Racial misclassification of American Indian and Alaska Native people in the Electronic Medical Record: an unexpected hurdle in a retrospective medical records cohort study

Ann Marie Rusk, Alanna M Chamberlain, Jamie Felzer, Yvonne Bui, Christi A. Patten, Christopher C. Destephano, Matthew A. Rank, Roberto P Benzo, Cassie C. Kennedy

Submitted to: Journal of Medical Internet Research
on: March 17, 2025

Table of Contents

Racial misclassification of American Indian and Alaska Native people in the Electronic Medical Record: an unexpected hurdle in a retrospective medical records cohort study

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

[1] Mayo Clinic in Arizona Scottsdale US
[2] Mayo Clinic Rochester US
[3] Emory University Atlanta US
[4] Mayo Clinic in Florida Jacksonville US

Corresponding Author:
Ann Marie Rusk MD

Mayo Clinic in Arizona
13400 E. Shea Blvd
Scottsdale
US

Abstract

Electronic health record (EHR) data represents a rich data source, however data accuracy must be considered prior to reporting health outcomes among American Indian and Alaska Native people.

(JMIR Preprints 17/03/2025:73086)
DOI: https://doi.org/10.2196/preprints.73086

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Authors:
Ann M. Rusk, M.D.; Alanna M. Chamberlain, Ph.D; Jamie Felzer, M.D., M.P.H; Yvonne Bui, M.D.; Christi A. Patten, Ph.D; Christopher C. DeStephano, M.D., M.P.H.; Matthew A. Rank, M.D.; Roberto P. Benzo. M.D.; Cassie C. Kennedy, M.D., M.S.
Title: Racial misclassification of American Indian and Alaska Native people in the Electronic Medical Record: an unexpected hurdle in a retrospective medical records cohort study
Abstract: Electronic health record (EHR) data represents a rich data source, however data accuracy must be considered prior to reporting health outcomes among American Indian and Alaska Native people.
Keywords: Electronic health records; Healthcare disparities, Indigenous Health
Introduction:
Indigenous North Americans (American Indian and Alaska Native [AI/AN] people) in the United States (US) have the shortest life expectancy among all racial or ethnic groups[1]. Disparate outcomes in health and survival are due to the confluence of factors that influence birth, health, life, and death, known as social determinants of health[2]. Addressing healthcare disparities among the AI/AN population requires equitable representation in public health data.[3] In a retrospective cohort study examining longitudinal cigarette smoking behaviors of Indigenous people residing in Olmsted County, Minnesota, the magnitude of racial misclassification in EHR data became a significant challenge in the verification of data accuracy.[4, 5] We describe methods used to assure equitable and accurate representation of this frequently underrepresented and mischaracterized population by harmonizing race data from multiple record sources.

Methods:
Individuals who had a vital record (birth or death certificate) or EHR data (provider history, and EHR flowsheet data, self-report, or nursing documentation) indicating AI/AN race were identified in a longitudinal cohort study (2006-2019) to assess cigarette smoking behaviors and pharmaceutical cessation aid uptake informed by race, sex, age, and an indexed social determinant of health measure. Inclusion criteria were AI/AN race, and availability of at least one year of smoking data. Exclusion criteria included non-AI/AN race, and no smoking data available between 2006-2019. Subjects were identified in the Rochester Epidemiology Project (REP), a medical-records linkage system inclusive of multiple healthcare delivery systems and population data for people residing in Olmsted County, Minnesota[6]. The REP has been established since 1966 and is inclusive of 99.9% of Olmsted County residents[7]. Data cleaning of all available records to resolve discordant records of AI/AN race included manual review of documentation narratives, exclusion of individuals using foreign language translation services, and review of vital records for the patient and records of parents and offspring when available. A non-AI/AN cohort was matched 1:1 on age (±5 years) and sex to the AI/AN cohort.

Results:
A total of 1271 individuals with at least one record indicating AI/AN race were identified. No smoking data was available for 124 individuals, and 24 individuals did not have EHR data between 2006 and 2019. A manual review of race and ethnicity data in the AI/AN study cohort revealed 25 individuals reporting immigration to the US from a country outside of North America, and 200 individuals required use of a foreign language interpreter for a language originating outside of the North American continent, most frequently reported regions of origin included the Indian Subcontinent and Southeast Asia. Final data cleaning resulted in a cohort of 898 AI/AN patients, demonstrating racial misclassification of 225/1271 (17.7%) in the primary cohort,[5] as demonstrated in (Figure 1). The smoking prevalence for race-misclassified individuals (n=225) revealed an annual smoking prevalence between 8-23%, compared to annual smoking prevalence between 39-47% in the AI/AN study cohort (n=898) demonstrated in (Figure 2).

