Validity of race and ethnicity codes in Medicare administrative data compared to gold-standard self-reported race collected during routine home health care visits

Olga F. Jarrín, PhD, RN,
School of Nursing, Division of Nursing Science, Institute for Health, Health Care Policy, and Aging Research, Rutgers, The State University of New Jersey, 112 Paterson Street, New Brunswick, NJ 08901

Abner N. Nyandege, PhD,
Institute for Health, Health Care Policy, and Aging Research, Rutgers, The State University of New Jersey

Irina B. Grafova, PhD,
School of Public Health, Rutgers, The State University of New Jersey, 683 Hoes Lane West, Room 310, New Brunswick, NJ 08901

XinQi Dong, MD, MPH,
Institute for Health, Health Care Policy, and Aging Research, Rutgers, The State University of New Jersey

Haiqun Lin, MD, PhD
School of Nursing, Division of Nursing Science, Rutgers, The State University of New Jersey, 180 University Avenue, Newark, NJ 07102

Abstract

Background: Misclassification of Medicare beneficiaries' race/ethnicity in administrative data sources is frequently overlooked and a limitation in health disparities research.

Objective: To compare the validity of two race/ethnicity variables found in Medicare administrative data (EDB and RTI race) against a gold-standard source also available in the Medicare data warehouse: the self-reported race/ethnicity variable on the home health Outcome and Assessment Information Set (OASIS).

Subjects: Medicare beneficiaries over the age of 18 who received home health care in 2015 (N = 4,243,090).

Measures: Percent agreement, sensitivity, specificity, positive predictive value (PPV), and Cohen's kappa coefficient.

Results: The EDB and RTI race variable have high validity for Black race and low validity for American Indian/Alaskan Native race. While the RTI race variable has better validity than the

olga.jarrin@rutgers.edu.

EDB race variable for other races, kappa values suggest room for future improvements in classification of Whites (0.90), Hispanics (0.87), Asian/Pacific Islanders (0.77), and American Indian/Alaskan Natives (0.44).

Discussion: The status quo of using 'good-enough for government' race/ethnicity variables contained in Medicare administrative data for minority health disparities research can be improved through the use of self-reported race/ethnicity data, available in the Medicare data warehouse. Health services and policy researchers should critically examine the source of race/ethnicity variables used in minority health and health disparities research. Future work to improve the accuracy of Medicare beneficiaries' race/ethnicity data should incorporate and augment the self-reported race/ethnicity data contained in assessment and survey data, available within the Medicare data warehouse.

INTRODUCTION

Improving minority health and reducing health dispartities is a national priority.[1,2] Recent attention has been placed on addressing confounding of observational data and the use of sophisticated causal modeling methods in health disparities research.[3] However, monitoring and reducing disparities requires accurate data on race and ethnicity that is not consistently available.[4–6] Administrative data sources of race/ethnicity data are limited with regards to completeness and accuracy, making self-reported data the preferred source and gold standard.[7] Despite this, even when self-reported race/ethnicity data is available, an administrative data source is frequently used in research on disparities in healthcare quality and outcomes.[8–10] The completeness and accuracy of race/ethnicty data is especially problematic for Asian Americans and Pacific Islanders (AAPI), as well as for American Indians and Alaskan Natives (AIAN).[11–13] As a result, incomplete and inaccurate race/ethnicity data limit our understanding of the sources of disparities in healthcare access, quality, and outcomes as well as evaluation of changes in minority health over time.

Administrative data, including insurance plan enrollment and demographic information, is contained in the Medicare Beneficiary Summary File (MBSF). The MBSF contains two separate race variables. The first is from the Medicare enrollment database (EDB), and originates from Social Security Administration records. Prior to 1980, the Social Security Administration (SSA) collected voluntary race data using the categories: white, black, other, and unknown (for people who did not respond). "A further limitation in the racial and ethnic data contained in Medicare beneficiary files is that when the Center for Medicare and Medicaid Services (CMS) obtains the enrollee information from the SSA master beneficiary record, it receives information only on the retiree, not the retiree's spouse. Instead, the race of the beneficiary is simply assigned to the spouse."[14] CMS has made multiple efforts to fill in missing data including a postcard survey of people with Hispanic surname or country of birth, and use of race/ethnicity data from Medicaid for dual-eleigible beneficiaries from 32 states. However, despite these efforts, the EDB race variable is known to severely undercount Hispanics, Asian Americans/Pacific Islanders, and American Indians/Alaskan Natives.[15] Due to these limitations, analyses using race/ethnicity data from the enrollment file (EDB) are generally restricted to the identification of differences between black and white patient populations.[9,16] The second race variable was created a decade ago by

researchers at the Research Triangle Institute (RTI) to improve classification of Hispanics and Asians/Pacific Islanders.[17,18] The RTI race imputation algorithm utilizes lists of Hispanic and Asian/Pacific Islander names from the U.S. Census, and simple geography (residence in Puerto Rico or Hawaii) to improve on the EDB race code.[17] The RTI race variable is used by the Centers for Medicare & Medicaid Services' in reports on health disparities in the Medicare population and in studies which include focus on Hispanic and Asian/Pacific Islander populations.[19,20]

