Race and Ethnicity in Oracle EHR Real-World Data: A Contrast to Self-Reported Data in the United States

SA77

Rascon Velasco V, Berliner E, Jaffe DH, Beckwith SM, Furegato M, Brignoli L

Oracle Life Sciences, Real World Evidence

Introduction

Race and ethnicity are strongly related to health status, healthcare access, and quality of care; thus, the accuracy of their reporting is important to better understand health disparities in research. The degree of validity and reliability usually depends on the method used for data collection, and it is known that self-reporting is considered the most reliable one[1]. Alternatively, Electronic Health Records (EHRs) is one of the usual data sources where population-level demographic data can be retrieved, yet race and ethnicity tend to be inadequately reported.[2,3,4] Therefore, to better understand the quality of race and ethnicity reported in Oracle EHR Real-World Data, we used self-reported data from a large population-based US survey to identify the level of concordance of these variables in our EHR.

Objective

To contrast race and ethnicity data reported in Oracle EHR Real-World Data (OERWD) with a US self-reported survey.

Methods

Data Sources

Data from OERWD (Data extraction – June 2023, n=105,832,841) was linked to data from the US National Health and Wellness Survey (NHWS) (2015-2022, n=292,391).
OERWD is a national, de-identified, person-centric dataset that enables organizations to leverage robust clinical data from contributing healthcare organizations. OERWD combines inpatient and outpatient clinical encounters across a variety of health systems for over 100 million patients total throughout the nation.

The US NHWS is a cross-sectional online survey administered every year to adults, aged 18 years old and more, members of a consumer panel. Each year, the US NHWS survey captured patient reported data on approximately 75,000 individuals, that are representative of the US population in terms of gender, age and race/ethnicity. Data collected from 2015 to 2022 have been pulled together to a total of 292,391 individuals who participated in NHWS.

Linkage

Tokens were created for both datasets, using Datavant HIPAA-certified de-identified linking software. Tokens were generated from Personally Identifiable Information (PII) : last name, first name, gender and date of birth. The software uses a proprietary probabilistic matching algorithm to find the matches between the two datasets.
Several steps occurred when performing the linkage:

1. **Deduplication:** an initial linkage was performed, and we identified non-unique identifiers that matched more than one individual in the other dataset. Those duplicates (n=15,558) were removed from the analysis.
2. **Inconsistencies in the NHWS dataset:** individuals who participated in several waves of NHWS and had discrepancies in terms of race and ethnicity answers or year of birth were excluded (n=346 and n=65 individuals respectively)
3. **Inconsistencies in the OERWD dataset**: individuals without a valid date of last encounter were excluded (n=1,053)
4. **Inconsistencies between both datasets**: Individuals with non-matching year of birth in both datasets were excluded (n=752)

Once all these steps were completed, we identified 29,992 linked individuals.

Data of Interest

Race was categorized as American Indian/Alaskan Native, Asian, Black/African American, Native Hawaiian/Other Pacific Islander, White, Other, and Unknown. A Mixed category was added if races were selected in NHWS.

Ethnicity was reported as Non-Hispanic, Hispanic/Latino, and Unknown.
Descriptive statistics were derived.
In NWHS, providing race and ethnicity data were mandatory.

Results

We identified 29,992 individuals, 64.7% were women with an overall mean age at last encounter of 46.2 years (SD=19.6) ranging from 0 to 89 years old.
Details of race and ethnicity in both datasets are presented in Figure 1 and 2.

Figure 1. Repartition of race in NHWS and OERWD

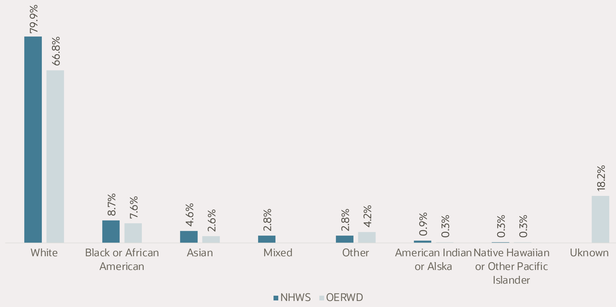
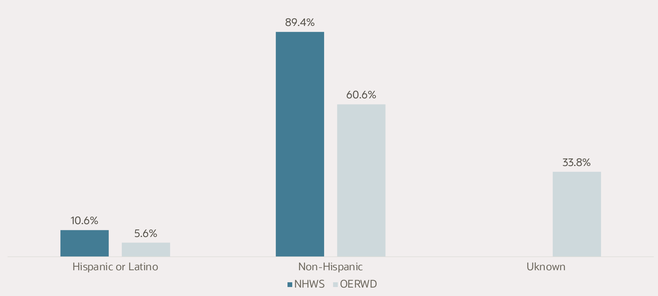


Figure 2. Repartition of ethnicity in NHWS and OERWD



For race, 72.3% (n=21,683) of patients had concordant responses in both datasets, 9.5% (n=2,852) were discordant, and 18.2% (n=5,457) were unknown in OERWD (Figure 3). The highest proportions of matched responses were observed among patients identifying as White (78.6%), Black/African American (74.0%) and Asian (51.1%) (Table 1). For all races, except Asian and White, when answers were discordant, most individuals were considered White in OERWD.
Regarding ethnicity, 58.0% (n=17,396) of patients had concordant answers, 8.2% (n=2,470) were discordant, and 33.8% (n=10,126) were unknown in OERWD (Figure 3).

