Improving Health Outcomes: The Implementation of Race and Ethnicity Composite Values from Data Submitted to the Maine Health Data Organization

Human Services Research Institute HSRI

Authors
Leanne Candura, MPH, Vice President, Population Health, HSRI; Karynlee Harrington, Executive Director of the Maine Health Data Organization; and Dani Saunders, MPH, HSRI

MHDO Maine Health Data Organization
Information | Insight | Improvement

Objective

Our analysis provides insights into how a data assessment on race and ethnicity was used to develop a flexible de-identified person-level composite that provides data users with high-quality demographic information, using data from two comprehensive sources:

1. **Maine Health Data Organization's (MHDO) All-Payer Claims Data (APCD)** – includes records on Maine's insured population (encompassing 90% of Maine's overall population)
2. **MHDO Hospital Encounter Data** – represents approximately 77% of Maine's population

By creating a de-identified person-level composite that accurately captures how individuals self-identify, data users can perform health care-related analysis that identify strengths and areas of improvement for racially diverse populations to assist with efforts addressing health inequities.

Background

The MHDO APCD includes medical, pharmacy, and dental claims data from commercial and public payors. Beginning with 2021 data submissions, race and ethnicity data was a required element for eligibility records, allowing the inclusion of up to three race categories, three ethnicity categories, and a Hispanic indicator.

MHDO's Hospital Inpatient and Outpatient Encounter Data includes all inpatient and outpatient services of the hospital and services provided by hospital-owned specialty groups and primary care practices. Race and ethnicity information has been largely reported to MHDO by Maine hospitals for over a decade and allows the inclusion of one race value and one ethnicity value, with additional options for race refused, race unknown, and ethnicity unknown.

In 1977, the Office of Management and Budget (OMB) established a minimum standard for the collection of race and ethnicity information for federal surveys[1]. These standards include:

- Self-reported
- Multiple selections
- 5 racial groups (American Indian or Alaskan Native, Asian, Black, Native Hawaiian or Pacific Islander, and White)
- 2 ethnic groups (Hispanic and non-Hispanic) The goal for the development of the person-level composite was to include these criteria.

Populations Studied

Data from the MHDO APCD was based on month-level records from medical, dental, and pharmacy data for 2021 and 2022. Data from the MHDO Hospital Encounter data was based on inpatient and outpatient hospital visits with a discharge date between 2018 and 2022.

Study Design

In the spring of 2022, using a deidentified index key (or PersonID) at the individual level, data in the MHDO APCD was merged to MHDO Hospital Encounter data. Data availability and completeness was assessed and consistency for deidentified individuals was examined within and across data sources. Consistency was defined as a deidentified PersonID having the same race(s) and/or ethnicity 100% of the time after 'unknown' or 'refused to answer' options are eliminated. Since the Hospital Encounter submission layout does not currently have options to include multiple races, the calculation was modified so at least one of the race options matched 100% of the time among individuals with multiple races included in the APCD. This assessment was rerun to include additional data years to determine if there were additional findings that should be taken into consideration for next steps.

The next phase was to develop the logic for the de-identified person-level composite based on the findings from the assessment.
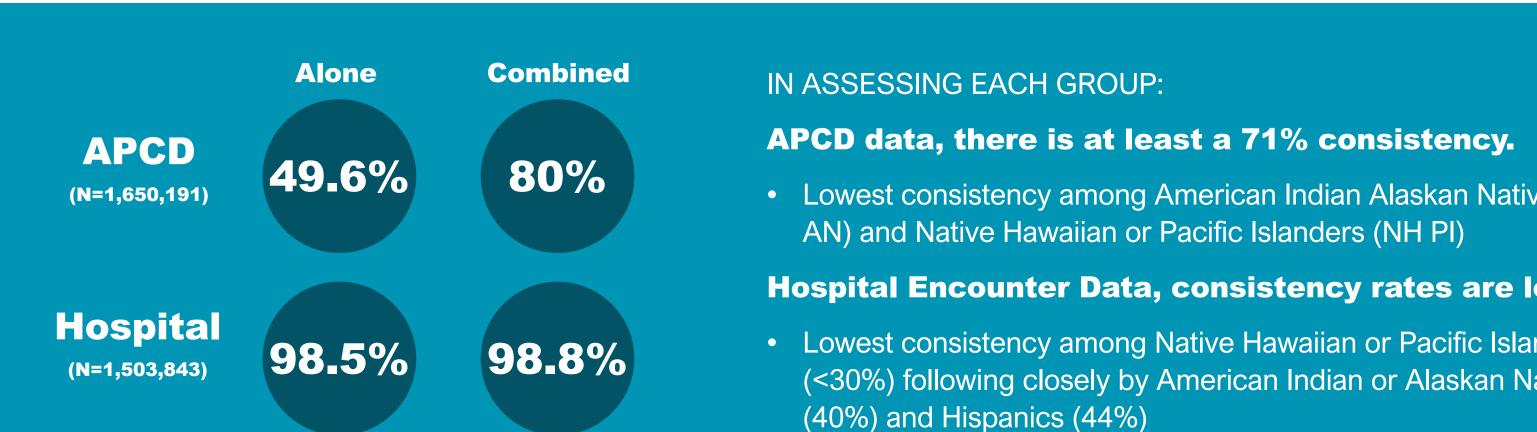
Limitations & Implications