In contrast to administrative data sources, national surveys of Medicare beneficiaries include self-reported race and ethnicity. Examples of survey datasets that contain self-reported race/ethnicity include the Medical Expenditure Panel Survey (MEPS), the Medicare Current Beneficiary Survey (MCBS), and the Health and Retirement Survey (HRS). Additionally, the Consumer Assessment of Healthcare Providers and Systems (CAHPS) patient experience datasets contain self-reported race ethnicity data. Finally, self-reported race/ethnicity data is collected as part of post-acute and long-term care assessments including the Outcome and Assessment Information Set (OASIS) used in home health care (the gold-standard in this study), the Minimum Dataset (MDS) used in nursing homes, the Inpatient Rehabilitation Facility-Patient Assessment Instrument (IRF-PAI), and the Medicare Health Outcomes Survey (HOS) used in Programs of All-Inclusive Care of the Elderly and with a random sample of Medicare Advantage plan subscribers.

While patient experience survey data (CAHPS) has been used to validate race/ethnicity variables contained in administrative sources, the use of self-reported race/ethnicity data collected as a routine part of healthcare delivery has received less attention. The main objective of this analysis is to compare the agreement and accuracy of two sources of race and ethnicity information contained in the Medicare data warehouse: 1) the Enrollment Database (EDB) race variable which originates from Social Security Administration data; 2) the Research Triangle Institute (RTI) race variable imputed from name and geography; with a gold-standard: the self-reported race and ethnicity data collected by Registered Nurses and Physical Therapists during routine home health care assessments as part of the Outcome and Assessment Information Set (OASIS).[21] For added context, the accuracy and agreement measures are stratified by sex, patterns of misclassification errors are explored, and we compare our findings with earlier studies using survey data as the gold standard.

METHODS

Data Source and Patient Population

The study population included all Medicare beneficiaries, 18 years and older, who received home health care in 2015 (4,243,090 people). Two data sources containing three race/ethnicity variables for our sample of Medicare beneficiaries were linked using the unique Chronic Conditions Warehouse (CCW) beneficiary identification number for the entire study population: The 2015 Medicare Beneficiary Summary File (MBSF) containing the Enrollment Database (EDB) race variable and Research Triangle Institute (RTI) race variable; and the 2015 Outcome and Assessment Information Set (OASIS) containing the 'gold-standard' self-reported race/ethnicity for all home health care patients. All three race variables (EDB, RTI, OASIS) were available for the entire study population.

During the initial home health care visit by a registered nurse or licensed physical therapist, as part of the standardized OASIS assessment, race/ethnicity data are obtained by self-report (a caregiver may answer if the patient is unable) and allows for multiple answers to be recorded. The directions for this question include the words "Mark all that apply" and the response choices are: 1) American Indian or Alaska Native, 2) Asian, 3) Black or African-American, 4) Hispanic or Laino, 5) Native Hawaiian or Pacific Islander, and 6) White.

For the purposes of this paper, and for consistency with the EDB and RTI race variable categories, beneficiaries who self-identified as either or both 1) Asian and 2) Native Hawaiian or Pacific Islander were classified as Asian American/Pacific Islander (AAPI). The vast majority (99.73%) of home health beneficiaries had only a single race/ethnicity recorded, and we restricted our study to this population. Details of the remaining 11,720 people (0.27% of study population) who identified with two or more racial/ethnic groups are included for the interested reader as a brief Appendix. Our final study sample consisted of 4,231,370 adult Medicare beneficiaries who received home health care in 2015. The study was approved by the Institutional Review Board of [replace with the authors' academic institution].

Statistical Analyses

Datasets were linked at the patient level using the unique beneficiary identification code assigned for this purpose by CMS. For each person, the analytic file contained the three race variables (EDB, RTI, OASIS) which were recoded (so that the value 1 had the same meaning in each dataset) and also dummy coded for calculation of single-race kappa statistic. All analyses were completed by the second author using SAS statistical software (version 9.4) and the first author using Stata 15.0 to ensure reproducibility and confirm final results were error-free. We first assessed the agreement and validity of the EDB race and RTI race variables compared to self-reported race/ethnicity data from the home health Outcome and Assessment Information Set (OASIS). Analyses of sensitivity, specificity, positive predictive value (PPV), and Cohen's kappa coefficient were calculated for the full sample and for each sex separately (Table 1).

