

Race and Ethnicity Covariates

Race and ethnicity are a good illustration of the trade-offs between clinical and research data

- Race and ethnicity data in prospectively collected research datasets are usually reliable and valid
 - Question: What kind of validity is this?
- BUT, the racial and ethnic distributions in research datasets often aren't representative of broader populations of interest
 - In other words, findings from these prospectively collected data sets may have limited _____ with respect to race and ethnicity?

Let's take a look at the race data from SHHS

- Group 1: Look at published literature or census data to get some basic info on the population-level proportions of black, white, and “other” (since these are the categories in SHHS)
- Group 2: Using the data explorer, find the proportions of black, white, and other in the SHHS dataset. How did you find this information?
- Group 3: How often are race data missing in the SHHS dataset, and how did you find this info? If you have time, are race data missing completely at random, or is there a relationship between missingness of race and outcome (any_cvd) or one of the other primary covariates?

As a class, compare these distributions. Think about the following:

- Are the distributions relatively similar, or different in a meaningful way?
- If they are different, what could be the possible repercussions in understanding results from SHHS?
- What could be possible reasons for any differences you observe?

What about race and ethnicity in EHR data?

- Likely to be more representative of the actual population, with some limitations
 - Question: How might you expect the demographics of a hospital-based sample to differ from the broader population with respect to race and ethnicity?
- But what about the quality of the race and ethnicity data?

Quality of EHR race and ethnicity data?

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

70.9% of black patients are correctly identified as black in the EHR. Therefore, **29.1%** of black patients are *not* identified as black.

79.3% of patients who prefer Spanish are identified as preferring Spanish in the EHR. Therefore, **20.7%** of patients who prefer Spanish do not have this preference in the EHR.

If a patient's preferred language is listed as Spanish in the EHR, there's a **63.9%** chance that they actually prefer Spanish. Therefore, if a patient is identified as preferring Spanish, there's a **36.1%** chance they prefer a different language.

Discussion questions about quality of EHR race and ethnicity data

- Based on your knowledge of how race and ethnicity are recorded in the EHR, what possible reasons could you think of for this disagreement?
- Would you expect to see similar rates of agreement and disagreement across different institutions?
- How much do you think this matters in reuse scenarios? (Also worth considering impact at the point of care)
- Missing and incorrect race and ethnicity data could potentially impact internal validity. What does this mean for external validity?