Enhancing the granularity of Native Hawaiian and Pacific Islander (NHPI) data at the United States Department of Veterans Affairs using unstructured data and an expanded Race/Ethnicity Lexicon

Benjamin Viernes[1,3], Patrick R Alba[1,2], Qiwei Gan[1,2], Elizabeth E Hanchrow[1], Mengke Hu[1,2], Gregorio Coronado[1,3], Scott L DuVall[1,2], Kalani Raphael[1-3]

1 – *VA Salt Lake City Health Care System, Salt Lake City, UT, USA*
2 – Department of Internal Medicine, University of Utah Medical School, Salt Lake City, UT, USA
3 – Center for Native Hawaiians, Pacific Islander, and US Affiliated Pacific Islander Veterans. VA Pacific Islands Healthcare System. Honolulu, HI, USA

Background

Native Hawaiian and Pacific Islander (NHPI) populations have unique genetic, environmental, and cultural factors that impact their health outcomes. Current medical research often underrepresents or inadequately categorizes these groups, resulting in a lack of nuanced understanding and effective healthcare interventions. Existing race, ethnicity, and nationality categories in standardized data structures also fail to capture the diversity within NHPI communities, although the breadth of cultures, peoples, languages, geographies, and origins is vast across the pacific cultures generally categorized under United States Office of Management and Budget (OMB) Statistical Policy Directive (SPD) 15 race classification: “Native Hawaiian or Pacific Islander.”[1] Further, the combination of Native Hawaiians (the largest group of NHPI) with other pacific islanders obscures any disparities unique to smaller populations of pacific islanders.

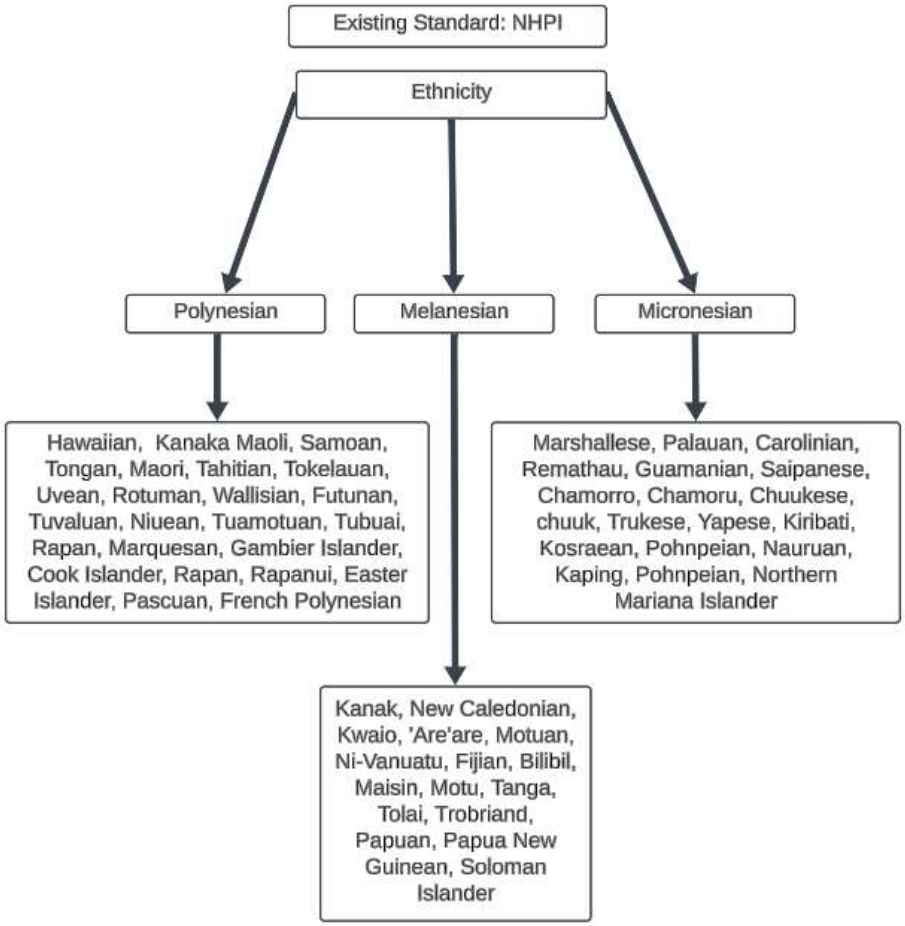
Current electronic health records contain accessible clinical text for large patient populations. Frequently this text contains granular information that may or may not be uniformly documented. This analysis sought to accomplish two aims: first, to identify whether more granular categories could be determined for subsets of the NHPI population that could be useful for understanding the heterogeneity present across the larger population; second, to identify whether populations for which race is ‘Unknown’ or for which only vague legacy race categories exist (i.e., ‘Asian or Pacific Islander’) could be classified within the NHPI population.

We hypothesize that clinical notes can be utilized in two ways, first to disaggregate legacy race and ethnicity standard OMB categories and second, to identify patients with missing race and ethnicity altogether.

Methods

We created a lexicon of terms that map to initial standards of Polynesian, Melanesian, and Micronesian. We then extract all instances of mapped key terms from clinic notes within our cohort. These terms encompassed specific geographic locations of islands and island groups throughout Polynesia, Melanesia, and Micronesia, as well as languages spoken throughout the Pacific. An example of the terms identified can be seen in figure 1; these terms and hierarchies are constantly being expanded and reconsidered.

