Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals

RECEIVED 7 July 2015
REVISED 8 September 2015
ACCEPTED 14 September 2015
PUBLISHED ONLINE FIRST 11 December 2015

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD UNIVERSITY PRESS

Simon J Craddock Lee[1,2], James E Grobe[1], Jasmin A Tiro[1,2]

ABSTRACT

Background Measurement of patient race/ethnicity in electronic health records is mandated and important for tracking health disparities.

Objective Characterize the quality of race/ethnicity data collection efforts.

Methods For all cancer patients diagnosed (2007–2010) at two hospitals, we extracted demographic data from five sources: 1) a university hospital cancer registry, 2) a university electronic medical record (EMR), 3) a community hospital cancer registry, 4) a community EMR, and 5) a joint clinical research registry. The patients whose data we examined ($N = 17\,834$) contributed 41 025 entries (range: 2–5 per patient across sources), and the source comparisons generated 1–10 unique pairs per patient. We used generalized estimating equations, chi-squares tests, and kappas estimates to assess data availability and agreement.

Results Compared to sex and insurance status, race/ethnicity information was significantly less likely to be available ($\chi^2 > 8043$, $P < .001$), with variation across sources ($\chi^2 > 10\,589$, $P < .001$). The university EMR had a high prevalence of "Unknown" values. Aggregate kappa estimates across the sources was 0.45 (95% confidence interval, 0.45–0.45; $N = 31\,276$ unique pairs), but improved in sensitivity analyses that excluded the university EMR source ($\kappa = 0.89$). Race/ethnicity data were in complete agreement for only 6988 patients (39.2%). Pairs with a "Black" data value in one of the sources had the highest agreement (95.3%), whereas pairs with an "Other" value exhibited the lowest agreement across sources (11.1%).

Discussion Our findings suggest that high-quality race/ethnicity data are attainable. Many of the "errors" in race/ethnicity data are caused by missing or "Unknown" data values.

Conclusions To facilitate transparent reporting of healthcare delivery outcomes by race/ethnicity, healthcare systems need to monitor and enforce race/ethnicity data collection standards.

Keywords: electronic medical record, data quality, race and ethnicity, cancer registry

BACKGROUND AND SIGNIFICANCE

In the United States, patient race and ethnicity are used as universal demographic variables in health and medical research.[1] Collecting racial and ethnic information is crucial for tracking disparities in healthcare, evaluating race or ethnicity as a potential confounder, understanding how sociocultural factors relate to disease, and facilitating quality improvement efforts to address racial/ethnic disparities in healthcare delivery.[2–4] Irrespective of debates over how to best conceptualize and measure race and ethnicity,[5] it is critically important to characterize the quality of the existing data on these demographics to determine whether current data collection standards are being implemented appropriately or need reinforcement.[5–13]

Large computerized databases such as cancer registries and electronic medical records (EMRs) have changed the landscape of healthcare practice, policy, and research. Increasing emphasis on the "meaningful use" of electronic health information technology has increased the utilization of these data sources for multiple purposes.[14] For example, hospitals are now required to provide race and ethnicity information on all patients with a cancer diagnosis, and not-for-profit hospitals commonly integrate race and ethnicity data into their community needs assessments for community-benefit reporting.[15] Unfortunately, the quality of these data is unclear.[16,17] Without establishing the quality of the data collected, health systems are limited in their ability to draw inferences about the association of race and ethnicity with various healthcare outcomes. Moreover, not knowing the quality of the data collected may hamper quality improvement initiatives that want to use these data to demonstrate transparency and accountability to the clinical populations and community stakeholders they serve.[18]

Within healthcare systems, patient identity data on demographic information such as race/ethnicity, insurance status, and sex are routinely collected at various points of patient intake and registration in both ambulatory and in-patient settings. For a single institution, we would expect there to be agreement between patient identity data from different clinical information systems, such as an EMR and a cancer registry, which takes its baseline from the parent EMR and routinely supplements it with data collected by patient surveys and telephone interviews. Although important work has been done to examine data accuracy (eg, clinical information system data comparison to self-report), prior studies have only analyzed a central data source or a single setting.[11,17,19] Furthermore, studies that have examined central data sources, like state cancer registries or the Surveillance, Epidemiology, and End Results (SEER) program, have not pursued subanalyses down to the level of the contributing facility.[6,20] To our knowledge, prior studies have not compared patient race/ethnicity data across multiple data sources and settings.

Correspondence to Simon J Craddock Lee, PhD, MPH, Department of Clinical Sciences, University of Texas, Southwestern Medical Center, Harold C. Simmons Comprehensive Cancer Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9066, USA; simoncraddock.lee@utsouthwestern.edu; Tel: (214) 648-2410

For numbered affiliations see end of article.

OBJECTIVE

This study assessed the quality of race and ethnicity data across multiple clinical information system databases. Data quality can be examined on multiple dimensions, and various taxonomies have been suggested to guide quality assessments.[21] We assessed the extent of data availability, or completeness, along with agreement, or concordance, across five clinical information systems that store race and ethnicity data for two healthcare institutions. In addition, we characterized the types of discrepancies in race and ethnicity data observed across data sources as well as potential comparators, including sex and insurance status.

