Accuracy of Race, Ethnicity, and Language Preference in an Electronic Health Record

Elissa V. Klinger, MS¹, Sara V. Carlini, BA¹, Irina Gonzalez, TTS, MD¹, Stella St. Hubert, BA¹, Jeffrey A. Linder, MD, MPH^{1,4}, Nancy A. Rigotti, MD^{2,4}, Emily Z. Kontos, MS, ScD³, Elyse R. Park, Ph.D., MPH^{2,4}, Lucas X. Marinacci, BS¹, and Jennifer S. Haas, MD, MSPH^{1,3,4}

¹Division of General Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, USA; ²Tobacco Research and Treatment Center and General Medicine Division, Massachusetts General Hospital, Boston, MA, USA; ³Department of Social and Behavioral Sciences, Harvard School of Public Health, Boston, MA, USA; ⁴Harvard Medical School, Boston, MA, USA.

BACKGROUND: Collection of data on race, ethnicity, and language preference is required as part of the “meaningful use” of electronic health records (EHRs). These data serve as a foundation for interventions to reduce health disparities.

OBJECTIVE: Our aim was to compare the accuracy of EHR-recorded data on race, ethnicity, and language preference to that reported directly by patients.

DESIGN/SUBJECTS/MAIN MEASURES: Data collected as part of a tobacco cessation intervention for minority and low-income smokers across a network of 13 primary care clinics ($n=569$).

KEY RESULTS: Patients were more likely to self-report Hispanic ethnicity (19.6 % vs. 16.6 %, $p<0.001$) and African American race (27.0 % vs. 20.4 %, $p<0.001$) than was reported in the EHR. Conversely, patients were less likely to complete the survey in Spanish than the language preference noted in the EHR suggested (5.1 % vs. 6.3 %, $p<0.001$). Thirty percent of whites self-reported identification with at least one other racial or ethnic group, as did 37.0 % of Hispanics, and 41.0 % of African Americans. Over one-third of EHR-documented Spanish speakers elected to take the survey in English. One-fifth of individuals who took the survey in Spanish were recorded in the EHR as English-speaking.

CONCLUSION: We demonstrate important inaccuracies and the need for better processes to document race/ethnicity and language preference in EHRs.

KEY WORDS: disparities; race; ethnicity; health information technology.

J Gen Intern Med 30(6):719–23

DOI: 10.1007/s11606-014-3102-8

© Society of General Internal Medicine 2014

INTRODUCTION

Racial and ethnic disparities in both access to and quality of care are well documented in the context of U.S. health care.^{1,2} Inequalities in health care may persist on some level as a function of the failure to systematically collect accurate information about, and subsequently respond to, the needs of racial/

ethnic minority patient populations. A first step toward achieving more equitable health care is the systematic collection of accurate data on patient race, ethnicity, and language preference;³ these data can create an actionable foundation upon which to understand where and why disparities exist.⁴ Electronic health records (EHRs) provide an opportunity to collect such information systematically to enable provision of targeted interventions for particular individuals and populations. To this end, incentives for the “meaningful use” of EHRs require the collection of patient demographic data including race, ethnicity, and language preference, with the explicit intent to improve quality and safety, reduce health disparities, and improve population health.⁵ As providers and health systems transition to accountable care by managing the health of populations, the routine and accurate collection of these data may promote equity.³ In 1997, the federal Office for Management and Budget updated standards for the collection of race and ethnicity data by federal and state government agencies. In 2007, Massachusetts implemented regulations requiring all acute care hospitals, but not ambulatory practices, to collect patient self-reported race, ethnicity, and preferred language, using a standardized approach with specific categories.⁶

Despite recognition that such data can improve care, reliable data collection around race, ethnicity, and language preference is uncommon, even in settings that disproportionately treat minority and immigrant populations.⁷ Rates of misclassification and missing data in EHRs were high prior to the implementation of meaningful use.⁸ Structured patient self-reporting of these data directly into an EHR offers the potential to improve the accuracy of these data, as has been demonstrated for other domains like the completion of health maintenance items.⁹ We compared patient self-reported information about race, ethnicity, and language preference collected using an automated phone call to that recorded in the EHR.