Data collection practices from hospitals and payors are currently unknown, making it difficult to determine if gaps in the availability of racial categories is related to data reporting or data collection standards from commercial and public payors.
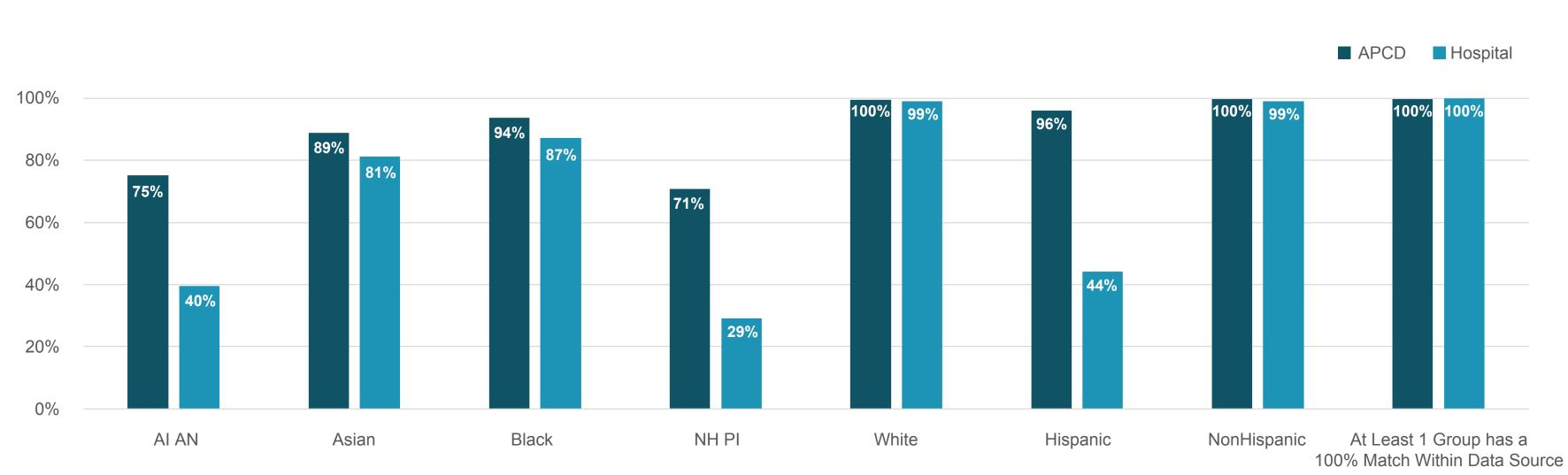
When assessing racial and ethnic disparities, understanding the quality of the data being used to examine demographic characteristics is an essential first step. Systems should collect and offer the ability to submit data based on the OMB or Centers for Disease Control (CDC) standards with options to include multiple races. This assessment informed the logic decisions made in the development of a de-identified person-level directory by providing additional insights into where there may be limitations or underreporting of minority populations.

