TITLE PAGE

Enhanced Identification of Hispanic Ethnicity Using Clinical Data: A Study in the Largest Integrated United States Health Care System

Authors: Pedro Ochoa-Allemant, MD[1]; Janet P. Tate, ScD[1,2]; Emily C. Williams, PhD, MPH[3,4]; Kirsha S. Gordon, PhD, MS[1,2]; Vincent C. Marconi, MD[5,6]; Kara M.K. Bensley, PhD, MSc[7]; Christopher T. Rentsch, PhD, MPH[1,2,8]; Karen H. Wang, MD, MHS[9,10]; Tamar H. Taddei, MD[2,11]; Amy C. Justice, MD, PhD[1,2,12] for the VA Family of EHR Cohorts (VACo Family)

Institutions:

1 Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

2 VA Connecticut Healthcare System, US Department of Veteran Affairs, West Haven, CT, USA

3 Denver-Seattle Center of Innovation for Veteran-Centered and Value-Driven Care, VA Puget Sound Health Services Research & Development, Seattle, WA, USA

4 Department of Health Services, University of Washington, Seattle, WA, USA

5 Emory University, Atlanta, GA, USA

6 Atlanta Veterans Affairs Medical Center, Atlanta, GA, USA

7 Department of Public Health, Bastyr University, Kenmore, WA, USA

8 Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK

9 Equity Research and Innovation Center, Section of General Internal Medicine, Yale School of Medicine, New Haven, CT, USA

10 Center for Medical Informatics, Yale School of Medicine, New Haven, CT, USA

11 Section of Digestive Diseases, Yale School of Medicine, New Haven, CT, USA

12 Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA

Corresponding Author:

Amy C. Justice, MD, PhD

VA Connecticut Healthcare System

950 Campbell Avenue, 11-ACSLG

West Haven, CT 06516-2770, USA

amy.justice2@va.gov

Tel: 203-932-5711 x3541

Disclosure:

No conflicts of interest to disclose.

Funding:

National Institutes of Health: NCI R01-CA206465; NIAAA U24-AA020794, U01-AA020790, U24-AA022001, U10 AA013566-completed; Department of Veterans Affairs, Office of Research and Development, Million Veteran Program MVP000.

Abbreviations:

BMI, body mass index; CDW, Corporate Data Warehouse; CI, confidence interval; CMS, Centers for Medicare and Medicaid Services; COVID-19, coronavirus disease 2019; CVD, cerebrovascular disease; EHR, electronic health record; FY, fiscal year; HIV, human immunodeficiency virus; HCC, hepatocellular carcinoma; ICD, International Classification of Diseases; NAACCR, North American Association of Central Cancer Registries; NCI, National Cancer Institute; NH, non-Hispanic; NHIA, NAACCR Hispanic Identification Algorithm; OMB, Office of Management and Budget; RTI, research triangle institute; RUCA, rural-urban commuting area; SEER, Surveillance, Epidemiology, and End Results Program; U.S., United States; VA, Veterans Affairs; VINCI, VA Informatics and Computing Infrastructure; VACS, Veterans Aging Cohort Study.

Electronic word count: 3,116 words (including only main text)

Number of references: 41

Number of Tables/Figures: 3 tables and 2 figures

Number of text pages: 15

ABSTRACT (word count: 244/250)

Background: Collection of accurate Hispanic ethnicity data is critical to evaluate disparities in health and health care. However, this information is often inconsistently recorded in electronic health record (EHR) data.

Objective: To enhance capture of Hispanic ethnicity in the Veterans Affairs (VA) EHR and compare relative disparities in health and health care.

Methods: We first developed an algorithm based on surname and country of birth. We then determined sensitivity and specificity using self-reported ethnicity from the 2012 Veterans Aging Cohort Study (VACS) survey as the reference standard and compared this to the RTI race variable from the Medicare administrative data. Finally, we compared demographic characteristics and age-and sex-adjusted prevalence of conditions in Hispanic patients among different identification methods in the VA EHR 2018-2019.

Results: Our algorithm yielded higher sensitivity than either EHR-recorded ethnicity and the RTI race variable. In 2018-2019, Hispanic patients identified by the algorithm were more likely to be older, had a race other than White, and foreign-born. The prevalence of conditions was similar between EHR and algorithm ethnicity. Hispanic patients had higher prevalence of diabetes, gastric cancer, chronic liver disease, hepatocellular carcinoma, and HIV than non-Hispanic White patients. Our approach evidenced significant differences in burden of disease among Hispanic subgroups by nativity status and country of birth.

Conclusions: We developed and validated an algorithm to supplement Hispanic ethnicity information using clinical data in the largest integrated U.S. healthcare system. Our approach enabled clearer understanding of demographic characteristics and burden of disease in the Hispanic Veteran population.