METHODS

Setting and Data Sources

Leveraging the state-mandated cancer surveillance function of certified cancer registries,[22,23] we utilized two major hospital cancer registries in the Dallas metropolitan area: one affiliated with a university hospital, the other affiliated with a community-based safety-net hospital. We also pulled data from the respective institutions' EMRs as well as the information system that tracks clinical research enrollment at both institutions. The fact that a significant number of safety-net patients cross over to the university hospital for cancer care presented us with a rare opportunity to examine data agreement across multiple institutional data sources. This study was approved by the University of Texas Southwestern Institutional Review Board (STU# 032011-133) and conducted with the approval of the Parkland Office of Research Administration.

Abstraction Process

Each registry supplied demographic records for all cancer patients diagnosed from January 1, 2007 to December 31, 2010; this date range was selected to avoid any policy-driven changes associated with Affordable Care Act reform. We used unique patient identification numbers to request the following demographic data from the EMRs and clinical research databases: race, ethnicity, sex, age, marital status, and insurance status. This process resulted in five sets of demographic data for the cancer patient population: 1) the university cancer registry, 2) the university hospital EMR, 3) the community cancer registry, 4) the community hospital EMR, and 5) the clinical research registry. To remove duplicate entries in each database, a combination of variables, including medical record number, name, and date of birth, was used to identify unique patients, and then link them across data sources. Patients were excluded if they had entries in only one source (ie, cancer registry patients with no matching data in the other four sources). Thus, a single patient could have distinct demographic information in two to five sources.

Coding Race/Ethnicity

Protocols varied for how staff at each institution recorded demographic information, including race and ethnicity.[24] US Census guidelines promulgated by the US Office of Management and Budget (OMB) recommend treating ethnicity (eg, Hispanic vs non-Hispanic) as a separate variable from race, but allowed for a combined format when necessary.[25] In our study, of the 5981 patients coded as "Hispanic" in the ethnicity category across the five databases, most (98%) had a race category code of "White." In addition, "Hispanic" repeatedly appeared as a data value recorded in the race field. Therefore, to allow for comparable variable coding across institutions and to simplify analyses, we aggregated race and ethnicity into one variable, RE.[26] The aggregate coding system ranked data values such that Hispanic ethnicity trumped other racial groups (ie, race information was dropped for the Hispanic data entries), resulting in the following values: Hispanic, non-Hispanic White, non-Hispanic Black, non-Hispanic Asian, Other, or Unknown.

Statistical Analyses

To examine associations between the quality of race/ethnicity data and other variables (eg, data source), we used generalized estimating equations, a population-based, generalized, linear mixed modeling approach. Generalized estimating equations allow for the appropriate specification of distributions (eg, logistic regression analyses for the binary variable data availability) and takes into account the complex, correlated structure of the data that results when a single patient can contribute to multiple data sources.[27,28]

To examine data availability, we coded the presence/absence (1/0) of information for a given demographic variable (eg, race/ethnicity, sex) for each unique patient across all five data sources and computed the proportion of patients with data available from each data source. Variables that were unpopulated or coded as "Unknown" were classified as absent, and the value "Other" was coded as present. The data availability of other demographic variables was also analyzed, to compare against the race/ethnicity results.

To evaluate concordance, we compared the available race/ethnic data values for each patient across multiple sources. We chose this method because the data sources did not have consistent policies or protocols governing how staff should collect race/ethnicity data (eg, patient self-reporting is the recognized gold standard of such data collection).[11] Thus, the data quality of a single source could only be evaluated against the aggregate of all the other sources.[29,30] Patients could have information in two to five sources, creating 1 to 10 unique pairs for comparison. For each patient, concordance was calculated as a simple proportion (the total number of concordant pairs divided by the total number of unique pairs).[31] Aggregate kappa statistics were calculated by taking the weighted mean of all the pairwise kappa values.[31] A sensitivity analysis was performed to examine kappa estimates with and without the unpopulated/"Unknown" data values.

To characterize disagreement in race/ethnicity data, we created a summary graph illustrating the proportion and type of disagreements associated with each race/ethnicity category.[29,31] In this case, a pair was the unit of analysis, not a unique patient. For example, a patient contributing information from five sources may have, respectively, a race/ethnicity description of White, Black, White, Black, and Other, representing two concordant pairs and eight discordant pairs. We calculated two proportions specific to each race/ethnic category – overall disagreement and race/ethnicity pair-specific disagreement. In both calculations, the denominator was the total number of observed pairs in which a particular race/ethnicity category was represented. For overall disagreement, the numerator was the total number of discordant pairs. For each race/ethnicity pair-specific disagreement, the numerator was the number of pairs with that race/ethnicity pattern (eg, the proportion Black-Hispanic disagreement = the total number of Black-Hispanic pairs divided by the total number of pairs with Black listed at least once in a pair).[31,32]

RESULTS

Study Population

Table 1 summarizes the derivation of the final sample and the overlap between sources. We excluded 62 patients for whom data appeared in only one data source and, thus, did not have sufficient data available for comparison. The merged dataset contained 17 834 unique patients that contributed 41 025 data entries (range: 2–5 entries for each patient, from each source). By linking entries associated with the

