethnicity categories. Further work should also explore the discrepancies between staff-reported vs self-recorded race, as those discrepancies may demonstrate internal and external biases, which ultimately affects the quality of the information collected. Additionally, people's perceptions on how race and ethnicity information will be used might lead to discrepant answers in different scenarios.

In summary, race and ethnicity provide valuable information for precision medicine and critical information for efforts to eliminate socially determined health disparities. However, the quality of these data is concerning. While the use of genetics is not feasible at a population level, the use of patient-facing tools has the potential of dramatically improving its quality and ultimately facilitate disease risk assessment and identification and monitoring of health disparities.

This study has several limitations. The data from HCUP and Optum were deidentified; therefore, patients may be represented in the database more than once. The datasets only captured 1 race and 1 ethnicity for each patient, not capturing cases in which patients might self-identify as more than 1 race or ethnicity. Not all databases captured Native American/Alaska Natives and Pacific Islander/Native Hawaiian individuals; therefore, these individuals were treated as "Other" in the analysis. Additionally, each one of the institutions included as part of the observational databases might use a different approach for collecting race and ethnicity data, including patient-facing forms; however, this information is not available for analysis.

CONCLUSION

Our study demonstrates that collection of race and ethnicity, particularly among diverse populations, can be problematic. Poor data quality for race and ethnicity can negatively impact clinical care decisions that are based on disease risk-adjustment models incorporating race and ethnicity. Moreover, incomplete or inaccurate race and ethnicity data prevent public health professionals and policymakers from measuring and reducing racial and ethnic healthcare disparities. To address these challenges, we recommend that patient-recorded data be used to improve quality and completeness of race and ethnicity.

FUNDING

This work was supported by Agency for Healthcare Research and Quality grant numbers R01HS021816 (FCGP, DKV), R01HS023704 (HS), R01HS024713 (HS); by the National Institute of General Medical Sciences grant number R01GM107145 (NPT); and by the U.S. National Library of Medicine R01LM006910 (GH).

AUTHOR CONTRIBUTIONS

FCGP and DKV were involved in conceptualization; FCGP and DKV in methodology; FCGP, PR, HS, AP, MMS, GH, NPT, DKV, and SYW in investigation; FCGP and DKV in writing the original draft; FCGP, PR, HS, AP, MMS, GH, NPT, and DKV in writing, review, and editing; and DKV in funding acquisition.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Adler NE, Stead WW. Patients in context–EHR capture of social and behavioral determinants of health. *N Engl J Med* 2015; 372 (8): 698–701.
2. Hasnain-Wynia R, Pierce D, Pittman MA. *Who, What, When, Where: The Current State of Data of Collection on Race and Ethnicity in Hospitals*. New York: Commonwealth Fund; 2004.