Figure 1: Working Body of NHPI Race and Ethnicity Descendants



The existence of a key term in clinical notes alone does not reflect a patients race or ethnicity. To better understand the contexts in which these terms exist, we developed an annotation schema to classify the context of NHPI terms found in notes, with contexts including (1) Is the term referencing the patient's race/ethnicity? (2) Did it refer to a location? (3) Did it refer to a language? In the latter two contexts, additional questions asked annotators to identify how the location or language related to the origin/heritage/nationality/identity of the patient (e.g., "Was the patient born in/from the location?", "Was the patient a native speaker of the language?", etc.). Clinical annotators then reviewed and labeled a sample of the extracted terms creating an initial set of labels for evaluation. These annotated instances were split with 326 cases for training, 54 for validation, and 55 for testing. A preliminary NLP model was trained using a transformer-based architectures.

15,252,145 Veterans who utilize care in the Department of Veterans Affairs (VA) and have electronic health record data were identified for inclusion. Data from the VA Corporate Data Warehouse (CDW) and VA OMOP CDM were used to identify structured race data and associated clinical notes.

Results

Preliminary NLP Model:

The model performs well on classifying a term indicating the patients' race/ethnicity with a precision of 0.96, recall of 0.95, and F1 score of .95. Using this preliminary model, we then randomly sampled 1000 cases per unique term found in the lexicon within the training cohort of documents. Due to the relative infrequency of some terms, this set of terms totaled 30,746 cases. We then predict on these cases with the initial model finding that 13,387 of these terms could be classified as identifying the patients' race/ethnicity, showing that while the existence of key terms alone cannot positively identify a patient's race/ethnicity, we find that with this preliminary NLP model these more granular concepts are frequently documented as race/ethnicity within patient notes.

NHPI Documentation and Disaggregation:

Of Veterans who utilize care in the VA, 669,066 were identified with the Race/ethnicity categorization of Asian Pacific Islander (API) or Native Hawaiian or Pacific Islander (NHPI). Of those VA patients, 150,586 were found to have NHPI-related terms. Of those API patients with NHPI-related terms, 115,869 patients had terms classified as NHPI documentation by the NLP model. Combined with structured data, NLP classified 165,437 patients as having NHPI documentation. Of those with documentation, NLP was able to classify documentation for 35,554 as Polynesian, 12,917 as Micronesian, and 2,208 as Melanesian.

We also identified 3,674,646 VA patients with "unknown" or missing race; of these patients, 71,930 were found to have key terms within their notes related to NHPI ethnicity and race. 18,530 patients with "unknown" or missing race had terms classified as NHPI documentation by the NLP model. Of these 18,530 patients with NHPI documentation, NLP was able to classify documentation for 3,542 as Polynesian, 2,303 as Micronesian, and 391 as Melanesian.

Conclusions

Expanding race, ethnicity, and nationality categories to include more detailed NHPI subgroups in standardized data structures can further advance medical research and improve health outcomes for these populations. The analysis of clinical text from EHRs revealed that NHPI populations are frequently aggregated into broader racial categories, such as "Asian or Pacific Islander" and "Native Hawaiian or Other Pacific Islander" obscuring critical health trends specific to these groups and more granular subsets.[4]

Although the historical construct of race, the current usage and standards, and the data available are all limitations in the utility of disaggregating race data for identifying health disparities, the identification of expanded, more granular categories will enhance the utility of health data, facilitate more focused interventions, and promote health equity. While this proof of concept model and results are promising, the next steps for this work include the annotation of a larger set of terms and further model development with the goal of eventual inclusion of this data as standardized concepts within the VA's OMOP CDM.

References

1. Office of Management and Budget. Revisions to OMB's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. Federal Register Notice. Available at: https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and. Published 2024.
2. Mau MK, Sinclair K, Saito EP, Baumhofer KN, Kaholokula JK. Cardiometabolic health disparities in Native Hawaiians and other Pacific Islanders. Epidemiol Rev. 2009;31(1):113-129.
3. Evan T Sholle, Laura C Pinheiro, Prakash Adekkanattu, Marcos A Davila, Stephen B Johnson, Jyotishman Pathak, Sanjai Sinha, Cassidie Li, Stasi A Lubansky, Monika M Safford, Thomas R Campion, Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation, Journal of the American Medical Informatics Association, Volume 26, Issue 8-9, August/September 2019, Pages 722–729, https://doi.org/10.1093/jamia/ocz040
4. Quint J, Matagi C, Kaholokula JK. The Hawai'i NHPI Data Disaggregation Imperative: Preventing Data Genocide Through Statewide Race and Ethnicity Standards. Hawaii J Health Soc Welf. 2023 Oct;82(10 Suppl 1):67-72. PMID: 37901675; PMCID: PMC10612414.
5. United States Government Accountability Office Report to Congressional Committees: VETERANS AFFAIRS Actions Needed to Improve Access to Care in the U.S. Territories and Freely Associated States. GAO-24-106364 VA Benefits in Territories and FAS May 23, 2024
6. Nelson A. Unequal treatment: confronting racial and ethnic disparities in health care. J Natl Med Assoc2002; 94 (8): 666–8.
7. U.S. Department of Health and Human Services. National Healthcare Disparities Report 2011. Publication 12-0006. Rockville, MD: Agency for Healthcare Research and Quality; 2012.
8. McGarry ME , McColley SA. Minorities are underrepresented in clinical trials of pharmaceutical agents for cystic fibrosis. Ann Am Thorac Soc2016; 13 (10): 1721–5.