Research Letter

Racial Misclassification of American Indian and Alaska Native People in the Electronic Medical Record: An Unexpected Hurdle in a Retrospective Medical Record Cohort Study

Ann Marie Rusk[1], MD; Alanna M Chamberlain[2], PhD; Jamie Felzer[3], MPH, MD; Yvonne Bui[2], MD; Christi A Patten[2], PhD; Christopher C Destephano[4], MPH, MD; Matthew A Rank[1], MD; Roberto P Benzo[1], MD; Cassie C Kennedy[2], MS, MD

[1]Mayo Clinic in Arizona, Scottsdale, AZ, United States
[2]Mayo Clinic, Rochester, MN, United States
[3]Emory University, Atlanta, GA, United States
[4]Mayo Clinic in Florida, Jacksonville, FL, United States

Corresponding Author:

Ann Marie Rusk, MD
Mayo Clinic in Arizona
13400 E. Shea Blvd
Scottsdale, AZ 85259
United States
Email: rusk.ann@mayo.edu

Abstract

Electronic health record data represent a rich data source; however, data accuracy must be considered prior to reporting health outcomes among American Indian and Alaska Native people. Using a hybrid approach to harmonizing data from multiple sources represents a valid method of assessing data integrity in this population.

J Med Internet Res 2025;27:e73086; doi: 10.2196/73086

Keywords: electronic health records; EHR; health care disparities; Indigenous health; American Indian or Alaska Native; social determinants of health

Introduction

Indigenous North Americans (American Indian and Alaska Native [AI/AN] people) in the United States have the shortest life expectancy among all racial or ethnic groups [1]. Disparate health and survival outcomes are influenced by social determinants of health (SDOHs)—factors that influence birth, health, life, and death—including health behaviors and systemic factors (eg, health care access) [2]. Addressing health care disparities requires equitable representation in public health data. In a retrospective cohort study that examined longitudinal cigarette smoking behaviors of Indigenous people in Olmsted County, Minnesota—a county without access to Indian Health Service clinics or hospitals—the magnitude of racial misclassification in electronic health record (EHR) data became an unexpected hurdle for the study team [3]. Most AI/AN people reside in urban areas or off reservation lands [4]. Understanding this population's health behaviors is critical to informing interventions. Herein, we describe methods for harmonizing race data from multiple record sources to assure this frequently underrepresented and mischaracterized population's accurate representation.

Methods

Individuals with vital records (birth or death certificate) or EHR data (provider histories, EHR flowsheets, self-reports, or nursing documentation) indicating AI/AN race were identified in a longitudinal cohort study (2006-2019) to assess smoking behaviors and pharmaceutical cessation aid uptake by race, sex, age, and indexed SDOHs [3]. Inclusion criteria were AI/AN race and availability of ≥1 year of smoking data. Exclusion criteria included non-AI/AN race and no smoking data available from 2006 to 2019. Patients were identified in the Rochester Epidemiology Project—a medical-record linkage system (established since 1966) inclusive of multiple health care delivery systems and population data for 99.9% of Olmsted County residents [5,6]. AMR, JF, and YB conducted data cleaning for all available records to resolve discordant records of AI/AN race; this included manual review of narrative EHR data, exclusion of individuals who used foreign language translation services, and review of patients' vital records and records of parents and offspring. A sex- and age-matched (±5 years) non-AI/AN cohort was compared to the study cohort.

Results

In total, 1271 individuals with ≥1 record indicating AI/AN race were identified; 148 were excluded (missing smoking data: n=124; no 2006-2019 EHR data: n=24). Manual review of the AI/AN cohort's race and ethnicity data revealed 25 individuals who reported immigration from a non–North American country, and 200 individuals required foreign language interpreters for languages originating outside of North America (primarily languages originating from the Indian subcontinent and Southeast Asia). Final data cleaning resulted in a cohort of 898 AI/AN patients, demonstrating 17.7% (225/1271) racial misclassification [3] (Figure 1). The annual smoking prevalence for race-misclassified individuals (n=225) ranged between 8% and 23%; that for the AI/AN cohort (n=898) ranged between 39% and 47% (Figure 2). The matched cohort included 1780 individuals (White: n=1483, 83.3%; Black/African American: n=105, 5.9%; Asian: n=96, 5.4%; Hawaiian/Pacific Islander: n=4, 0.2%; other: n=68, 3.8%; declined: n=8, 0.4%; unknown: n=16, 0.9%) [3].