METHODS

Overview

As part of a randomized controlled trial (RCT) of a multifaceted tobacco treatment intervention designed to identify

Received March 21, 2014

Revised October 8, 2014

Accepted October 30, 2014

Published online December 20, 2014

and recruit low-to-moderate income and minority smokers, we collected demographic data directly from patients using an interactive automated phone call. The primary purpose of the call was to proactively reach out to eligible smokers and offer participation and enrollment by briefly summarizing the aims of the study, obtaining consent, and ensuring eligibility; the final portion of this 5–10 min automated call collected standardized information about race and ethnicity. We then compared these patient self-reported data to existing, coded data in the EHR. The Partners Human Research Committee approved the study protocol and the trial was registered at ClinicalTrials.gov (NCT01156610).

Setting and Eligibility

Patients were recruited for the RCT between November 2011 and June 2013 from 13 primary care practices in Eastern Massachusetts affiliated with Brigham and Women's Hospital and Massachusetts General Hospital. The practice settings included six community health centers, two community-based practices, four hospital-based practices, and one medical home, all of which share a common EHR called the Longitudinal Medical Record, an internally developed, web-based, fully functional EHR that has been certified by the Certification Commission for Healthcare Information Technology (CCHIT). Eligible patients were at least 18 years old, had made a visit to their primary care provider within the last 60 days (to ensure that their smoking status was current), and were noted in the EHR as a current smoker in either the patient problem list or a coded smoking status field. Individuals were eligible for this study if the race field alone and not the ethnicity field was coded in the EHR as white, black, or Hispanic and they lived in a census tract with a low or moderate median household income. Additionally, individuals were required to have a primary language of English or Spanish noted in the EHR, as the intervention was available in only those languages. Patients with missing or unknown race field data or primary language data were ineligible for the study.

EHR-Derived Data on Race, Ethnicity, and Language Preference

At the time of enrollment, patient race, ethnicity, language preference, and other demographic data entered the EHR by means of a centralized, phone-based registration process for outpatient primary care patients, regardless of the settings in which they received their primary care. The system was navigated by a live operator who had completed a brief registration survey by phone. While automated prompts in the online registration system reminded staff to collect these data, variability is likely to have existed in how this information was requested, the responses elicited, and what was actually recorded. Race and ethnicity were captured in separate EHR data fields, with the "race" field category more narrowly defined (e.g., black/African American, Hispanic or Latino, white); however, possible responses for this field conflated

race and ethnicity, especially for patients of Hispanic ethnicity. The race field allowed for only one selected option, including an "other" category, although patients designated as "other" did not meet eligibility for the RCT. The "ethnicity" field, which we looked at only after we had assembled our cohort based on the race field, conflated multiple dimensions of identity, such as culture, ancestry, and country of birth. Response categories in the ethnicity field included over 100 options (e.g., Afghanistani to Zairean). Language preference was asked and recorded for spoken communication only.

Patient Self-Reported Data on Race/Ethnicity and Language Preference

Patients eligible for the trial were mailed an informational letter in their EHR-designated language with instructions on how to opt out. Patients who did not opt out within 2 weeks were contacted using an automated phone call generated with an interactive voice response system (IVRS), and received up to 15 call attempts over a 2-week period. If the phone was answered and hung up twice by anyone in the household, this was considered a "passive" opt-out. Patients who answered the phone heard a brief informed consent script and were asked to confirm their interest in participating. The IVRS call, available to patients in their choice of either English or Spanish, confirmed smoking status, and asked a series of questions about self-reported race and ethnicity that allowed participants to report multi-racial/ethnic status. The question was structured as follows: "Now I will ask you about your race and ethnicity. Please say yes or no after you hear the following categories. You may answer yes to more than one. Are you White? Black or African-American? Are you Hispanic or Latino? Asian? Other?" Language preference for spoken communication was inferred from the language in which the patient completed the automated call.