KEYWORDS: Algorithms; Ethnic groups; Hispanic Americans; United States Department of Veterans Affairs

INTRODUCTION

A social and political construct, Hispanic ethnicity was first introduced as a term in the 1970 United States (U.S.) Census[1] and it continues to be used in national and state reporting systems.[2] The U.S. Office of Management and Budget (OMB) defines "Hispanic" as a person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race.

The Hispanic population represents the largest minority group in the U.S.[3] and bears a disproportionate burden of chronic conditions compared with non-Hispanic populations.[4,5] Collection of accurate ethnicity data is critical to evaluate disparities in healthcare utilization, burden of disease, and outcomes.[6,7] However, this information is often inconsistently documented in electronical health records (EHRs),[8,9] and Hispanic patients may fear to disclose their ethnicity due to discrimination.[10] Therefore, restricting analyses to available data may lead to biased conclusions.[11-13]

Several efforts have been made to improve the completeness and accuracy of Hispanic ethnicity in health databases. The North American Association of Central Cancer Registries (NAACCR) developed the Hispanic Identification Algorithm (NHIA) to enhance the identification of Hispanic patients and estimate cancer incidence and mortality rates.[14,15] The Research Triangle Institute (RTI) developed an algorithm for the Centers for Medicare and Medicaid Services to improve the accuracy of race and ethnicity data[16,17] and is currently used in reports on health disparities.[18] Recent efforts have explored diverse approaches such as Bayesian techniques, machine learning, and genetic ancestry data to address unreliable information.[19-22]

Despite consistent growth of the Hispanic Veteran population,[23] similar work to improve the completeness and accuracy of ethnicity data in the Veterans Affairs (VA) has not occurred. Therefore, we sought to develop an algorithm to supplement Hispanic ethnicity information and evaluate differences in demographic characteristics and prevalence of disease.

METHODS

Data Source

We used data from the VA, which comprises over 1,200 points of care nationwide, including hospital, medical centers, and community outpatient clinics. All care is recorded in the VA Corporate Data Warehouse, a national repository containing clinical and administrative data. This study was approved by the Institutional Review Boards of the VA Connecticut Healthcare System and Yale University.

Study Datasets

We first included all VA patients who completed the 2012 Veterans Aging Cohort Study (VACS) survey (Table 1).[24] Next, we linked three data sources to obtain ethnicity data: (1) 2012 VACS survey containing self-reported ethnicity which constituted the reference standard; (2) VA EHR containing recorded ethnicity; and (3) Medicare Beneficiary Summary File containing the RTI race variable.[16] We then obtained data for all patients who had at least one inpatient or outpatient visit in the VA for each fiscal year (FY) from October 1, 2017, to September 30, 2019 (FY 2018-2019). Relevant variables were extracted and defined as detailed on eTable 1 and 2 in the Supplement.

Data Analysis

Using birthplace and surname, we developed an algorithm to enhance identification of Hispanic ethnicity. Specific steps were as follows: (1) Identify birthplace: If birthplace was from a country associated with a high probability of Hispanic ethnicity, then patient was considered Hispanic; and (2) Link surnames with 2010 Census Surname List: If surname was considered

"heavily Hispanic" (>75%),[14] then patient was considered Hispanic. We did not include patients born in the Philippines, as the country has a high prevalence of Spanish surnames but low probability of Hispanic ethnicity.

We then used self-reported Hispanic ethnicity from the 2012 VACS survey as the reference standard to determine sensitivity and specificity. Sensitivity indicated how a method correctly identified Hispanic individuals and was calculated as [True positive/ (True positive + False negative) x 100]. Specificity indicated how a method correctly identified non-Hispanic individuals and was calculated as [True negative/ (True negative + False positive) x 100]. Next, we combined EHR and algorithm ethnicity, identifying patients as Hispanic if either source indicated this, hereafter referred to as combined. Lastly, we compared EHR-recorded, algorithm-derived, combined ethnicity, and the RTI race variable with self-report for all survey respondents.

Using FY 2018-2019, we compared demographic characteristics of Hispanic patients according to EHR, algorithm, and combined ethnicity. Next, we applied direct standardization to calculate age- and sex-adjusted prevalence of selected conditions and compared them among identification methods.[25,26] The reference population was the entire FY 2018-2019 sample, using 10-year age groups, combining those 80 years and older. Age was set according to the age on September 30, 2018, the mid-point of the dataset. Finally, we compared prevalence among Hispanic patients by nativity status and subgroups.

All analyses were conducted using SAS statistical software version 9.4 (SAS Institute). *P* values were 2-sided, and statistical significance was $P < 0.05$.