Principal Findings

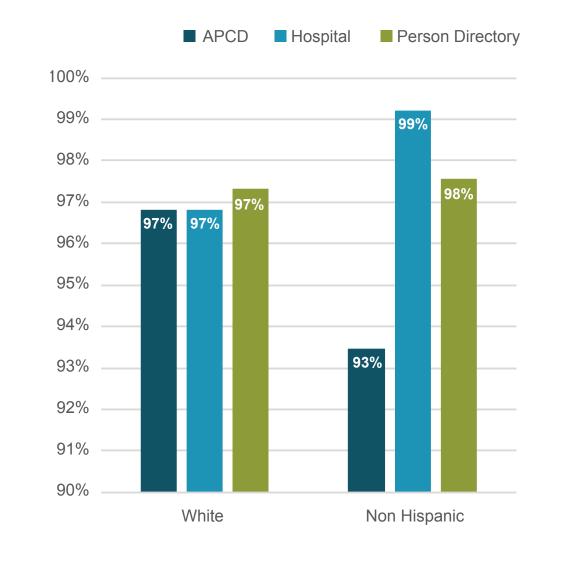### Percentage of Individuals with Race/Ethnicity Data by System

	Alone	Combined
APCD (N=1,650,191)	49.6%	80%
Hospital (N=1,503,843)	98.5%	98.8%

IN ASSESSING EACH GROUP:

APCD data, there is at least a 71% consistency.
- Lowest consistency among American Indian Alaskan Natives (AI AN) and Native Hawaiian or Pacific Islanders (NH PI)

Hospital Encounter Data, consistency rates are lower.
- Lowest consistency among Native Hawaiian or Pacific Islanders (<30%) following closely by American Indian or Alaskan Natives (40%) and Hispanics (44%)

Consistency of Race/Ethnicity Data by Data Source

Legend: APCD, Hospital

	AI AN	Asian	Black	NH PI	White	Hispanic	NonHispanic	At Least 1 Group has a 100% Match Within Data Source
APCD	75%	89%	94%	71%	100%	96%	100%	100%
Hospital	40%	81%	87%	29%	99%	44%	99%	100%
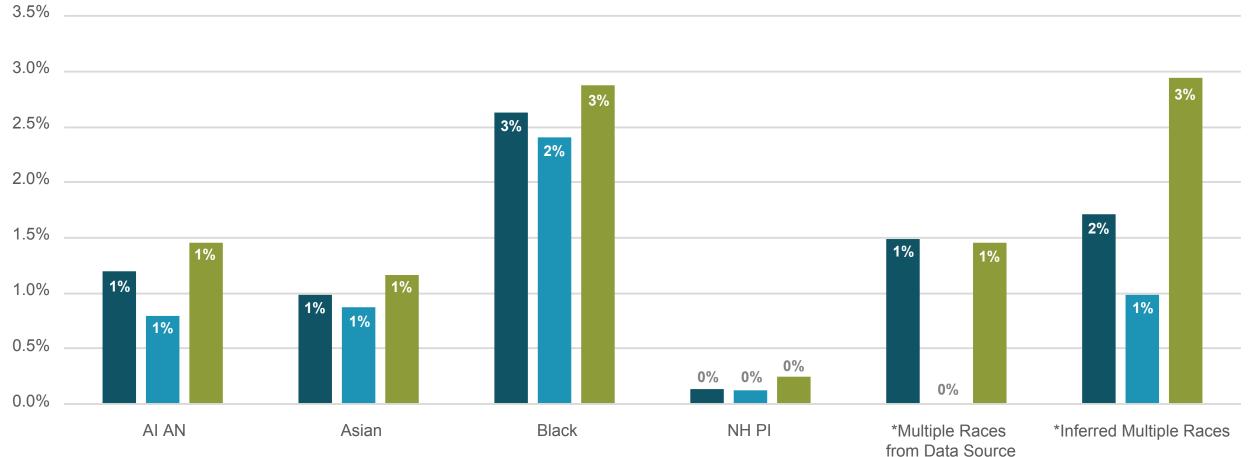
Data Set Comparisons

Data Set	Pros	Cons
APCD	• Ability to capture multi-racial identities. • More consistently reported information on minority populations.	• Only ~50% of individuals have viable race/ethnicity data. • Data is only available from 2021 onward.
Hospital Encounter	• Almost 100% completely populated. • More robust data set for race/ethnicity information.	• Unable to capture multi-racial identities. • Less consistently reported information on minority populations which may be related to the inability to submit multiracial identities.