Table 1: Derivation of study sample and patient overlap across the five data sources

Derivation process	Source					Total
	1	2	3	4	5	
	University cancer registry	University hospital EMR	Community cancer registry	Community hospital EMR	Clinical research registry	
Initial number of entries	14 704		6166			
Cleaned number of entries[a]	14 122		5932			
Match with other data sources[b]	14 078	14 060	5914	5909	1064	41 025 unique entries
Pairwise overlap with each source						31,276 unique pairs
With source 2 – university hospital EMR	14 060					
With source 3 – community cancer registry	2158	2143				
With source 4 – community hospital EMR	2153	2139	5909			
With source 5 – clinical research registry	934	930	425	425		
Overlap across five sources[c]						17,834 unique patients
In only two sources	14 911 (83.6%) contributing one pair each					
In only three sources	783 (4.4%) contributing three pairs each					
In only four sources	1846 (10.4%) contributing six pairs each					
In all five sources	294 (1.6%) contributing 10 pairs each					

EMR, electronic medical record. [a]Erroneous and duplicate entries removed. [b]Entries that had at least one valid match in another source, based on patient identifying information. [c]Counts of unique patients by how many sources contributed demographic data for the patient.

same patient, we identified 31 276 pair-wise comparisons of race/ethnicity information across sources, with 2923 patients (16.4%) contributing entries from three or more sources. Table 2 presents the demographic characteristics of the samples obtained from each source and aggregated across all sources.

Availability

As shown in Figure 1, data availability varied significantly by source and demographic variable. Compared with sex and insurance status, the race/ethnicity variables were significantly less likely to be available in the data sources ($\chi^2 > 8043$, $P < 0.001$). Furthermore, the availability of race/ethnicity data varied significantly across sources ($\chi^2 > 10\ 589$, $P < 0.001$), with a more than three-fold difference in rates between the university hospital EMR and the other four sources (28.1% vs 96.2%; relative risk: 3.4, 95% confidence interval [CI], 3.3-3.5). The interaction between a demographic variable and a source was significant, such that race/ethnicity data availability varied markedly across sources, whereas sex and insurance data availability were consistently high across all sources ($\chi^2 > 1857$, $P < 0.001$). In addition to variation in whether a race/ethnicity variable was available, the number of different race and ethnicity categories themselves varied across sources (data not shown).

Concordance

Initial analyses, using all five data sources, suggested poor agreement in race/ethnicity coding. For the total sample, the proportion of race/ethnicity concordant pairs was 0.54 (95% CI, 0.53-0.55). When "Unknown" values from all five sources were included, the aggregate kappa estimate for race/ethnicity agreement was 0.45 (95% CI, 0.45-0.45; $N = 31\ 276$ unique pairs). In contrast, the kappa statistic for sex was near perfect, at 0.98, and the kappa statistic for insurance was low, at 0.38 (95% CI, 0.38-0.39).

Table 3 summarizes aggregate kappa estimates and 95% confidence intervals for each data source, including and excluding "Unknown" data values (shaded and unshaded values, respectively), as well as kappa estimates for each source pair. Analyses controlling for data source revealed significant variation across sources ($\chi^2 > 16\ 703$, $P < 0.001$), with the university hospital EMR having significantly lower agreement with the other sources, predominately due to the number of patients with "Unknown" race/ethnicity values. Pairwise agreement scores involving the university hospital EMR and including "Unknown" values were poor, ranging from 0.16-0.23 (Table 3). In contrast, pairwise agreement scores not involving the university hospital EMR were generally high, ranging from 0.85-0.92. Aggregate kappa estimates dramatically improved when we conducted sensitivity analyses that either excluded "Unknown" values (unshaded values in Table 3) or excluded the university hospital EMR data source ($\kappa = 0.89$, 95% CI, 0.88-0.90). However, adding this exclusion criterion eliminated over 44% of the total race/ethnicity pairs. Supplemental analyses also examined the more granular coding, beyond the commonly recognized OMB categories available in the two cancer registries (25 different categories) and found a high level of agreement of 0.85 (95% CI, 0.84-0.87) and a kappa statistic of 0.80 (95% CI, 0.78-0.81).

Characterizing Race-Ethnicity Disagreement

Of the 31 276 possible pairs, 13 905 included one or more "Unknown" values. Given the above kappa statistic findings, we chose to exclude pairs that included an "Unknown" value. Figure 2 illustrates the distribution of agreement and disagreement for each race/ethnicity pair by