RESULTS

Algorithm Evaluation

A total of 3,810 patients responded the 2012 VACS survey and had available ethnicity information. Using self-reported Hispanic ethnicity from the survey as the validation standard, the sensitivity of EHR ethnicity was 59% among those self-identified as Hispanic, and the specificity was 98% among those self-identified as non-Hispanic (Table 2). Notably, the algorithm yielded a sensitivity of 80% compared to 75% for the RTI race variable. Combining Hispanic patients identified by either EHR or algorithm achieved the highest sensitivity (84%).

Algorithm Implementation

A total of 7,156,670 patients received care in the VA FY 2018-2019 (Table 1). Of these, the algorithm identified 74,029 additional Hispanic individuals, a 16% increase. The algorithm reclassified as Hispanic 1.6% of non-Hispanic patients, 6.9% of patients with unknown ethnicity, and 63.2% of patients with conflicting ethnicity information (eTable 3 in the Supplement).

Compared to EHR, Hispanic patients identified by the algorithm were more likely to be older (59 years [IQR, 41-71] vs. 57 years [IQR, 39-71]), foreign-born (36.2% vs. 33.2%), and had a race other than White (31.8% vs. 27.4%) (Table 3).

The prevalence of chronic conditions in Hispanic patients was similar between EHR and algorithm **(Table 4)**. Compared to non-Hispanic Whites, Hispanic patients identified by the algorithm had a higher prevalence of diabetes (30.2% vs. 22.2%), chronic live disease (7.2% vs. 5.0%), hepatocellular carcinoma (29.3 vs. 14.6 per 10,000), gastric cancer (8.4 vs. 4.8 per

10,000), and HIV (50.1 vs. 25.5 per 10,000). Foreign-born Hispanic patients have a higher prevalence of hypertension, gastric cancer, and HIV than U.S.-born Hispanic and non-Hispanic White patients (Figure 1). Among Hispanic subgroups, the prevalence of diabetes, hypertension, cerebrovascular disease, and chronic liver disease was highest among Puerto Ricans compared to other Hispanic subgroups and non-Hispanic White patients (eTable 4 in the Supplement).

DISCUSSION

We developed and validated a simple algorithm based on country of birth and surname to supplement EHR ethnicity in the largest integrated U.S. healthcare system. The application of our method in the VA suggests that sole reliance on EHR data is problematic. First, the sensitivity of EHR ethnicity is only 59% compared to 80% for the algorithm. Second, EHR data underestimated the number of Hispanic patients. Third, Hispanic patients identified by the algorithm differed in age, race, and nativity status. Fourth, although the prevalence of sentinel health conditions was similar between EHR and algorithm ethnicity, our approach revealed substantial differences among Hispanics by nativity status and subgroups.

Hispanic ethnicity, as well as other racialized groups, is a social construct derived in the last century for political reasons, which groups people based on skin color and, to some extent, shared heritage but with limited biological significance.[27] Social privileges are allotted to some at the expense of others, provides foundation for discrimination and racism, and shapes lived experiences and access to resources that determine disparities in morbidity and mortality.[28,29]A recent example occurred for Hispanic individuals during the COVID-19 pandemic.[30]

The U.S. military is not immune to this social phenomenon, and minority Veterans, including Hispanics, have higher rates of psychiatric disorders due to stress associated with racism, racial bias on officer promotion rates, and decreased access to the VA healthcare system.[31-33] Therefore, the ability to measure race and ethnicity is essential for exposing disparities and inequities in health and healthcare stemming from social stratification.

The present study builds on efforts made to improve identification of Hispanic ethnicity in health databases (eTable 5 in the Supplement),[14,16] which rely on complex methods, bringing challenges in code maintenance, programming language transitions, and communication.[19-21,34]

Our easily reproducible algorithm based solely on surname and country of birth, demonstrated higher sensitivity and specificity than EHR ethnicity. Furthermore, the performance of our approach was superior to the RTI race variable from the Medicare administrative data.

We revealed that EHR ethnicity underestimates the number of Hispanic patients as evidenced on previous studies.[17,35] Inaccurate and missing ethnicity data is frequently non-random and may be attributed to perceived discrimination for minority patients in health care,[33] including the VA.[36] For instance, Hispanic patients are significantly less comfortable disclosing their ethnicity, very concerned it could be used to discriminate against themselves, and less likely to go to a health facility that collects ethnicity information.[10,37] We also demonstrated that Hispanic patients identified by our algorithm were more frequently older, foreign-born, and had a race other than White. These findings are consistent with a prior study demonstrating significant differences in demographic characteristics between patients with missing race and ethnicity from those with structured EHR data.[38] Given that fears of discrimination may play a critical role, researchers and policy makers need to be aware of the extent of unreliable ethnicity information to evaluate for potential biased findings resulting in health policy inaction or inefficient allocation of resources.