Figure 1. Inclusion and exclusion criteria applied to identify a cohort of American Indian or Alaska Native individuals in the Rochester Epidemiology Project from 2006 to 2019.

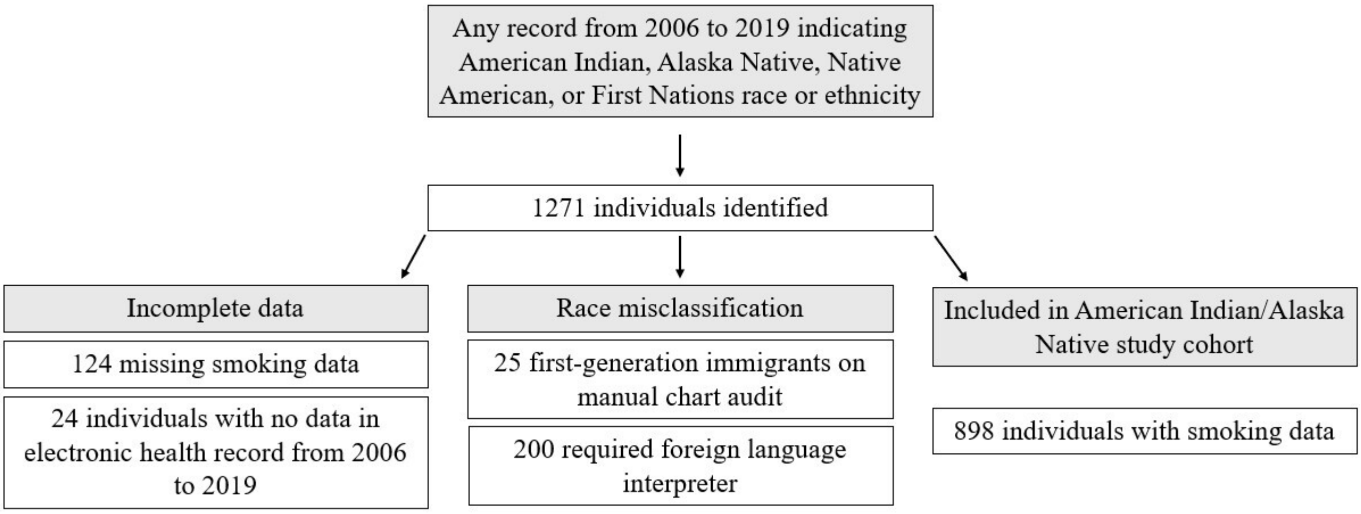
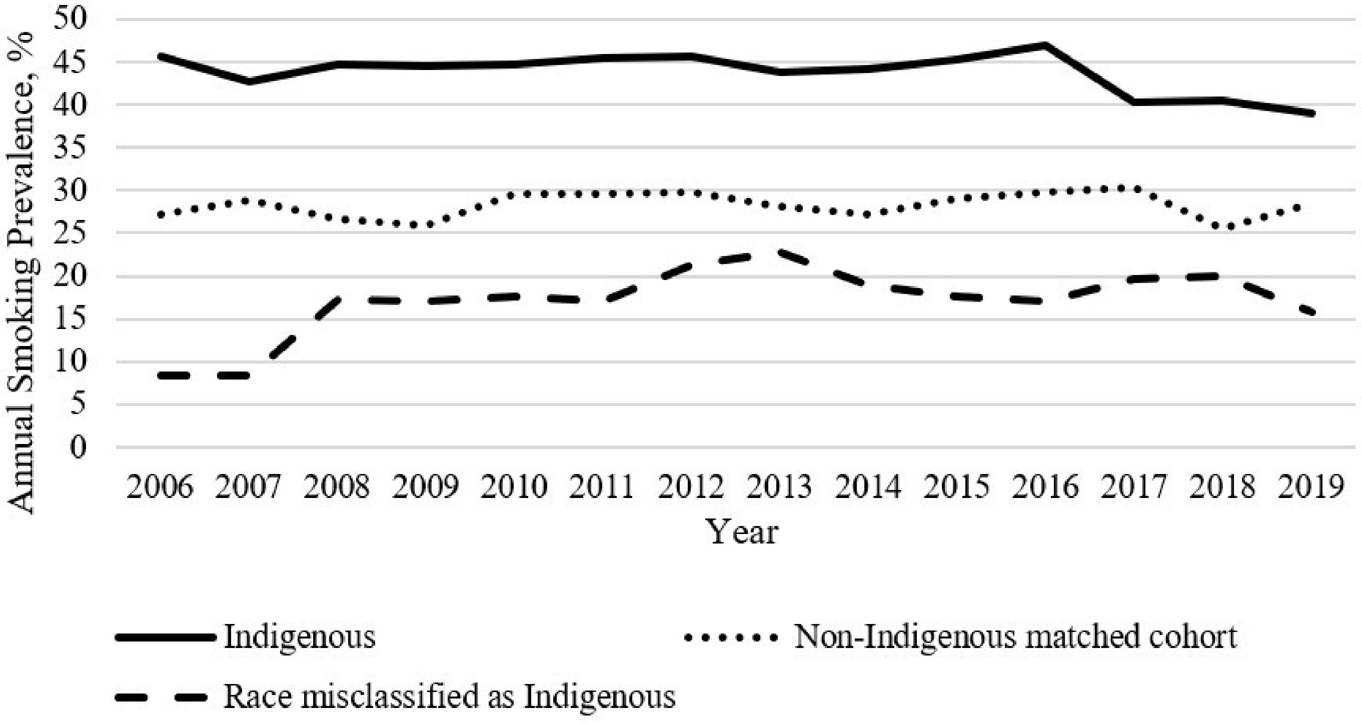