Discussion:

Harmonization of vital records and multiple EHR data sources proved essential during cohort identification in this study. The magnitude of disagreement (17.7%) in this study was higher than other studies inclusive of AI/AN populations, including a review of longitudinal mortality data among AI/AN people in the state of Washington demonstrating race misclassification of 12%[8]. Without careful cohort validation, the smoking prevalence of this study would have been falsely lower due to lower smoking prevalence among race-misclassified individuals. Until implementation of standardized data entry has occurred, additional methods to review the accuracy of race demographic data are necessary[9]. A methodology to address race misrepresentation and improve the quality of race data includes data linkage, the combination of information belonging to the same individual across different data sources[10]. Utilizing a hybrid approach when conducting AI/AN cohort validation, which consists of a manual review of narrative documentation, vital records, and EHR input across multiple health systems, represents a potential method for use in smaller epidemiological studies. In addition to manual review to ensure high data quality, the conduct of studies seeking to address disparities experienced by 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, whose members also expressed the critical importance of accurate race data. The design and conduct of studies utilizing EHR data that include the AI/AN population must include careful consideration of the accuracy of demographic data to ensure accurate representation.

Acknowledgements:
The content of this article is the sole responsibility of the authors and does not represent the views of the National Institute of Health, Mayo Clinic, or the Robert D. and Patricia E. Kern Center for the Science of Healthcare Delivery. The authors thank the Healthy Nations Advisory Board for their input and oversight of this study. Grant Support: This study was supported by funding from the Robert D. and Patricia E. Kern Center for the Science of Healthcare Delivery, the Rochester Epidemiology Project Scholarship, the Robert A. Winn Career Development Award, and the American Thoracic Society Fellowship in Health Equity and Diversity. The Rochester Epidemiology Project (REP) medical records-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.
References:

1. Arias E, Xu J, Kochanek K. United States Life Tables, 2021. Natl Vital Stat Rep. 2023 Nov;72(12):1-64. PMID: 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. 2019 Jan;109(S1):S16-S20. PMID: 30699025. doi: 10.2105/AJPH.2018.304883.
3. Bauer UE, Plescia M. Addressing disparities in the health of American Indian and Alaska Native people: the importance of improved public health data. Am J Public Health. 2014 Jun;104 Suppl 3(Suppl 3):S255-7. PMID: 24754654. doi: 10.2105/AJPH.2013.301602.
4. Rusk AM, Giblon RE, Chamberlain AM, Patten CA, Felzer JR, Bui YT, et al. Smoking Behaviors Among Indigenous Pregnant People Compared to a Matched Regional Cohort. Nicotine Tob Res. 2022 Oct 17. PMID: 36250476. doi: 10.1093/ntr/ntac240.
5. Rusk AM, Giblon RE, Chamberlain AM, Patten CA, Felzer JR, Bui YT, et al. Indigenous Smoking Behaviors in Olmsted County, Minnesota: A Longitudinal Population-Based Study. Mayo Clin Proc. 2022 Oct;97(10):1836-48. PMID: 36202495. doi: 10.1016/j.mayocp.2022.03.019.
6. 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. 2011 May 1;173(9):1059-68. PMID: 21430193. doi: 10.1093/aje/kwq482.
7. 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. 2012 Feb;87(2):151-60. PMID: 22305027. doi: 10.1016/j.mayocp.2011.11.009.
8. Jenine Dankovchik MH, David L. Nordstrom, Elizabeth Knaster. Racial Misclassification and Disparities in Mortality among AI/AN and Other Races, Washington. Agency for Healthcare Research and Quality, Rockville, MD2014 [updated 8/19/2014; cited 2024 12/30/2024]; Available from: https://hcup-us.ahrq.gov/datainnovations/raceethnicitytoolkit/or26.jsp.
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. 2023 Feb 24;192(3):483-96. PMID: 35938872. doi: 10.1093/aje/kwac146.
10. Chin MK, Doan LN, Russo RG, Roberts T, Persaud S, Huang E, et al. Methods for retrospectively improving race/ethnicity data quality: a scoping review. Epidemiol Rev. 2023 Dec 20;45(1):127-39. PMID: 37045807. doi: 10.1093/epirev/mxad002.