Although the prevalence of conditions in Hispanic patients was similar between EHR and algorithm, our study showed remarkable differences compared to non-Hispanic White patients. Consistent with prior reports,[39] Hispanic individuals were more likely to have diabetes, chronic liver disease, hepatocellular carcinoma, gastric cancer, and HIV. Most studies evaluating health and health care disparities have not included Hispanic Veterans.[40,41] This is particularly concerning given the higher prevalence of conditions such as chronic liver

disease among Hispanic patients, and those with hepatocellular carcinoma experience reduced or delayed healthcare utilization.[42] Furthermore, our method provided detailed and comprehensive information of this heterogeneous population, revealing substantial differences in disease burden among Hispanic patients by nativity and subgroups.

Our study has several strengths. First, we used data from the largest integrated U.S. healthcare system which could inform strategies for enhancing Hispanic identification in other healthcare systems. Second, our algorithm is the first to provide detailed data of Hispanic Veterans according to nativity status and country of birth. This could be meaningful for future studies to evaluate if aggregation of Hispanics as a single group masks heterogeneity in health outcomes. Lastly, since data on race and ethnicity is incomplete and inaccurate in the EHR, we believe the output of our algorithm could be readily incorporated as a variable in the VA data. For example, Medicare administrative data already incorporates the RTI race variable to facilitate both a better understanding of, and more effective interventions on, ethnic differences and disparities in health care access and care quality.

Our study has limitations. First, Veterans receiving care in the VA are predominantly men, older, and have higher prevalence of comorbidities compared to the general U.S. population. Second, Hispanic identity is multidimensional and multifaceted, and evolves over time. However, we followed the definition of ethnicity from the OMB standards.[2] Third, our algorithm did not use maiden name, as this was not available, leading to potential misclassification of women. Moreover, the surname list is from the 2010 Census, however, this was the only list available. Fourth, there is a risk for over-ascertainment of Hispanic ethnicity when combining EHR and algorithm ethnicity, as well as misclassification when considering Hispanic surnames if over 75% identify as Hispanic. Nevertheless, we followed

methods applied in previous algorithms and were able to improve the sensitivity without loss of specificity.[14,16] Finally, the prevalence of conditions does not consider missed cases such as those who did not show up for care or whose diagnosis was not recorded.

In conclusion, we developed and validated an algorithm to supplement ethnicity information in the largest integrated U.S. healthcare system. Our method enabled clearer understanding of demographic characteristics and burden of disease among Hispanic Veterans. Complete and accurate ethnicity data is crucial to understand health disparities in this population given the downstream effects on the identification of risk factors, outcomes, and prognosis.

Acknowledgments

This work uses data provided by patients and collected by the VA as part of their care and support. The views and opinions expressed in this paper are those of the authors and do not necessarily represent those of the Department of Veteran Affairs or the United States Government.