Any record from 2006 to 2019 indicating American Indian, Alaska Native, Native American, or First Nations race or ethnicity
1271 individuals identified
Incomplete data
124 missing smoking data
24 individuals with no data in electronic health record from 2006 to 2019
Race misclassification
25 first-generation immigrants on manual chart audit
200 required foreign language interpreter
Included in American Indian/Alaska Native study cohort
898 individuals with smoking data

Figure 2. Annual smoking prevalence.



Discussion

Harmonization of vital records and multiple EHR sources proved essential, as the magnitude of race misclassification (17.7%) in this study was higher than that in other AI/AN population studies, including a review of mortality data among AI/AN people in Washington State (12%) [7]. Without cohort validation, this study's smoking prevalence would have been falsely lower due to lower smoking prevalence among race-misclassified individuals. Smoking behavior misrepresentation in medical literature would further exacerbate health care disparities in this underrepresented population. The Centers for Medicare & Medicaid Services have recognized the need to standardize data entry, releasing resources for health care organizations to improve demographic accuracy [8]. Until standardized data entry is implemented, additional methods for validating historical race data are necessary [9]. Data linkage—the harmonization of an individual's data across different sources—represents a valid methodology [10]. Using a hybrid approach to AI/AN cohort validation—manual review of narrative documentation, vital records, and EHR input across multiple health systems—represents a potential method for smaller epidemiological studies. This study's limitations included the inability to link data with tribal registries or Indian Health Service data and the time required to manually review records. Our methods may be used for counties where AI/AN individuals lack access to tribal health facilities (eg, Olmsted County). Besides manual data review, studies including AI/AN people should be conducted in concert with AI/AN people and tribes. This study was designed and conducted with oversight by an AI/AN community advisory board that expressed the critical importance of accurate race data. Studies using EHR data inclusive of AI/AN people should include measures for ensuring accurate race data and representation.

Acknowledgments

The content of this article is the sole responsibility of the authors and does not represent the views of the National Institutes of Health, Mayo Clinic, or the Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery. The authors thank the Healthy Nations Advisory Board for their input and oversight of this study. This study was supported by funding from the Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, the Rochester Epidemiology Project (REP) Scholarship, the Robert A. Winn Career Development Award, and the American Thoracic Society Fellowship in Health Equity and Diversity. The REP medical-record linkage system is supported by the National Institute on Aging (NIA, AG 058738), the Mayo Clinic Research Committee, and fees paid annually by REP users. There was no use of generative artificial intelligence for any version of drafts, outlines, writing, text, or figures in this paper.