RESEARCH AND APPLICATIONS

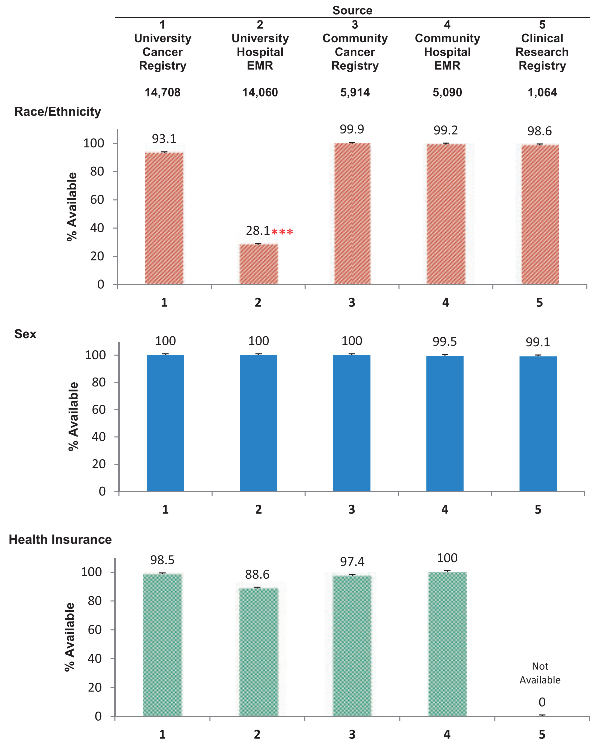
Table 2: Characteristics of patient demographic entries across data sources

Characteristics	Source					All entries[a]
	1	2	3	4	5	
	University cancer registry	University hospital EMR	Community cancer registry	Community hospital EMR	Clinical research registry	
	$N = 14\,708$	$N = 14\,060$	$N = 5914$	$N = 5090$	$N = 1064$	$N = 41\,025$
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
Race/Ethnicity						
White	8688 (61.7)	2579 (18.3)	1482 (25.1)	1535 (26.0)	597 (56.1)	14 881 (36.3)
Black	2174 (15.4)	635 (4.5)	2154 (36.4)	2194 (37.1)	253 (23.8)	7410 (18.1)
Asian	499 (3.5)	136 (1.0)	273 (4.6)	271 (4.6)	36 (3.4)	1215 (3.0)
Hispanic	1621 (11.5)	394 (2.8)	1955 (33.1)	1852 (31.3)	159 (14.9)	5981 (14.6)
Other	128 (0.9)	207 (1.5)	207 (0.8)	12 (0.2)	4 (0.4)	397 (1.0)
Unknown	968 (6.9)	10 109 (71.9)	4 (0.1)	45 (0.8)	15 (1.4)	11 141 (27.2)
Sex						
Female	6722 (47.7)	6728 (47.9)	3122 (52.8)	3111 (52.6)	639 (60.1)	20 322 (49.5)
Male	7356 (52.3)	7330 (52.1)	2791 (47.2)	2770 (46.9)	415 (39.0)	20 662 (50.4)
Unknown	0 (0.0)	2 (0.0)	1 (0.0)	28 (0.5)	10 (0.9)	41 (0.1)
Insurance status						
No insurance	1074 (7.6)	1396 (9.9)	3781 (63.9)	483 (8.2)		6734 (16.9)
Charity	0 (0.0)	349 (2.5)	0 (0.0)	2376 (40.2)		2725 (6.8)
Medicaid	683 (4.9)	460 (3.3)	698 (11.8)	1315 (22.3)		3156 (7.9)
Medicare	4414 (31.4)	3969 (28.2)	942 (15.9)	1379 (23.3)		10 704 (26.8)
Private/Military	7924 (54.6)	6282 (44.7)	338 (5.7)	356 (6.1)		14 668 (36.7)
Unknown	215 (1.5)	1604 (11.4)	155 (2.6)	0 (0.0)		1974 (4.9)
Marital status						
Single		3366 (23.9)	2252 (38.1)	2184 (37.0)		7802 (30.1)
Married/Partner		6061 (43.1)	2033 (34.4)	1980 (33.5)		10 074 (38.9)
Separated/Divorced/ Widowed		1320 (9.4)	1495 (25.2)	1699 (28.8)		4514 (17.5)
Unknown		3313 (23.6)	134 (2.3)	46 (0.8)		3493 (13.5)
Age						
<18	252 (1.8)	227 (1.6)	10 (0.2)	12 (0.2)		501 (1.3)
18–40	1212 (8.6)	1205 (8.6)	877 (14.8)	842 (14.2)		4136 (10.4)
41–60	4992 (35.5)	4824 (34.3)	2907 (49.2)	2848 (48.2)		15 571 (39.0)
61–80	6461 (45.9)	6561 (46.7)	1964 (33.2)	2012 (34.0)		16 998 (42.5)
>80	1161 (8.2)	1243 (8.8)	156 (2.6)	167 (2.8)		2727 (6.8)
Unknown	0 (0.0)	0 (0.0)	0 (0.0)	28 (0.5)		28 (0.1)

EMR, electronic medical record. Grayed boxes indicate demographic variables that were not reported by the data source. [a]The total number of entries is lower for demographic fields in which a source did not report on that variable (eg, $N = 39\,961$ for age).

each category (eg, the percentage of White-White concordant pairs and the percentage of White-Black discordant pairs among all pairs that included “White” as a data value). Further analyses of the distribution found that only 39.2% (6988 patients) were in complete agreement across all sources (eg, all pairwise comparisons were concordant). Pairs that included a “Black” data value exhibited the highest agreement (95.3%), whereas those pairs that included an “Other” value exhibited the lowest agreement across sources (11.1%).