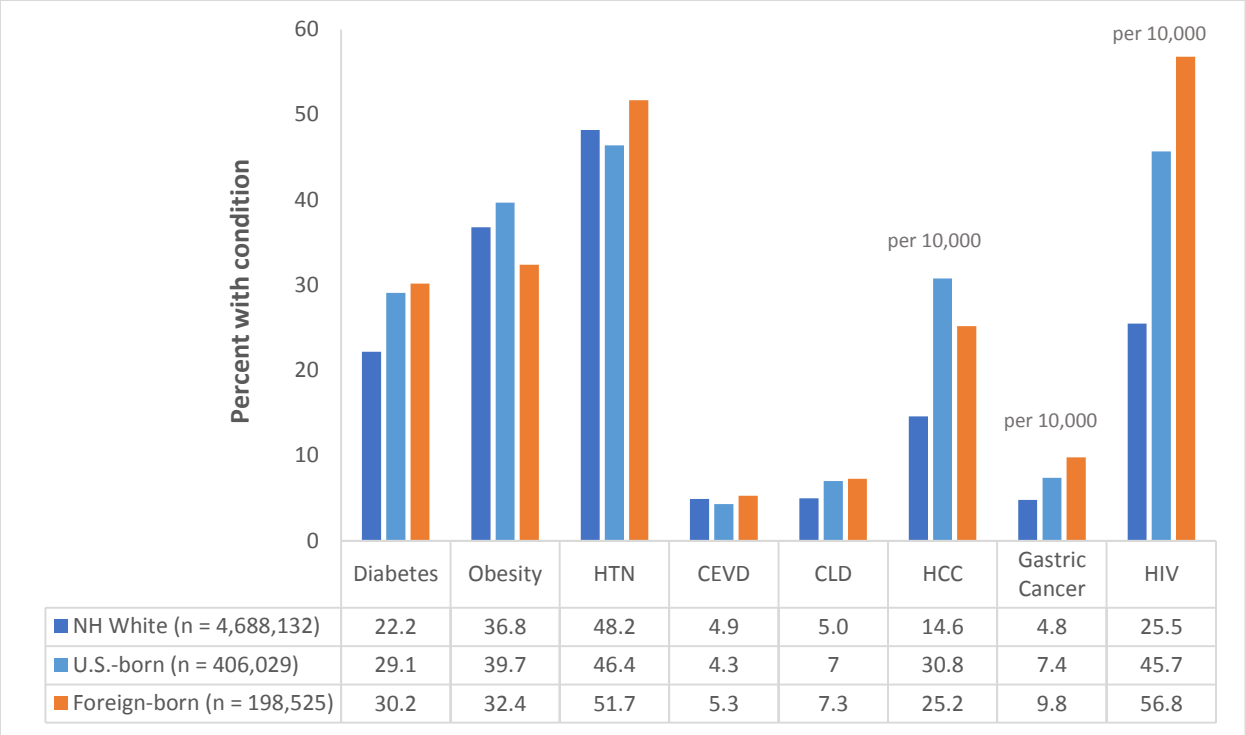
REFERENCES

1. U.S. Census Bureau. 1970 Census of Population, Subject Reports: Persons of Spanish Origin. Updated October 8, 2021. Accessed November 10, 2021, https://www.census.gov/library/publications/1973/dec/pc-2-1c.html
2. Executive Office of the President, Office of Management and Budget. *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity*. Vol. 62. 1997. *Federal Register*. October 30, 1997. https://www.whitehouse.gov/wp-content/uploads/2017/11/Revisions-to-the-Standards-for-the-Classification-of-Federal-Data-on-Race-and-Ethnicity-October30-1997.pdf
3. U.S. Census Bureau. Annual Estimates of the Resident Population by Sex, Age, Race, and Hispanic Origin for the United States and States: April 1, 2010 to July 1, 2018. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk
4. Escarce JJ, Morales LS, Rumbaut RG. The Health Status and Health Behaviors of Hispanics. In: National Research Council (US) Panel on Hispanics in the United States. In: Tienda M, Mitchell F, eds. *Hispanic and the Future of America*. National Academies Press (US); 2006.
5. U.S. Department of Health and Human Services Offices of Minority Health. Cancer and Hispanic Americans. Accessed September 6, 2021, https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4&lvlid=61
6. Waldo DR. Accuracy and Bias of Race/Ethnicity Codes in the Medicare Enrollment Database. *Health Care Financ Rev*. Winter 2004;26(2):61-72.
7. Magana Lopez M, Bevans M, Wehrlen L, Yang L, Wallen GR. Discrepancies in Race and Ethnicity Documentation: a Potential Barrier in Identifying Racial and Ethnic Disparities. *J Racial Ethn Health Disparities*. Sep 8 2016;doi:10.1007/s40615-016-0283-3
8. Ng JH, Ye F, Ward LM, Haffer SC, Scholle SH. Data On Race, Ethnicity, And Language Largely Incomplete For Managed Care Plan Members. *Health Affairs*. Mar 1 2017;36(3):548-552. doi:10.1377/hlthaff.2016.1044
9. Polubriaginof FCG, Ryan P, Salmasian H, et al. Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc*. Aug 1 2019;26(8-9):730-736. doi:10.1093/jamia/ocz113
10. Baker DW, Hasnain-Wynia R, Kandula NR, Thompson JA, Brown ER. Attitudes toward health care providers, collecting information about patients' race, ethnicity, and language. *Med Care*. Nov 2007;45(11):1034-42. doi:10.1097/MLR.0b013e318127148f
11. Labgold K, Hamid S, Shah S, et al. Estimating the Unknown: Greater Racial and Ethnic Disparities in COVID-19 Burden After Accounting for Missing Race and Ethnicity Data. *Epidemiology*. Mar 1 2021;32(2):157-161. doi:10.1097/EDE.0000000000001314
12. Kressin NR, Chang B-H, Hendricks A, Kazis LE. Agreement between administrative data and patients' self-reports of race/ethnicity. *Am J Public Health*. 2003;93(10):1734-9. doi:10.2105/ajph.93.10.1734
13. Bierman AS, Lurie N, Collins KS, Eisenberg JM. Addressing Racial And Ethnic Barriers To Effective Health Care: The Need For Better Data. *Health Affairs*. 2002;21(3):91-102. doi:10.1377/hlthaff.21.3.91
14. NAACCR Race and Ethnicity Work Group. NAACCR Guideline for Enhancing Hispanic/Latino Identification: Revised NAACCR Hispanic/Latino Identification Algorithm