Statistical Analysis

We calculated the sensitivity, specificity, and positive predictive value (PPV) for EHR-derived information compared to patient self-report of race and ethnicity (African American vs. non-African American; Hispanic vs. non-Hispanic). We also compared language preference in the EHR and the language each respondent used to complete the survey. Statistical analyses were conducted using SAS version 9.3 (Cary, NC), with $p < 0.05$ as the criteria for statistical significance.

RESULTS

We called 8,545 individuals between November 2011 and September 2013. Of these, 706 (8 %) answered and agreed to participate in the trial. Of the 706 participants, 569 (81 %) provided self-reported information on race and ethnicity; the analysis on language preference was restricted to these patients. The median age of participants was 50 years (19–77

range); more than 70 % were female (Table 1), reflecting the demographics of the study practices. Nearly 10 % of patients had not completed high school and only 28 % indicated that they had completed college. Approximately 15 % of participants were foreign-born.

Patients were more likely to self-report Hispanic ethnicity (19.6 % vs. 16.6 %, $p<0.001$), and African American race (27.0 % vs. 20.4 %, $p<0.001$) than was reflected in the EHR. Conversely, patients were less likely to complete the survey in Spanish than the language preference noted in the EHR would have suggested (5.1 % vs. 6.3 %, $p<0.001$). Thirty percent of whites also self-reported identification with at least one other racial or ethnic group, as did 37.0 % of Hispanics, and 41.0 % of African Americans.

Overall, strong and statistically significant agreement ($p<0.001$) was observed between EHR-documented and self-reported race. While the sensitivity of EHR-recorded race compared to patient self-reported race for African Americans was modest (70.9 %), the specificity and PPV were high (98.8 % and 95.5 % respectively) (Table 2). A similar pattern was seen for Hispanics (83.8 %, 99.8 %, and 98.9 % respectively). The sensitivity was high for whites (93.8 %), as was the specificity and PPV (97.0 % and 98.3 % respectively). For Spanish language preference, the sensitivity was modest (79.3 %), the specificity high (97.6 %), and the PPV more modest (63.9 %). Only 1 % of EHR-documented English speakers chose to take the phone survey in Spanish, whereas 36 % of EHR-documented Spanish speakers elected to take the survey in English. One-fifth of individuals who took the survey in Spanish were recorded in the EHR as English-speaking.

Table 1. Demographics ($n=569$)

	Source of data:		<i>p</i> value
	Electronic health record	Self-report	
Age, years (median, range)		50 (19–77)	
Sex			
Female		400 (70.3 %)	
Ethnicity			
Hispanic ^a	94 (16.6 %)	111 (19.6 %) ^b	< 0.0001
Race			
White	355 (62.4 %)	372 (65.4 %) ^c	< 0.0001
Black ^a	112 (20.4 %)	151 (27.0 %) ^d	< 0.0001
Language preference			
Spanish	36 (6.3 %)	29 (5.1 %)	< 0.0001
Education ^a			
Did not complete high school		52 (9.2 %)	
High school graduate		148 (26.3 %)	
Some college		208 (36.9 %)	
College graduate		155 (27.5 %)	
Born in the US		476 (85.0 %)	

^a Missing IVR: Hispanic 3; Black 9; Education 6; Born in the US 9

^b Twenty-eight Hispanics also reported white; 13 also reported black

^c 111 whites reported two or more race categories

^d 62 blacks reported two or more race categories

Table 2. Accuracy of Electronic Health Record Documented Race, Ethnicity and Language Preference Compared with Self-Report

Electronic health record documented:	Compared with self-report		
	Sensitivity	Specificity	Positive predictive value
Race/ ethnicity:			
Black	70.9	98.8	95.5
Hispanic	83.8	99.8	98.9
White	93.8	97.0	98.3
Language preference:			
Spanish	79.3	97.6	63.9

DISCUSSION

Although we demonstrate strong agreement between EHR-documented and self-reported race, ethnicity, and language preference, we also demonstrate important inaccuracies and the need for better processes to document these data in EHRs. Our data suggest that, even in a state with regulations promoting the collection of self-reported data about race and ethnicity in acute care settings, EHR data may “undercount” individuals who identify themselves as African American or Hispanic in ambulatory settings. In addition, one-third of individuals coded in the EHR as Spanish-speaking opted to participate in this study in English. These results demonstrate the importance of identifying the individual language preferences of multilingual patients, who may have differing preferences based on the specific content and method of contact.