Figure 1: Proportion of patients with availability of race/ethnicity, sex, and health insurance variables across the five data sources. Panel 1: Race/Ethnicity. Panel 2: Sex. Panel 3: Health Insurance. Each vertical bar represents point estimates for a specific source, with 95% confidence intervals denoted by the error bars. Source 5 did not record insurance information. *** Denotes that data availability for a given source is significantly different from the other sources ($P < .001$).



Patients coded as White, Hispanic, or Asian in one or more sources had similar levels of agreement (range: 80.8–87.4%). Some race/ethnicity disagreement patterns were more common than others (eg, Black-Hispanic discordant vs Black-White discordant).

In Figure 2, agreement (green) and disagreement rates for each race/ethnicity category are plotted. The calculations of these rates are based on the proportion of concordant or discordant pairs divided by all of the pairs containing that race/ethnicity category (RE); thus, discordant pairs are represented in both race/ethnicity categories (eg, a Black-Hispanic pair is counted in both the Black and Hispanic bars).

DISCUSSION

The availability of race and ethnicity data and agreement between the five data sources varied markedly, primarily due to significantly poor performance by the university hospital EMR data source on these metrics. Further, the degree of data collection for race and ethnicity also varied compared with other variables that are important to healthcare delivery; we found a three-fold difference in availability when comparing documentation of race/ethnicity information to information on sex and insurance status between the best- and worst-performing data sources. We also found high levels of disagreement in race/ethnicity information, primarily among patients categorized as “Unknown” or “Other.” Excluding patients whose race/ethnicity was classified as “Unknown” from the sensitivity analyses substantially increased agreement but reduced the sample size by almost half. Further, pervasive “Unknown”, or unpopulated, race and ethnicity fields result in data quality that was worse than if the data had been generated at random (ie, kappa estimates were <0.5). These findings affirm prior studies that examined missing data[7,33] and illustrate the loss of information that results when health systems do not prioritize consistent race/ethnicity data collection practices.

Examining the patterns of disagreement between data sources revealed that the data for individuals coded as “Other” (the least frequent category to appear in the data) were of the poorest quality and may not align with the common aggregate categories promulgated by OMB (Hispanic/non-Hispanic, American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White).[34] That is, use of the category “Other” may represent uncertainty on the part of data collectors or resistance by patients or families to using the common categories offered. The frequency of mismatches between “White” and “Hispanic” lends support to recommendations for changing how information on Hispanic ethnicity is elicited from patients.[30,35,36] Collectively, our findings suggest that the categorization of patients’ racial/ethnic identity might be better accomplished using a more granular and robust coding system (eg, code multiple race and/or ethnicity categories), as suggested by national efforts toward the standardization of racial and ethnicity tracking.[2,37,38]

In our study, the cancer registries recorded more granular race and ethnicity information, with over 21 different race/ethnicity terms present in the data (eg, Thai, Dominican, and “Guam”). After we aggregated and simplified the race/ethnicity coding schemes across all of the different sources,[21,37] the data quality remained good. Collecting more granular data at the point of service and recording it in the patient’s EMR could better characterize all individuals and describe local variations in terminology.[38] Standardized concatenation of granular data into simpler aggregates by institutional health information management would permit broader data analyses and reporting.[39] Common standards would also allow for greater interoperability of data fields between institutions.[40] The ability to track such granular information can be important for public health interventions when there are population-level, inter- and intra-ethnic disparities at work.[41–45]

Although collecting race/ethnicity data has been mandated by civil rights and Medicare legislation, institutional incentives and the enforcement of data collection policies needed to drive staff behavior vary. For instance, the lack of attention the university hospital, where most patients were privately insured, gives to collecting race and ethnicity data is apparent from that institution’s EMR data. In contrast, communication with organizational leaders at the community hospital identified local institutional practices that emphasized routine and consistent collection of race and ethnicity information from patients. For example, the care of a majority of patients at the community hospital was financed through local medical assistance or other public programs, and local funding authorities expect race/ethnicity population aggregation to be a part of routine reporting from hospitals. The variation we found across sources, together with patterns across other sociodemographic information (sex, insurance status), suggests that the actions of external forces that enforce data collection, such as state regulation of professional accreditation, are important to achieve better data quality.

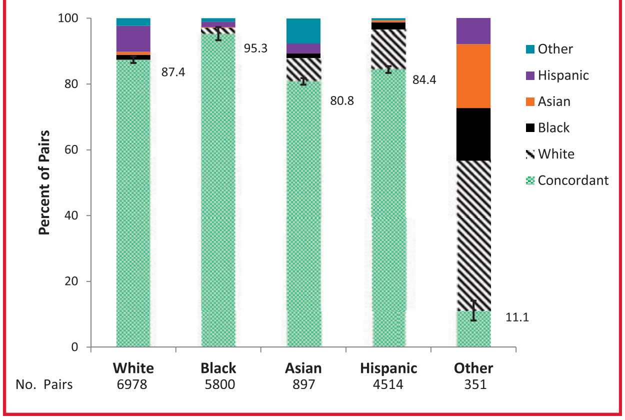
Race and ethnicity data collection is effectively voluntary for most health systems, despite numerous system-level policy statements

Table 3. Aggregate and pairwise kappa estimates and 95% confidence intervals of race/ethnicity variable agreement by data source pair, including pairs with unknown race/ethnicity (bold) and excluding pairs with an unknown race/ethnicity (*italic*).