[NHIA v2.2.1]. Springfield (IL): North American Association of Central Cancer Registries. September 2011.
15. National Cancer Institute. Surveillance, Epidemiology, and End Results Program: Documentation for SEER Data, Race Recode Changes. Accessed September 6, 2021, 2021. https://seer.cancer.gov/seerstat/variables/seer/race_ethnicity/index.html
16. Bonito AJ BC, Eicheldinger C, Carpenter L. Creation of New Race-Ethnicity Codes and Socioeconomic Status (SES) Indicators for Medicare Beneficiaries. Final Report, Sub-Task 2. (Prepared by RTI International for the Centers for Medicare and Medicaid Services through an interagency agreement with the Agency for Healthcare Research and Policy, under Contract No. 500-00-0024, Task No. 21) AHRQ Publication No. 08-0029-EF. Rockville, MD, Agency for Healthcare Research and Quality. January 2008.
17. Eicheldinger C, Bonito A. More accurate racial and ethnic codes for Medicare administrative data. *Health care financing review*. 2008;29(3):27-42.
18. Centers of Medicare & Medicaid Services Office of Minority Health. The Mapping Medicare Disparities Tool: Technical Documentation. https://www.cms.gov/About-CMS/Agency-Information/OMH/Downloads/Mapping-Technical-Documentation.pdf
19. Hernandez SE, Sylling PW, Mor MK, et al. Developing an Algorithm for Combining Race and Ethnicity Data Sources in the Veterans Health Administration. *Mil Med*. Mar 2 2020;185(3-4):e495-e500. doi:10.1093/milmed/usz322
20. Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res*. Oct 2008;43(5 Pt 1):1722-36. doi:10.1111/j.1475-6773.2008.00854.x
21. Kim JS, Gao X, Rzhetsky A. RIDDLE: Race and ethnicity Imputation from Disease history with Deep LEarning. *PLoS Comput Biol*. Apr 2018;14(4):e1006106. doi:10.1371/journal.pcbi.1006106
22. Fang H, Hui Q, Lynch J, et al. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am J Hum Genet*. Oct 3 2019;105(4):763-772. doi:10.1016/j.ajhg.2019.08.012
23. National Center for Veterans Analysis and Statistics. Veteran Population Projection Model, Population Tables, Table 3L: VetPop2020 Living Veterans by Race/Ethnicity, Gender, 2020-2050. Updated 20220907. 2022. https://www.va.gov/vetdata/veteran_population.asp
24. Justice AC, Dombrowski E, Conigliaro J, et al. Veterans Aging Cohort Study (VACS): Overview and description. *Med Care*. Aug 2006;44(8 Suppl 2):S13-24. doi:10.1097/01.mlr.0000223741.02074.66
25. Park LS, Tate JP, Sigel K, et al. Time trends in cancer incidence in persons living with HIV/AIDS in the antiretroviral therapy era: 1997-2012. *AIDS*. Jul 17 2016;30(11):1795-806. doi:10.1097/QAD.0000000000001112
26. Gordis L. *Epidemiology*. Fifth edition. ed. Elsevier/Saunders; 2014:pp. 77-85.
27. Borrell LN, Elhawary JR, Fuentes-Afflick E, et al. Race and Genetic Ancestry in Medicine - A Time for Reckoning with Racism. *N Engl J Med*. Feb 4 2021;384(5):474-480. doi:10.1056/NEJMms2029562
28. Link BG, Phelan J. Social conditions as fundamental causes of disease. *J Health Soc Behav*. 1995;Spec No:80-94.
29. Phelan JC, Link BG. Is Racism a Fundamental Cause of Inequalities in Health? *Annual Review of Sociology*. 2015;41(1):311-330. doi:10.1146/annurev-soc-073014-112305