Figure 3. Overall concordance of race and ethnicity data reported in OERWD against self-reported data from NHWS

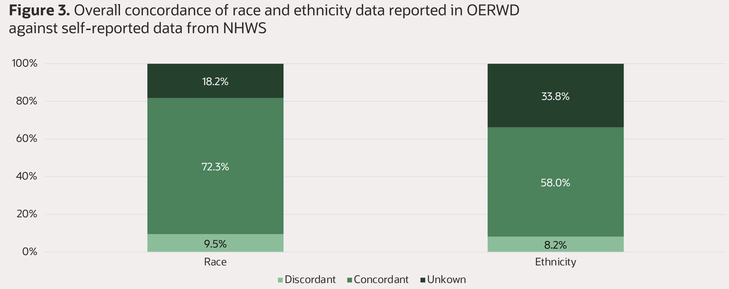


Table 1. Matching race registered in OERWD against self-reported answers in NHWS

OERWD	NHWS						
	American Indian or Alaskan Native	Asian	Black or African American	Mixed	Native Hawaiian or Other Pacific Islander	White	Other
American Indian or Alaskan Native	18.5%	0.8%	0.2%	1.2%	1.1%	0.1%	0.6%
Asian	0.4%	51.1%	0.1%	4.7%	6.4%	0.1%	1.0%
Black or African American	2.7%	0.6%	74.0%	16.7%	2.1%	0.6%	4.5%
Native Hawaiian or Other Pacific Islander	0.4%	1.7%	0.0%	1.9%	20.2%	0.1%	0.4%
White	51.4%	8.6%	7.7%	43.9%	34.0%	78.6%	43.5%
Other	7.7%	11.5%	4.0%	10.4%	12.8%	3.0%	18.2%
Unknown	18.9%	25.7%	14.0%	21.3%	23.4%	17.6%	31.9%
Total	100%	100%	100%	100%	100%	100%	100%

We investigated differences in proportion of matching by gender and age. For gender, we did not observe a big difference in the proportion of concordance across gender groups for both race (73.0% vs 71.0%, females and males respectively) and ethnicity (58.5% vs 57.1%, females and males respectively), as well as in proportion of discordance of race (9.5% vs 9.6%, females and males respectively) and ethnicity (8.6% vs 7.6%, females and males respectively). However, concordant answers increased with age at last encounter from 46.4% (0 to 14 years old) to 86.0% (75 years old and more) for race (Figure 5) and from 10.3% to 81.5% for ethnicity (Figure 6).

Figure 5. Concordance of race by age group

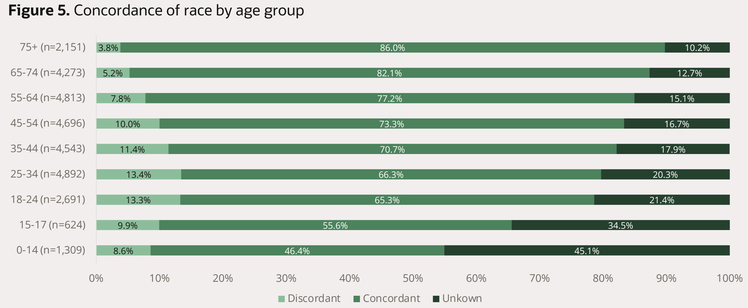
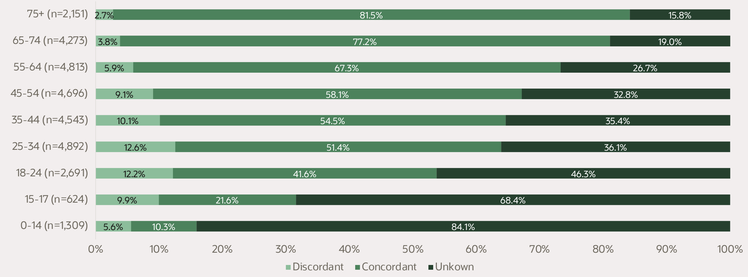


Figure 6. Concordance of ethnicity by age group



Conclusions

Our study showed adequate concordance between EHR-recorded race and ethnicity and self-reported data indicating reliability of these variables when reported in EHR. Moreover, we observed an increase in concordance of race and ethnicity with age whereas similar proportions of concordance was observed across gender groups.
Further research is needed to better understand reasons for underreporting.

References

1. Soucie J, Buckley OB, Albanese K, Harrington R, Hudson Scholle S. Current Health Plan Approaches to Race and Ethnicity Data Collection and Recommendations for Future Improvements. NCQA. 2023. https://www.ncqa.org/wp-content/uploads/2023/03/Current-Health-Plan-Approaches-to-Race-and-Ethnicity-Data-Collection-and-Recommendations-for-Future-Improvements_Final.pdf
2. Yemane L, Mateo CM, Desai AN. Race and Ethnicity Data in Electronic Health Records—Striving for Clarity. JAMA Netw Open. 2024;7(3):e240522. doi:10.1001/jamanetworkopen.2024.0522
Salhi RA, Macy ML, Samuels-Kalow ME, Hogikyan M, Kocher KE. Frequency of Discordant Documentation of Patient Race and Ethnicity. JAMA Netw Open. 2024;7(3):e240549. doi:10.1001/jamanetworkopen.2024.0549
4. Egede LE. Race, ethnicity, culture, and disparities in health care. J Gen Intern Med. 2006 Jun;21(6):667-9. doi: 10.1111/j.1525-1497.2006.0512.x. PMID: 16808759; PMCID: PMC1924616.