			Pairwise Kappa[b] between Data Sources									
			1		2		3		4		5	
	Aggregate Kappa[a] (95% CI)		University Cancer Registry		University Hospital EMR		Community Cancer Registry		Community Hospital EMR		Clinical Research Registry	
1 University Cancer Registry	*0.87*	(.87-.87)			*0.84*	(.83-.86)	*0.88*	(.86-.89)	*0.89*	(.87-.90)	*0.9*	(.87-.93)
		8952				3795		2137		2118		902
	0.35	(.35-.35)										
		19305										
2 University Hospital EMR	*0.86*	(.86-.86)					*0.91*	(.88-.93)	*0.93*	(.91-.96)	*0.86*	(.80-.91)
		5508						682		678		353
	0.18	(.18-.18)	**0.16**	(.15-.17)								
		19272		14060								
3 Community Cancer Registry	*0.91*	(.91-.91)							*0.93*	(.92-.93)	*0.91*	(.87-.94)
		9104								5863		422
	0.76	(.76-.77)	**0.86**	(.85-.88)	**0.22**	(.21-.24)						
		10635		2158		2143						
4 Community Hospital EMR	*0.92*	(.92-.92)									*0.93*	(.90-.96)
		9080										421
	0.77	(.76-.77)	**0.87**	(.85-.88)	**0.23**	(.21-.25)	**0.92**	(.91-.93)				
		10626		2153		2139		5909				
5 Clinical Research Registry	*0.9*	(.90-.90)										
		2098										
	0.65	(.64-.66)	**0.85**	(.81-.88)	**0.21**	(.19-.24)	**0.9**	(.86-.93)	**0.92**	(.88-.95)		
		2714		934		930		425		425		

CI, confidence interval; EMR, electronic medical record. [a]Aggregate kappa estimate was calculated by taking the weighted mean of all the pairwise kappa values across the five data sources. [b]Pairwise kappa estimates illustrate the agreement between two data sources.

Figure 2: Disagreement analyses stratified by race-ethnicity category, excluding pairs containing an “Unknown” value. 95% CI values denote error bars.



stipulating collection practices.[24,35] Our findings point to the need for stronger industry controls that require institutions to enforce policies regarding the day-to-day collection of race and ethnicity information from patients.[46] Cancer registry accreditation fosters adherence to minimal acceptable standards of data collection and, consequently, may produce race and ethnicity data of better quality, as our study suggests. Quality improvement initiatives should, similarly, incorporate routine audits of race-ethnicity data collection practices to strengthen existing systems of checks and balances that are appropriate to the particular provider setting and the communities it serves.

A key limitation of our study concerns generalizability, because our design drew from only two institutions in one locality. However, the healthcare facilities that contributed data to this study serve a densely-populated and diverse metropolitan area, which produced a robust study sample with ample variability in race and ethnicity. As institutions with independent governance structures with substantial overlap as regards cancer care, these sites present an important opportunity to demonstrate the effects of variation in administrative data collection practices. Additionally, although the aggregation of race and ethnicity data was justified both analytically and by the demographics of our study population, doing so could result in the loss of significant information about more heterogeneous Hispanic populations (eg, Dominican Americans, who may identify as ethnically Hispanic and racially Black).

Race and ethnicity data are fraught with social meanings that directly influence the social processes by which health system staff elicit and record these data.[24] Clinical staffs have reported several barriers to collecting race/ethnicity information from patients, including concerns about patient privacy, the legality of soliciting such information, and potential resistance from patients.[36,47] Our analyses suggest that possible bias toward the imputation of "race" according to perceived phenotypes can influence which data categories are more likely to be populated.[48] Developing institutional policies and training procedures to implement consistent patient self-reporting and explicitly discontinuing staff imputation of patients' race/ethnicity information is critical to such data's accuracy and quality.[24] Studies indicate that systematic elicitation of such information can be swift and accurate.[49] Our analyses suggest that even minimal enforcement of existing institutional policies, such as those directing front-line staff to request and record patients' race/ethnicity information consistently, could markedly improve overall data quality.[35,50]

Significance

Collectively, our results provide evidence that high-quality race/ethnicity data are attainable using existing collection systems. Furthermore, our agreement analysis and the overwhelming influence of "Unknown" values on data quality suggest that many of the "errors" inherent in race/ethnicity data collection can be directly addressed. Discrepancies and disparities in race and ethnicity data quality across the five sources we examined were striking. In any other field of science and medicine, an individual data source found to contribute such poor information would command remedial action. In real-world settings involving actual patient data, we cannot simply exclude sources of data, as we did here, that fail to meet acceptable standards of data quality.[51] As representations of actual patient status or clinical experience, poor-quality data in these systems demand intervention.

CONCLUSION

Healthcare systems should monitor and routinely enforce data quality standards, because transparent reporting of healthcare delivery and outcomes by patient race/ethnicity is critical to building trust between healthcare systems and the diverse communities that they serve.

