Concordance Rates in Racial and Ethnic Identification: Insights Based on Linked Electronic Health Records and the American Community Survey

Rocio Rosa-Lebron[1], Aubrey Limburg[2], Timothy S. Carey[3], Victoria Udalova[2], Barbara Entwisle[1,4]

Research Question

To what extent does race and ethnicity as recorded in electronic health records (EHRs) correspond with race and ethnicity as recorded in the American Community Survey (ACS)?

Background

EHR-derived data (henceforth EHRs) are a potentially valuable source for the study of clinical correlates, social determinants, health disparities, and population health. However, despite many strengths, EHRs:

- have limited information on the social determinants of health
- are often missing race and ethnicity information
- have considerable variation in how race and ethnicity data are collected
- represent a non-random sample of the population.

ACS is strong in coverage of social determinants, race and ethnicity, and sample design, but weak in measurement of health. Integrating the two leverages complementary strengths.

Data and Methods

EHRs from a large public integrated health system in North Carolina (2016-2019) linked to American Community Survey (ACS, 2001-2017) microdata.

- Data use agreement; IRB review; health system approval
- Secure transfer of EHR-derived data to U.S. Census Bureau
- Data ingested and anonymized identifiers (i.e., PIKs) assigned
- PIKs used to match EHR-derived data to ACS microdata
- 14.57% (n=29,000) of selected patients matched to ACS

EHR characteristics

- Measures:
  - Patients (or proxy) provided race and ethnicity
  - Clerks instructed to accept whatever answer they were given
- Case selection:
  - Disproportionate stratified random sample of ~200,000 patients
  - 25-74 years of age
  - At least 2 visits between 2016-2019

ACS characteristics

- Source:
  - Annual cross-sectional probability survey of about ~2 million housing units
  - Missing responses can be allocated or imputed
- Measures:
  - Reference persons report on their own race and ethnicity as well as others living at that address (mix of respondent and informant data)

Table 1. Race and Ethnicity Classification in EHR and ACS Data

EHR Data	ACS Data
Ethnicity	
Hispanic or Latino	Hispanic, Latino, or Spanish origin [Mexican, Mexican American, Chicano; Puerto Rican; Cuban]
Not Hispanic or Latino	Not Hispanic or Latino, or Spanish origin
Missing/Patient Refused	Missing (N/A)
Unknown	
Race	
White or Caucasian	White
Black or African American	Black or African American
American Indian or Alaska Native (AIAN)	American Indian or Alaska Native [print tribe] (AIAN)
Asian	Asian [Asian Indian; Japanese; Chinese; Korean; Filipino; Vietnamese; Other Asian (print)]
Native Hawaiian or Other Pacific Islander (NHPI)	Native Hawaiian or Pacific Islander [Native Hawaiian; Guamanian or Chamorro; Samoan; Other Pacific Islander (print)] (NHPI)
Other Race	Some Other Race
Missing / Patient Refused	Missing (this is not present in ACS data due to allocation methods)

Note: The population of individuals identifying as Native Hawaiian or Other Pacific Islander in North Carolina is very small (<0.5%), they were grouped with the Other race group for subsequent analyses.

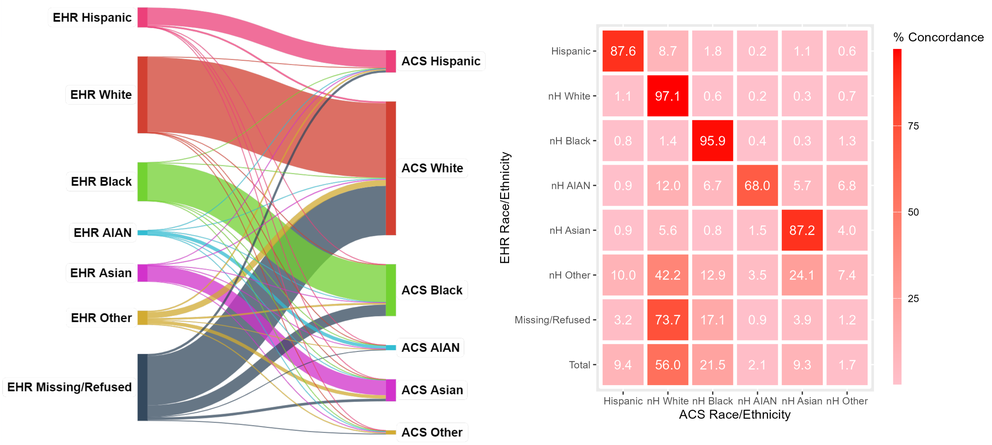
Concordance: Race Alone vs Race/Ethnicity Combined

- EHR → ACS: concordance rates similar between race alone and race/ethnicity (R/E) combined
- ACS → EHR: concordance increases with R/E combined compared to race alone for non-Hispanic (nH) White, nH Black, and nH AIAN patients