We should note that both the EHR-derived demographic data and the information collected as part of our RCT are patient self-reported, but we suspect that the quality of demographic data in the EHR is limited both by human factors in the capture of these data and by the way that the separate race and ethnicity data fields are structured. We also have limited information on how historic demographic data were elicited and entered into the EHR prior to current centralized processes. While our institution utilizes a centralized live-operator phone-based registration process for new primary care patients, how operators are trained to prompt for collection of these data and record the responses is likely to vary. To add to this complexity, definitions for what constitutes race, ethnicity, and other self-identifiers are fluid, debatable, and problematic to individuals and to institutions. The fact that the “race” field at our institution includes both race and ethnicity in a single field highlights this confusion at a systems level. Our data collection tool via automated phone calls did prompt for race and ethnicity in a single question, but also allowed for patients to self-identify with one or all possible categories, thereby allowing multiple race and ethnicities to be captured in a single field.

Our results suggest similar rates of misclassification of race and ethnicity in the EHR when compared to work completed

before the implementation of meaningful use criteria for EHRs; that said, missing data were reported less often.⁸ Our results are consistent with findings for the concordance between self-reported race and ethnicity and that abstracted by medical record reviewers (i.e., before documentation in coded fields),¹⁰ and that recorded in health care claims.¹¹ As EHR data are increasingly being used to manage the health of populations, ensuring the accurate recording of these characteristics will be crucial to the success of interventions designed to reduce health disparities. Health disparities may not be alleviated solely by broad quality improvement and health care reform initiatives, but may require specific, targeted interventions.^{12,13} The systematic collection of accurate data on patient race, ethnicity, and language preference is necessary to monitor interventions designed to promote equity. The widespread deployment of EHRs offers the opportunity to assess and improve the health of populations of patients, but also may be used to augment traditional public health tools like the Behavioral Risk Factor Surveillance Survey, by helping to define areas with unaddressed health needs.¹⁴ Our data suggest that requiring coded documentation alone may not be sufficient to ensure high quality data for these constructs.

These findings highlight the complexities of accurately capturing and recording race, ethnicity, and language preference in clinical settings. We found that a substantial proportion of this particular patient population self-identifies with more than one racial or ethnic group, a phenomenon not adequately reflected in EHR data. Explanations for this may include the limited structure of a registration process not adequately designed to elicit or record multi-racial and/or ethnic status. Second, reporting of language preference may be complex, as patients may not report that they speak Spanish because they are concerned about stigma or bias.¹⁵ Multi-lingual individuals may also prefer different languages for different tasks, which cannot be derived from a single question about language preference. The factors that mediate why multi-lingual patients may prefer to use one language for some activities and not others should be an area for future qualitative research.

Our study has several limitations. Participants were recruited into an RCT specifically designed to target low-income and minority patient populations of smokers. Our data come from a single primary care network, but include data from patients cared for at several primary care sites. Because our study design was directed at individuals coded as African American, Hispanic or white in the EHR, we are not able to address the accuracy of coding other racial and ethnic minority populations. We did not collect information about ethnic subgroups. Additional granularity may be important for promoting health equity,⁶ yet an exhaustive list of options may be overwhelming for both patients and providers.¹⁶ Our automated call was offered in a choice of English or Spanish, the two most common languages for this patient population. As language preference is a cornerstone of informed decision-making,

future attention should be paid to the complexity of this construct.