30. Mackey K, Ayers CK, Kondo KK, et al. Racial and Ethnic Disparities in COVID-19-Related Infections, Hospitalizations, and Deaths : A Systematic Review. *Ann Intern Med*. Mar 2021;174(3):362-373. doi:10.7326/M20-6306
31. Sohn L, Harada ND. Effects of racial/ethnic discrimination on the health status of minority veterans. *Mil Med*. Apr 2008;173(4):331-8. doi:10.7205/milmed.173.4.331
32. Burk J, Espinoza E. Race Relations Within the US Military. *Annual Review of Sociology*. 2012/08/11 2012;38(1):401-422. doi:10.1146/annurev-soc-071811-145501
33. Glass JE, Williams EC, Oh H. Racial/ethnic discrimination and alcohol use disorder severity among United States adults. *Drug Alcohol Depend*. Nov 1 2020;216:108203. doi:10.1016/j.drugalcdep.2020.108203
34. Boscoe FP, Schymura MJ, Zhang X, Kramer RA. Heuristic algorithms for assigning Hispanic ethnicity. *PLoS One*. 2013;8(2):e55689. doi:10.1371/journal.pone.0055689
35. Jarrin OF, Nyandege AN, Grafova IB, Dong X, Lin H. Validity of Race and Ethnicity Codes in Medicare Administrative Data Compared With Gold-standard Self-reported Race Collected During Routine Home Health Care Visits. *Med Care*. Jan 2020;58(1):e1-e8. doi:10.1097/MLR.0000000000001216
36. Hausmann LR, Jeong K, Bost JE, Kressin NR, Ibrahim SA. Perceived racial discrimination in health care: a comparison of Veterans Affairs and other patients. *Am J Public Health*. Nov 2009;99 Suppl 3:S718-24. doi:10.2105/AJPH.2008.150730
37. Baker DW, Cameron KA, Feinglass J, et al. Patients' attitudes toward health care providers collecting information about their race and ethnicity. *J Gen Intern Med*. Oct 2005;20(10):895-900. doi:10.1111/j.1525-1497.2005.0195.x
38. Sholle ET, Pinheiro LC, Adekkanattu P, et al. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *J Am Med Inform Assoc*. Aug 1 2019;26(8-9):722-729. doi:10.1093/jamia/ocz040
39. U.S. Department of Health and Human Services Offices of Minority Health. Hispanic/Latino Americans. Accessed February 19, 2022, https://www.minorityhealth.hhs.gov/omh/browse.aspx?lvl=3&lvlid=64
40. Kondo K, Low A, Everson T, et al. Health Disparities in Veterans: A Map of the Evidence. *Med Care*. Sep 2017;55 Suppl 9 Suppl 2:S9-S15. doi:10.1097/MLR.0000000000000756
41. Peterson K, Anderson J, Boundy E, Ferguson L, McCleery E, Waldrip K. Mortality Disparities in Racial/Ethnic Minority Groups in the Veterans Health Administration: An Evidence Review and Map. *Am J Public Health*. Mar 2018;108(3):e1-e11. doi:10.2105/AJPH.2017.304246
42. Rich NE, Hester C, Odewole M, et al. Racial and Ethnic Differences in Presentation and Outcomes of Hepatocellular Carcinoma. *Clin Gastroenterol Hepatol*. Feb 2019;17(3):551-559 e1. doi:10.1016/j.cgh.2018.05.039

Figure 1.



	Diabetes	Obesity	HTN	CEVD	CLD	HCC	Gastric Cancer	HIV
NH White (n = 4,688,132)	22.2	36.8	48.2	4.9	5.0	14.6	4.8	25.5
U.S.-born (n = 406,029)	29.1	39.7	46.4	4.3	7	30.8	7.4	45.7
Foreign-born (n = 198,525)	30.2	32.4	51.7	5.3	7.3	25.2	9.8	56.8

Abbreviations: CLD, chronic liver disease; CEVD, cerebrovascular disease; HCC, hepatocellular carcinoma; HIV, human immunodeficiency virus; HTN, hypertension; NH, non-Hispanic; U.S., United States.
*The prevalence of conditions does not consider missed cases such as those who did not show up for care or whose diagnosis was not recorded.

Table 1.

Variable	2012 VACS Survey (n = 4,409)	2018-2019 VA EHR (n = 7,156,670)
Sex		
Male	95.6	91.3
Female	4.4	8.7
Age, median (IQR), y	58 (52-63)	66 (51-74)
Age groups, y		
18-29	1.4	3.0
30-39	3.8	10.1
40-49	12.1	9.8
50-59	43.8	14.2
60-69	32.6	22.4
70-79	5.3	24.1
≥ 80	1.1	16.3
Ethnicity		
Hispanic	9.8	6.3
Non-Hispanic	76.6	88.0
Unknown/Conflicting	13.6	5.8
Race		
White	23.0	71.3
Black	69.3	16.3
AAPI	0.6	1.8
AIAN	0.0	0.7
Mixed	1.1	2.2
Unknown	6.0	7.7
HIV status		
Positive	52.9	0.4
Negative	47.1	99.6

Abbreviations: AAPI, Asian Americans and Pacific Islanders; AIAN, American Indian and Alaska Native; EHR, electronic health record; HIV, human immunodeficiency virus; IQR, interquartile range; NOS, not otherwise identified; U.S., United States; VA, Veterans Affairs; VACS, Veterans Aging Cohort Study.

[a] Includes American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and U.S. Virgin Islands.

Table 2.