Authors' Contributions

Conceptualization: AMR, AMC, CAP, CCK
Data curation: AMR, AMC
Formal analysis: AMR, JF, YB
Funding acquisition: AMR
Investigation: AMR, CCK
Methodology: AMR, CCK
Project administration: AMR
Resources: AMR, CCK
Software: CCK
Supervision: AMC, CAP, MAR, RPB, CCK
Validation: AMR, CCK
Visualization: AMR
Writing – original draft: AMR
Writing – review & editing: AMR, AMC, CAP, CCD, MAR, RPB, CCK

Conflicts of Interest

None declared.

References

1. Arias E, Xu J, Kochanek K. United States life tables, 2021. Natl Vital Stat Rep. Nov 2023;72(12):1-64. [Medline: 38048433]
2. Alvidrez J, Castille D, Laude-Sharp M, Rosario A, Tabor D. The National Institute on Minority Health and Health Disparities research framework. Am J Public Health. Jan 2019;109(S1):S16-S20. [doi: 10.2105/AJPH.2018.304883] [Medline: 30699025]
3. Rusk AM, Giblon RE, Chamberlain AM, et al. Indigenous smoking behaviors in Olmsted County, Minnesota: a longitudinal population-based study. Mayo Clin Proc. Oct 2022;97(10):1836-1848. [doi: 10.1016/j.mayocp.2022.03.019] [Medline: 36202495]

4. Public Use Microdata Sample. United States Census Bureau. 2019. URL: https://www.census.gov/programs-surveys/acs/microdata/access.html [Accessed 2025-06-06]

5. St Sauver JL, Grossardt BR, Yawn BP, Melton LJ 3rd, Rocca WA. Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. Am J Epidemiol. May 1, 2011;173(9):1059-1068. [doi: 10.1093/aje/kwq482] [Medline: 21430193]

6. St Sauver JL, Grossardt BR, Leibson CL, Yawn BP, Melton LJ 3rd, Rocca WA. Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester Epidemiology Project. Mayo Clin Proc. Feb 2012;87(2):151-160. [doi: 10.1016/j.mayocp.2011.11.009] [Medline: 22305027]

7. Dankovchik J, Hoopes M, Nordstrom DL, Knaster E. Racial misclassification and disparities in mortality among AI/AN and other races, Washington. Agency for Healthcare Research and Quality. 2012. URL: https://hcup-us.ahrq.gov/datainnovations/raceethnicitytoolkit/or26.jsp [Accessed 2024-12-30]

8. Inventory of resources for standardized demographic and language collection. Centers for Medicare & Medicaid Services. Mar 2024. URL: https://www.cms.gov/about-cms/agency-information/omh/downloads/data-collection-resources.pdf [Accessed 2024-05-06]

9. Martinez RAM, Andrabi N, Goodwin AN, Wilbur RE, Smith NR, Zivich PN. Conceptualization, operationalization, and utilization of race and ethnicity in major epidemiology journals, 1995-2018: a systematic review. Am J Epidemiol. Feb 24, 2023;192(3):483-496. [doi: 10.1093/aje/kwac146] [Medline: 35938872]

10. Chin MK, Đoàn LN, Russo RG, et al. Methods for retrospectively improving race/ethnicity data quality: a scoping review. Epidemiol Rev. Dec 20, 2023;45(1):127-139. [doi: 10.1093/epirev/mxad002] [Medline: 37045807]

Abbreviations

AI/AN: American Indian and Alaska Native
EHR: electronic health record
SDOH: social determinant of health

Edited by Javad Sarvestan; peer-reviewed by Joseph Pacheco, Sarah H Nash; submitted 17.03.2025; final revised version received 10.06.2025; accepted 11.06.2025; published 30.07.2025

Please cite as:
Rusk AM, Chamberlain AM, Felzer J, Bui Y, Patten CA, Destephano CC, Rank MA, Benzo RP, Kennedy CC
Racial Misclassification of American Indian and Alaska Native People in the Electronic Medical Record: An Unexpected Hurdle in a Retrospective Medical Record Cohort Study
J Med Internet Res 2025;27:e73086
URL: https://www.jmir.org/2025/1/e73086
doi: 10.2196/73086