3. Gail MH, Brinton LA, Byar DP, *et al*. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989; 81 (24): 1879–86.
4. Levey AS, Stevens LA, Schmid CH, *et al*. A new equation to estimate glomerular filtration rate. *Ann Intern Med* 2009; 150 (9): 604–12.
5. Stevens LA, Coresh J, Greene T, Levey AS. Assessing kidney function–measured and estimated glomerular filtration rate. *N Engl J Med* 2006; 354 (23): 2473–83.
6. Dorsey R, Graham G, Glied S, Meyers D, Clancy C, Koh H. Implementing health reform: improved data collection and the monitoring of health disparities. *Annu Rev Public Health* 2014; 35 (1): 123–38.
7. Douglas MD, Dawes DE, Holden KB, Mack D. Missed policy opportunities to advance health equity by recording demographic data in electronic health records. *Am J Public Health* 2015; 105 suppl 3: S380–8.
8. Kressin NR. Race/Ethnicity identification: vital for disparities research, quality improvement, and much more than “meets the eye”. *Medical Care* 2015; 53 (8): 663–5.
9. LaVeist TA, Gaskin D, Richard P. Estimating the economic burden of racial health inequalities in the United States. *Int J Health Serv* 2011; 41 (2): 231–8.
10. Buntin MB, Ayanian JZ. Social risk factors and equity in medicare payment. *N Engl J Med* 2017; 376 (6): 507–10.
11. Blustein J. The reliability of racial classifications in hospital discharge abstract data. *Am J Public Health* 1994; 84 (6): 1018–21.
12. Chakkalakal RJ, Green JC, Krumholz HM, Nallamothu BK. Standardized data collection practices and the racial/ethnic distribution of hospitalized patients. *Med Care* 2015; 53 (8): 666–72.
13. Gomez SL, Glaser SL. Misclassification of race/ethnicity in a Population-based Cancer Registry (United States). *Cancer Causes Control* 2006; 17 (6): 771–81.
14. Hamilton NS, Edelman D, Weinberger M, Jackson GL. Concordance between self-reported race/ethnicity and that recorded in a Veteran Affairs electronic medical record. *N C Med J* 2009; 70 (4): 296–300.
15. Lee SJC, Grobe JE, Tiro JA. Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals. *J Am Med Inform Assoc* 2016; 23 (3): 627–34.
16. Moscou S, Anderson MR, Kaplan JB, Valencia L. Validity of racial/ethnic classifications in medical records data: an exploratory study. *Am J Public Health* 2003; 93 (7): 1084–6.
17. Baker DW, Hasnain-Wynia R, Kandula NR, Thompson JA, Brown ER. Attitudes toward health care providers, collecting information about patients’ race, ethnicity, and language. *Med Care* 2007; 45 (11): 1034–42.
18. Berry C, Kaplan SA, Mijanovich T, Mayer A. Moving to patient reported collection of race and ethnicity data. *International J Health Care QA* 2014; 27 (4): 271–83.
19. Nelson NC, Evans RS, Samore MH, Gardner RM. Detection and prevention of medication errors using real-time bedside nurse charting. *J Am Med Inform Assoc* 2005; 12 (4): 390–7.
20. Office of Management and Budget. Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. *Statistical Policy Directive No 15*. October 1997. Washington, DC: Government Printing Office; 1997.
21. Healthcare Cost and Utilization Project (HCUP). Healthcare Cost and Utilization Project. https://www.hcup-us.ahrq.gov. Accessed August 21, 2017.
22. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Bleicher PD, Crown WH. Optum Labs: building a novel node in the learning health care system. *Health Aff (Millwood)* 2014; 33 (7): 1187–94.
23. Optum. Optum Data Assets. https://www.optum.com/content/dam/optum/resources/productSheets/5302_Data_Assets_Chart_Sheet_ISPOR.pdf. Accessed January 10, 2018.
24. Centers for Disease Control and Prevention. *National Health and Nutrition Examination Survey (NHANES). Vol. 2007*. Hyattsville, MD: National Center for Health Statistics; 2007.
25. Ng PC, Zhao Q, Levy S, Strausberg RL, Venter JC. Individual genomes instead of race for personalized medicine. *Clin Pharmacol Ther* 2008; 84 (3): 306–9.
26. Yudell M, Roberts D, DeSalle R, Tishkoff S. SCIENCE AND SOCIETY. Taking race out of human genetics. *Science* 2016; 351 (6273): 564–5.
27. Markus HR. Pride, prejudice, and ambivalence: toward a unified theory of race and ethnicity. *Am Psychol* 2008; 63 (8): 651–70.
28. Robbin A. The problematic status of US statistics on race and ethnicity: An “imperfect representation of reality. *J Gov Inf* 1999; 26 (5): 467–83.
29. Bhalla R, Yongue BG, Currie BP. Standardizing race, ethnicity, and preferred language data collection in hospital information systems: results and implications for healthcare delivery and policy. *J Healthc Qual* 2012; 34 (2): 44–52.
30. Sholle ET, Pinheiro LC, Adekkanattu P, *et al*. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *J Am Med Inform Assoc* 2019; 94 (8): 666–8.
31. Polubriaginof F, Salmasian H, Shapiro AW, *et al*. Patient-provided data improves race and ethnicity data quality in electronic health records. 2016. https://knowledge.amia.org/amia-63300-1.3360278/t001-1.3365273/f001-1.3365274/2498793-1.3365423/2498793-1.3365424.
32. Antonio MG, Petrovskaya O, Lau F. Is research on patient portals attuned to health equity? A scoping review. *J Am Med Inform Assoc* 2019; 27 (9650): 167–13.
33. Ancker JS, Hafeez B, Kaushal R. Socioeconomic disparities in adoption of personal health records over time. *Am J Manag Care* 2016; 22 (8): 539–40.