The capacity for patients to self-report their race, ethnicity, and language preference in their EHR could be improved and standardized through the use of patient portals that accommodate a multi-lingual population. Such efforts could include facilitating patient-reported data at point of care, either in a waiting room with the support of trained personnel, or prior to a visit. Collection of these data should address any patient concerns about the reasons for their use, while simultaneously offering safeguards for protection of this information.¹⁵ Improvement in the capture, recording, and storage of race and ethnicity as well as language preference is necessary in the healthcare setting, and could serve as a means to address health equity. To this end, efforts should include the systematic collection of such information directly from patients, to ensure that these data accurately reflect the complexity of these identity constructs.

Acknowledgements: Funding Sources: This work was conducted with support from the Lung Cancer Disparities Center at the Harvard School of Public Health (National Cancer Institute Award # P50 CA148596), Harvard Catalyst, the National Center for Research Resources, and the National Center for Advancing Translational Sciences (National Institutes of Health Award 1UL1 TR001102 and financial contributions from Harvard University and its affiliated academic health care centers). The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard University and its affiliated academic health care centers or the NIH.

Prior Presentations: A preliminary version of this work was presented at the national meeting of the Society for General Internal Medicine in April 2013 in Denver, Colorado.

Conflict of Interest: Nancy A. Rigotti, MD serves as an unpaid consultant to Pfizer and Allere and currently receives royalties from Up-to-Date. The authors declare that there are no other potential conflicts of interest to report.

Corresponding Author: Jennifer S. Haas, MD, MSPH; Division of General Medicine and Primary Care Brigham and Women's Hospital, 1620 Tremont Street, Boston, MA 02120, USA (e-mail: jhaas@partners.org).

REFERENCES

1. Agency for Healthcare Research and Quality. National Healthcare Disparities Report. Rockville, MD: U.S. Department of Health and Human Services; 2011.
2. Smedley BD, Stith AY, Nelson AR. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: Institute of Medicine; 2003.
3. 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.
4. Ulmer C, McFadden B, Nerenz D. Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. Washington, DC: National Academies Press; 2009.
5. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med*. 2010;363:501–504.
6. Weinick RM, Caglia JM, Friedman E, Flaherty K. Measuring racial and ethnic health care disparities in Massachusetts. *Health Aff*. 2007;26(5):1293–1302.

7. **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, part1):1501–1518.
8. **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.
9. **Staroselsky M, Volk LA, Tsurikova R, et al.** Improving electronic health record (EHR) accuracy and increasing compliance with health maintenance clinical guidelines through patient access and input. *Int J Med Inform.* 2006;75(10–11):693–700.
10. **West CN, Geiger AM, Greene SM, et al.** Race and ethnicity: comparing medical records to self-reports. *J Natl Cancer Inst Monogr.* 2005;35:72–74.
11. **McAlpine DD, Beebe TJ, Davern M, Call KT.** Agreement between self-reported and administrative race and ethnicity data among Medicaid enrollees in Minnesota. *Health Serv Res.* 2007;42(6 Pt 2):2373–2388.
12. **Zhu J, Brawarsky P, Lipsitz S, Huskamp H, Haas JS.** Massachusetts health reform and disparities in coverage, access and health status. *J Gen Intern Med.* 2010;25(12):1356–1362.
13. **Clark CR, Soukup J, Govindarajulu U, Riden HE, Tovar DA, Johnson PA.** Lack of access due to costs remains a problem for some in Massachusetts despite the state's health reforms. *Health Aff.* 2011;30(2):247–255.
14. **Linder JA, Rigotti NA, Brawarsky P, et al.** Use of practice-based research network data to measure neighborhood smoking prevalence. *Prev Chron Dis.* 2013;10:E84.
15. **Hasnain-Wynia R, Taylor-Clark K, Anise A.** Collecting race, ethnicity, and language data to identify and reduce health disparities: Perceptions of health plan enrollees. *Med Care Res Rev.* 2011;68(3):367–381.
16. **Lurie N, Fremont A.** Looking forward: cross-cutting issues in the collection and use of racial/ethnic data. *Health Serv Res.* 2006;41(4, Part 1):1519–1533.