	2012 VACS survey[a]			Accuracy measures[b]	
Method	**Hispanic**	**Non-Hispanic**	**Total**	**Sensitivity**	**Specificity**
VA EHR					
Hispanic	254	9	263	59	98
Non-Hispanic	93	3,298	3,391		
Unknown/Inconsistent	84	72	156		
Total	431	3,379	3,810		
Algorithm[c]					
Hispanic	343	20	363	80	99
Non-Hispanic	88	3,359	3,447		
Total	431	3,379	3,810		
Combined[d]					
Hispanic	361	26	387	84	99
Non-Hispanic	70	3,353	3,423		
Total	431	3,379	3,810		
RTI race[e]					
Hispanic	151	6	157	75	100
Non-Hispanic	51	1,687	1,738		
Total	202	1,693	1,895		

Abbreviations: EHR, electronic health record; RTI, research triangle institute, VA, Veterans Affairs; VACS, Veterans Aging Cohort Study.

[a] 2012 VACS survey has 599 (13.6%) individuals with missing ethnicity

[b] Accuracy measures: Sensitivity [True positive/ (True positive + False negative) x 100] and Specificity [True negative/ (True negative + False positive) x 100].

[c] Algorithm is based on surname and country of birth.

[d] Combined ethnicity is based on EHR and algorithm ethnicity.

[e] Imputation algorithm from the Centers for Medicare & Medicaid Services. Only 1,895 individuals merged between Medicare and 2012 VACS survey data.

Table 3.

Variable	Identified as Hispanic via EHR (n = 447,405)	Algorithm[a] (n = 521,434)	Combined[b] (n = 604,554)
Sex			
Male	89.4	89.9	88.3
Female	10.6	10.1	11.7
Age, median (IQR), y	57 (39-71)	59 (41-71)	58 (40-71)
Age group, y			
18-29	6.5	5.5	6.1
30-39	19.6	17.3	18.3
40-49	14.1	13.5	13.8
50-59	13.8	14.2	14.2
60-69	18.1	19.8	18.9
70-79	14.8	16.4	15.5
≥ 80	13.2	13.4	13.2
Race			
White	72.1	68.2	68.0
Black	3.9	3.6	4.4
AAPI	2.1	2.7	2.9
AIAN	1.5	1.5	1.6
Mixed	3.4	3.6	3.6
Unknown	17.0	20.4	19.5
Nativity			
U.S.-born	66.8	63.8	67.2
Foreign-born	33.2	36.2	32.8
Country of birth			
Puerto Rico	56.5	53.6	50.9
Mexico	14.0	13.9	13.2
Central America	5.2	7.2	6.8
South America	4.9	5.5	5.2
Cuba/DR	4.8	5.1	4.9
Spain	0.2	0.6	0.6
NOS	14.3	14.1	18.4
Geographic region			
Northeast	7.5	7.7	7.7
South	40.6	41.9	41.7
Midwest	6.2	6.3	6.7
West	34.7	34.0	34.7
U.S. territories[c]	11.1	10.1	9.1
Level of rurality			
Urban	85.5	84.9	84.5
Suburban	6.0	6.1	6.3
Rural	8.5	8.9	9.2

Abbreviations: AAPI, Asian Americans and Pacific Islanders; AIAN, American Indian and Alaska Native; DR, Dominican Republic; EHR, electronic health record; FY, fiscal year; IQR, interquartile range; NOS, not otherwise identified; U.S., United States.

[a] Algorithm is based on surname and country of birth.

[b] Combined ethnicity is based on EHR and algorithm ethnicity.

[c] Includes American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and U.S. Virgin Islands.

Table 4.

	Age- and sex-adjusted prevalence, %			
		Identified as Hispanic via		
Condition[a]	NH White (n = 4,688,132)	EHR (n =447,405)	Algorithm[b] (n = 521,434)	Combined[c] (n = 604,554)
Diabetes	22.2	29.9	30.2	29.6
Obesity	36.8	37.5	37.1	37.1
Hypertension	48.2	48.8	49.0	48.4
Cerebrovascular disease	4.9	4.7	4.8	4.7
Chronic liver disease	5.0	7.2	7.2	7.1
Hepatocellular carcinoma	14.6	28.5	29.3	28.9
Gastric Cancer	4.8	8.5	8.4	8.3
HIV	25.5	48.9	50.1	49.6

Abbreviations: EHR, electronic health record; FY, fiscal year; HIV, human immunodeficiency virus; NH, non-Hispanic.

[a] The prevalence of conditions does not consider missed cases such as those who did not show up for care at the VA or whose diagnosis was not recorded

[b] Algorithm is based on surname and country of birth.

[c] Combined ethnicity is based on EHR and algorithm ethnicity.