Demographic Breakdown by Data Source

Legend: APCD, Hospital, Person Directory

	White	Non Hispanic
APCD	97%	93%
Hospital	97%	99%
Person Directory	97%	98%

	AI AN	Asian	Black	NH PI	*Multiple Races from Data Source	*Inferred Multiple Races
APCD	1%	1%	2%	0%	1%	2%
Hospital	1%	1%	2%	0%	1%	1%
Person Directory	1%	1%	3%	0%	0%	3%

Race/Ethnicity Reporting Methods

Method	Pros	Cons
Rarest Race: Prioritizing less common race ethnicities first to assign one distinct race/ethnicity value. Order of priority is Native Hawaiian or Pacific Islander, Black, American Indian or Alaskan Native, Asian, Hispanic, and White.	• Minority populations would be less likely to be under reported since this method prioritizes lower frequency races. • Combines race and ethnicity into one easy to use category.	• Disregards individuals who self-identify with multiple races. • Data collection or entry errors may be possible and would persist as information is updated.
Most Frequent Race/Ethnicity: For each individual, a race and ethnicity are assigned based on what value shows up the most frequently.	• Accuracy of racial/ethnic identity could be assumed for populations that are commonly and consistently reported on (e.g., White or Black).	• Counts for minority populations decrease. • This is dependent on the data collection and update processes at the facility/payor level. • Would likely be primarily populated by the APCD since monthly eligibility data is submitted.
Most Recent Race/Ethnicity: For each individual, a race and ethnicity are assigned based on the value that shows up from the most recent data point.	• Data is assumed to be reflective of how the individual most recently identified.	• May be reflective of data entry or collection errors. • Counts for minority populations decrease. • This is dependent on the data collection and update processes at the facility/payor level for whichever data point is the most recent for an individual. • Would likely be primarily populated by the APCD since monthly eligibility data is submitted.
Custom Logic: A customized criteria for assigning race and ethnicity information based on performing a data quality assessment.	• Allows for flexibility for users in defining groups. • Customizable based on data quality findings and best practices. • Can be modified over time as additional data quality assessments are performed.	• Methodology may be difficult to replicate or may not be applicable to other projects.

Custom Logic Implemented

Stepwise Logic	Reason Behind Step
Step 1. A rolling 5-year data rule was applied, meaning data for the most recent five years is used as additional data gets pulled into the databases.	• Uses most recent data • Controls for data entry or submission errors by eliminating 'bad data' data over time
Step 2. For each person ID, a column is available for each of the minimum race and ethnicity categories with additional columns for 'Other Race,' 'Unknown Race,' and 'Unknown Ethnicity.' These columns are populated based on these categories ever being present in the data.	• Inclusion of all viable data • Allows end user flexibility in data group selections
Step 3. Row level data with only 'Other Race', 'Unknown', or 'Race Refused' were set aside.	• Viable data is used first
Step 4. Two data sets were created based on submitter history and applied to unpopulated rows in the following order: • 1st Submitters that had reported all minimum categories at least once • 2nd Remaining Submitters	• Prioritizes information from submitters that were known to collect all racial categories
Step 5. For individuals that had both Hispanic and Non-Hispanic reported, it was decided to defer to the rarest ethnicity: Hispanic.	• In these edge cases, at some point from either system they had chosen to identify as Hispanic
Step 6. Individuals that did not have information populated from any step above, the unknown race and ethnicity categories were populated.	
Step 7. Multiracial Flags are Created: • Source Flag (Based on Submitter Data) • Inferred Flag (Based on Merged Data)	• Allows User to choose multiracial definition
Step 8. Race-Refused Flag is Created based on most recent data submission being 'Race refused'.	• Allows User to choose how they want to handle this information

Consistency Across Data Sources

98.0% Race Data

97.1% Ethnicity Data

95.2% Both Race and Ethnicity Data

Conclusions

From the initial data assessment to this iteration, the same underlying data challenges were present with minimal changes. Although custom logic was implemented as opposed to a commonly used methodology, the logic fit our goals to allow for data analysts to customize their demographic examinations of the data within these data systems. This custom logic took elements from standard race and ethnicity logic where it made sense, but prioritized information based on the findings of the data quality assessment. The person-level composite allows for the best guess at the most accurate racial and ethnic identities for the individuals that appear within these data sets, but still allows for flexibility for the data users. MHDO's de-identified person-level composite enhances the MHDO's data in that it provides consistency across data sets for longitudinal analyses and allows for a better understanding and monitoring inequity that exists in health care.