CONTRIBUTORS

All authors meet all authorship requirements as stated in the Uniform Requirements for Manuscripts Submitted to Biomedical Journals. Design and data collection: S.C.L.; data analysis: J.G., J.T.; interpretation, writing and editing: S.C.L., J.G., J.T.

COMPETING INTERESTS

None.

FUNDING

This work was supported by a Special Interest Award for Cancer Disparities Research to Dr. Lee (ACS IRG 02-196), the University of Texas Southwestern Center for Translational Medicine (NIH UL1TR001105), and the University of Texas Southwestern Center for Patient-Centered Outcomes Research (AHRQ R24 HS022418).

ACKNOWLEDGEMENTS

We gratefully acknowledge the Office of Research Administration for Parkland Health & Hospital System. We thank Joan Cox (Parkland) and Alejandra Madrigales (University of Texas Southwestern) for their support and assistance with the cancer registry data extraction and the frank discussion of North American Association of Central Cancer Registries (NAACCR) standards. The authors thank Trisha Melhado for initial project management and data coordination.

REFERENCES

1. Thorlby R, Jorgensen S, Siegel B, Ayanian JZ. How health care organizations are using data on patients' race and ethnicity to improve quality of care. *Milbank Q.* 2011;89(2):226–255.
2. Mays VM, Ponce NA, Washington DL, Cochran SD. Classification of race and ethnicity: Implications for public health. *Annu Rev Public Health.* 2003;24(1):83–110.
3. Hahn RA. The state of federal health statistics on racial and ethnic groups. *JAMA.* 1992;267(2):268–271.
4. Manly JJ. Deconstructing race and ethnicity - Implications for measurement of health outcomes. *Medical Care.* 2006;44(11):S10–S16.
5. Ford ME, Kelly PA. Conceptualizing and categorizing race and ethnicity in health services research. *Health Serv Res.* 2005;40(5 Pt 2):1658–1675.
6. Clegg LX, Reichman ME, Hankey BF *et al.* Quality of race, Hispanic ethnicity, and immigrant status in population-based cancer registry data: implications for health disparity studies. *Cancer Causes Control.* 2007; 18(2):177–187.
7. Zaslavsky AM, Ayanian JZ, Zaborski LB. The validity of race and ethnicity in enrollment data for Medicare beneficiaries. *Health Serv Res.* 2012;47(3 Pt 2):1300–1321.
8. Wynia MK, Ivey SL, Hasnain-Wynia R. Collection of data on patients' race and ethnic group by physician practices. *N Engl J Med.* 2010;362(9):846–850.
9. Nerenz DR, Carreon R, Veselovskiy G. Race, ethnicity, and language data collection by health plans: findings from 2010 AHIPF-RWJF survey. *J Health Care Poor Underserved.* 2013;24(4):1769–1783.
10. Fiellin M, Chemerynski S, Borak J. Race, ethnicity, and the Seer database. *Med Pediatr Oncol.* 2003;41(5):413–414.
11. Blustein J. The reliability of racial classifications in hospital discharge abstract data. *Am J Public Health.* 1994;84(6):1018–1021.
12. Schrag D. Enhancing cancer registry data to promote rational health system design. *J Natl Cancer Inst.* 2008;100(6):378–379.
13. Johnson PJ, Blewett LA, Davern M. Disparities in public use data availability for race, ethnic, and immigrant groups national surveys for healthcare disparities research. *Medical Care.* 2010;48(12):1122–1127.
14. Shea CM, Malone R, Weinberger M *et al.* Assessing organizational capacity for achieving meaningful use of electronic health records. *Health Care Manage Rev.* 2014;39(2):124–133.
15. Rubin DB, Singh SR, Jacobson PD. Evaluating hospitals' provision of community benefit: an argument for an outcome-based approach to nonprofit hospital tax exemption. *Am J Public Health.* 2013;103(4):612–616.
16. Gomez SL, Le GM, West DW, Satariano WA, O'Connor L. Hospital policy and practice regarding the collection of data on race, ethnicity, and birthplace. *Am J Public Health.* 2003;93(10):1685–1688.
17. Gomez S, Glaser S. Quality of cancer registry birthplace data for Hispanics living in the United States. *Cancer Causes Control.* 2005;16(6):713–723.
18. Ramirez AG, Miller AR, Gallion K, San Miguel de MS, Chalela P, Garcia AS. Testing three different cancer genetics registry recruitment methods with Hispanic cancer patients and their family members previously registered in local cancer registries in Texas. *Community Genet.* 2008;11(4):215–223.
19. Gomez SL, Glaser SL. Misclassification of race/ethnicity in a population-based cancer registry (United States). *Cancer Causes Control.* 2006;17(6): 771–781.
20. Armenti KR, Celaya MO, Cherala S, Riddle B, Schumacher PK, Rees JR. Improving the quality of industry and occupation data at a central cancer registry. *Am J Ind Med.* 2010;53(10):995–1001.

21. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *JAMIA.* 2013;20(1):144–151.
22. American College of Surgeons Commission on Cancer. *Facility Oncology Registry Data Standards (Revised for 2015).* Chicago, IL; 2014.
23. Koering SM. The role of the cancer registrar. How cancer registrars ensure quality data for public reporting. *J AHIMA.* 2008;79(3):66–67.
24. Lee SJC. Organizational practice and social constraints: Problems of racial identity data collectionin cancer care and research. In: Laura EG, Nancy L, eds. *Mapping "race": Critical Approaches to Health Disparities Research.* New Brunswich: Rutgers University Press; 2013: 87–103.
25. Wallman KK, Evinger S, Schechter S. Measuring our nation's diversity: developing a common language for data on race/ethnicity. *Am J Public Health.* 2000;90(11):1704–1708.
26. Agency for Healthcare Research & Quality. *Defining Categorization Needs for Race and Ethnicity Data: Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement.* Rockville, MD: AHRQ; 2015.
27. Lin L, Hedayat AS, Wu W. A unified approach for assessing agreement for continuous and categorical data. *J Biopharm Stat.* 2007;17(4):629–652.
28. Vanbelle S, Mutsvari T, Declerck D, Lesaffre E. Hierarchical modeling of agreement. *Stat Med.* 2012;31(28):3667–3680.
29. Ruddat I, Scholz B, Bergmann S *et al.* Statistical tools to improve assessing agreement between several observers. *Animal.* 2014;8(4):643–649.
30. Stewart SL, Swallen KC, Glaser SL, Horn-Ross PL, West DW. Comparison of methods for classifying Hispanic ethnicity in a population-based cancer registry. *Am J Epidemiol.* 1999;149(11):1063–1071.
31. Uebersax JS. A design-independent method for measuring the reliability of psychiatric diagnosis. *J Psychiatr Res.* 1982;17(4):335–342.
32. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol.* 1990;43(6):551–558.
33. Lin SS, O'Malley CD, Lui SW. Factors associated with missing birthplace information in a population-based cancer registry. *Ethn Dis.* 2001;11(4):598–605.
34. Laws MB, Heckscher RA. Racial and ethnic identification practices in public health data systems in New England. *Public Health Rep.* 2002;117(1):50–61.
35. Jorgensen S, Thorlby R, Weinick RM, Ayanian JZ. Responses of Massachusetts hospitals to a state mandate to collect race, ethnicity and language data from patients: a qualitative study. *BMC Health Serv Res.* 2010;10:352.
36. 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):1034–1042.
37. Ulmer C, McFadden B, Nerenz DR. *Race, Ethnicity, and Language Data:: Standardization for Health Care Quality Improvement.* Washington DC: National Academies Press; 2009.
38. 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–S388.
39. Gomez SL, Lichtensztajn DY, Parikh P, Hasnain-Wynia R, Ponce N, Zingmond D. Hospital practices in the collection of patient race, ethnicity, and language data: a statewide survey, California, 2011. *J Healthcare Poor Underserved.* 2014;25(3):1548–6869.
40. HRSA Bureau of Primary Health Care. *Uniform Data System Manual.* US Health Resources & Services Administration: Washington, DC; 2013. Report No.: OMB Control # 0195-1093.
41. Haile RW, John EM, Levine AJ *et al.* A review of cancer in US Hispanic populations. *Cancer Prevent Res.* 2012;5(2):150–163.
42. Fong M, Henson DE, Devesa SS, Anderson WF. Inter- and intra-ethnic differences for female breast carcinoma incidence in the continental United States and in the state of Hawaii. *Br Cancer Res Treat.* 2006; 97(1):57–65.
43. Ai AL, Appel HB, Huang B, Lee K. Overall health and healthcare utilization among Latino American women in the United States. *J Women's Health.* 2012;21(8):878–885.
44. Ai AL, Noel L, Appel HB, Huang B, Hefley WE. Overall health and health care utilization among Latino American men in the United States. *Am J Mens Health.* 2013;7(1):6–17.
45. Banegas MP, Leng M, Graubard BI, Morales LS. The risk of developing invasive breast cancer in Hispanic women. *Cancer.* 2013;119(7): 1373–1380.
46. Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. *Health Affairs.* 2014;33(8):1314–1322.
47. Hasnain-Wynia R, Van DK, Youdelman M *et al.* Barriers to collecting patient race, ethnicity, and primary language data in physician practices: an exploratory study. *J Natl Med Assoc.* 2010;102(9):769–775.
48. Brown J, Dane FC, Durham MD. Perception of race and ethnicity. *J Soc Beh Pers.* 1998;13(2):295–306.
49. Baker DW, Cameron KA, Feinglass J *et al.* A system for rapidly and accurately collecting patients' race and ethnicity. *Am J Public Health.* 2006;96(3):532–537.
50. Hasnain-Wynia R, Baker DW. Obtaining data on patient race, ethnicity, and primary language in health care organizations: current challenges and proposed solutions. *Health Serv Res.* 2006;41(4 Pt 1):1501–1518.
51. Barnhart HX, Yow E, Crowley AL *et al.* Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. *Stat Methods Med Res.* 2014, doi: 10.1177/0962280214534651.

AUTHOR AFFILIATIONS

[1]Department of Clinical Sciences, University of Texas, Southwestern Medical Center, Dallas, TX, USA

[2]Harold C. Simmons Comprehensive Cancer Center, Dallas, TX, USA