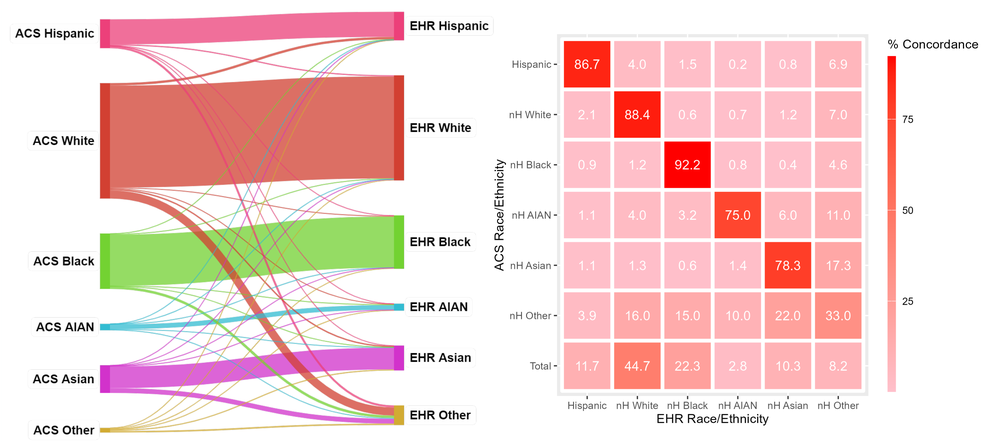
	EHR→ACS		ACS→EHR	
	Race alone	R/E Combined	Race alone	R/E Combined
White	98.0%	97.1%	78.4%	88.4%
Black	96.0%	95.9%	90.8%	92.2%
AIAN	68.0%	68.0%	70.0%	75.0%
Asian	87.3%	87.2%	78.2%	78.3%

- For subsequent analyses, we use race/ethnicity combined

Race/Ethnicity Concordance (EHR → ACS)

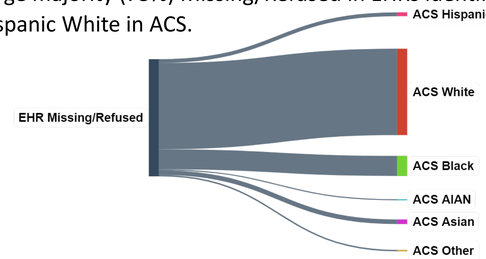


Race/Ethnicity Concordance (ACS → EHR)



Findings

- Concordance highest for non-Hispanic Black patients, lowest for non-Hispanic AIAN patients, with non-Hispanic White, Hispanic, and non-Hispanic Asian patients ranking in between, respectively.
- Except for AIAN patients, concordance is higher for EHR→ACS than for ACS→EHR.
- Large majority (73%) Missing/Refused in EHRs identified as non-Hispanic White in ACS.



Discussion

- US racial system based primarily on black-white binary. Less discordance because there is less ambiguity for Black and White patients than for other racial/ethnic groups.
- Combining race and ethnicity and giving primacy to ethnic identity helps address ambiguity of racial categories for Hispanics.
- For Asian patients, lower concordance could be a byproduct of how the race question is asked in ACS (e.g., more detailed race options).
- For AIAN patients, low concordance may reflect differences in ancestry claims and in federal and state recognition status of North Carolina's (NC) largest tribe, the Lumbees.
- ACS data can help to identify race and ethnicity for patients missing this information in EHRs.

Limitations

- Concordance analysis restricted to racial/ethnic categories as defined in EHRs (e.g., no category for mixed race).
- Analysis based on matched data from a single health system (which, while large, is one of several in NC) and a single state (with unique racial and ethnic history).
- Matched data differentially selective of racial and ethnic groups(PIK rates lower for Hispanic patients and "other" race); match rates lower for Black and Hispanic patients.

Significance

With the linked EHR-ACS data:

a) ACS data can be used to enhance data missing from the EHRs.
b) Use of EHRs can be pushed "upstream," to capture social determinants and better understand the more distal conditions that foster health.
c) EHRs can be studied in relation to a representative source population. Factors related to who appears in EHRs can be understood, modeled, and incorporated in studies of population health and health disparities.

Affiliations

1. Department of Sociology, University of North Carolina-Chapel Hill
2. Enhancing Health Data (EHealth) Program, U.S. Census Bureau
3. Department of Medicine, School of Medicine, University of North Carolina-Chapel Hill
4. Carolina Population Center, University of North Carolina-Chapel Hill

Acknowledgements: We are grateful for support from the Carolina Population Center (NICHD P2C HD050924), the UNC Translational and Clinical Sciences (TraCS) Institute (CTSA UL1TR002489), and the Enhancing Health Data (EHealth) program at the U.S. Census Bureau.

This presentation is to inform interested parties of ongoing research and to encourage discussion. Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product (Data Management System (DMS) number: P-7519212, Disclosure Review Board (DRB) approval number: CBDRB-FY23-POP001-0160; CBDRB-FY21-POP001-0087).