Supplementary Files

Figures

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

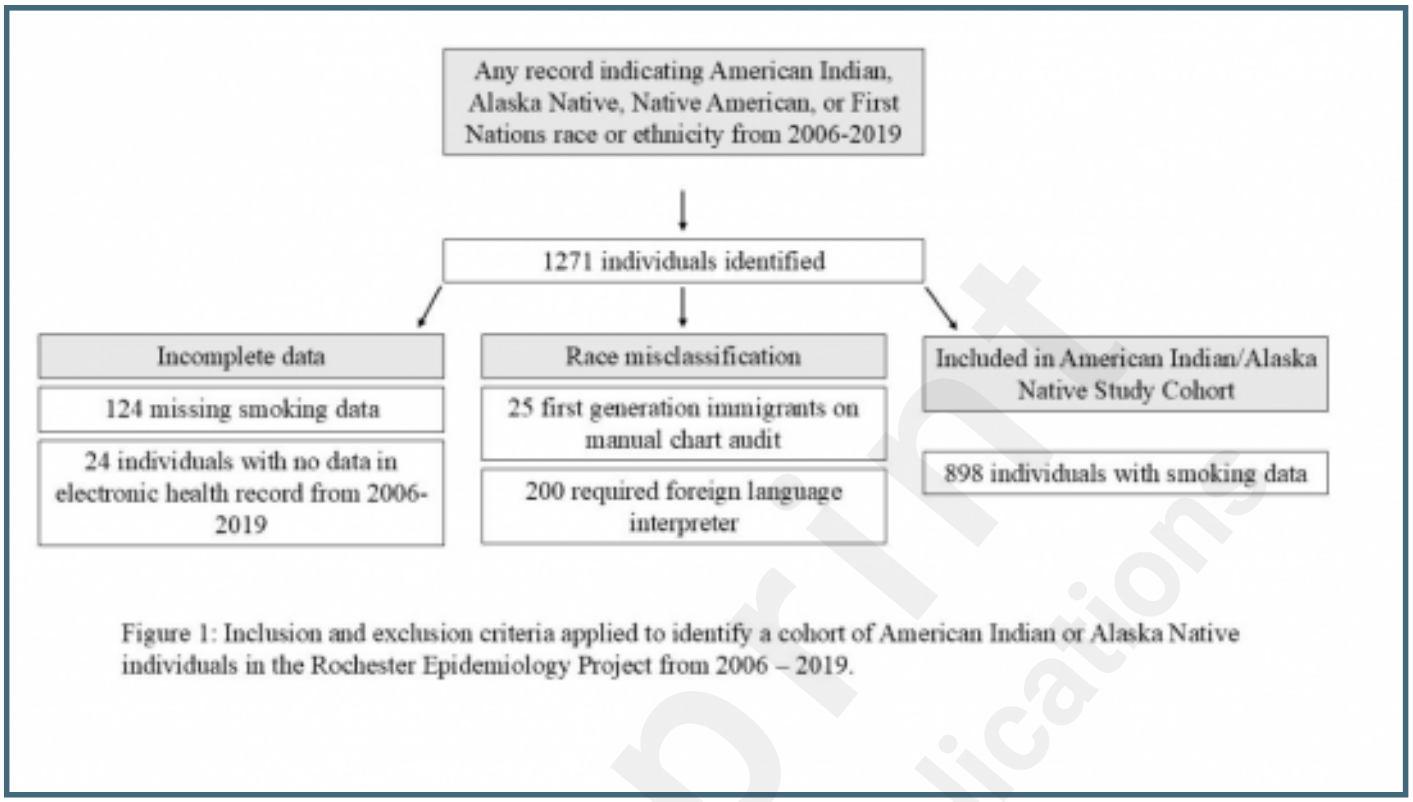

Figure 1: Inclusion and exclusion criteria applied to identify a cohort of American Indian or Alaska Native individuals in the Rochester Epidemiology Project from 2006 – 2019.

Annual smoking prevalence.