Sensitivity = [True Positive/(True Positive + False Negative)] *100

Specificity = [True Negative/(True Negative + False Positive)] *100

Positive Predictive Value = [True Positive/(True Positive + False Positive)] *100

Cohen's kappa statistic is a measure of interrater reliability that takes into account the frequency or rarity of belonging to a different racial/ethnic group. Values range from 1 (complete agreement) to –1 (complete disagreement).[22] As a point of reference, Landis and Koch have suggested a kappa coefficient greater than 0.81 indicates excellent agreement.[23] Both the overall kappa statistic and the individual race kappa statistics were calculated using the entire sample, including cases classified as other/unknown.

In the second set of analyses, the pattern of race/ethnicity misclassifications were explored for both the EDB and RTI race variables compared to OASIS gold-standard. Table 2 includes the raw data used to populate and calculate the overall sample statistics presented in

Table 1. Next, we focus on the subset of cases which were misclassified, highlighting the improvement of the RTI race variable compared to the EDB race variable (Table 3).

In the third set of analyses, differences in race/ethnicity categorization of RTI compared to OASIS race/ethnicity are compared for a subset of beneficiaries with dementia or diabetes (Table 4). We determined dementia or diabetes diagnosis status for our subset study population from the Medicare Beneficiary Summary File (MBSF) chronic conditions warehouse flags. This analysis highlights one aspect of race/ethnicity variable choice on study design and the resulting differences in frequency and prevalence of chronic disease burden within subpopulations.

RESULTS

Agreement and accuracy of Enrollment Data Base (EDB) and Research Triangle Institute (RTI) race variables with self-reported race/ethnicity from OASIS.

Both the EDB and RTI race variables have mutually exclusive categories, meaning that a person who is categorized as white or black is considered to be non-Hispanic. For this reason, in the text and tables, the term "white" refers to non-Hispanic whites, the term "black" refers to non-Hispanic blacks and African Americans, the term "AAPI" refers to non-Hispanic Asians and Pacific Islanders, and the term "AIAN" refers to non-Hispanic American Indians and Alaskan Natives.

In our analyses using OASIS race as the validation standard (shown in Table 1), the sensitivity of EDB and RTI race variables for non-Hispanic whites was high (96.9–97.9), however, the specificity of EDB race was low (79.6) compared to RTI race (95.5). In contrast, among people who self-identified as non-Hispanic black, the EDB and RTI race variables both perform similarly well, with high sensitivity (96.6–97.0) and high specificity (99.2–99.4). Among people who self-identified as Hispanic the original EDB variable had low sensitivity (36.2) but high specificity (99.8). In contrast, the RTI race variable had both good sensitivity (90.8) and high specificity (98.8). Among people who self-identified as non-Hispanic Asian, Hawaiian Native, or other Pacific Islander (AAPI), specificity of both the EDB and RTI race variables was high (99.6–99.8). However, the RTI race variable had better sensitivity (74.7) compared to the EDB race variable (62.6). Finally, among people who self-identified as non-Hispanic American Indian or Alaskan Native (AIAN) the sensitivity of the EDB and RTI race variables was low (43.0–43.2), while the specificity was high (99.8). The EDB classification of AIANs based on tribal membership registration results in fewer than half of people who self-identify as AIAN being correctly classified in Medicare administrative race/ethnicity data.

Sex differences in accuracy and agreement of race/ethnicity variables

The EDB race variable, originating from Social Security Administration records, is slightly more accurate for women compared to men except among AIANs ($k = 0.44$ vs. 0.46). In contrast the RTI race variable, imputed from U.S. Census name lists and residence in Hawaii or Puerto Rico, is less accurate for women compared to men among AAPIs ($k = 0.77$ vs.

0.79), Hispanics ($k = 0.85$ vs. 0.89), and AIANs ($k = 0.44$ vs. 0.46). See Table 1 for all accuracy and agreement statistics stratified by sex.

Patterns of over-classification and misclassification by race/ethnicity variables

The pattern of misclassification errors in the EDB and RTI race variables compared to self-reported race/ethnicity from the OASIS dataset are shown in Table 3. Using the original EDB race variable 190,434 people were misclassified as non-Hispanic white, with the majority (167,495/190,434 = 88%) self-identifying as Hispanic. In contrast, the RTI race variable mistakenly classifies a much smaller number (41,878) of minorities as being non-Hispanic white, with about half being Hispanic (21,941/41,878 = 52.4%). However, the RTI race variable misassigned non-Hispanic whites as Hispanic more than five times as often compared to the original EDB race variable (37,670 vs. 6,695), accounting for 78% of people misassigned as Hispanic by RTI race. Although smaller in number, non-Hispanic whites also comprise 80% of people misassigned by the RTI race variable as black, 77% who are misassigned as AAPI, and 84% of people misassigned as AIAN (Table 3).

Dementia and diabetes frequency and prevalence by race/ethnicity variables

To illustrate the potential impact of race/ethnicity misclassification on estimated size of health disparities and disease prevalence we calculated the number of beneficiaries with dementia and diabetes using each of the three race/ethnicity variables. When comparing the numbers of people with a diagnosis of dementia or diabetes the largest net differences were among the Hispanics, Asians/Pacific Islanders, and American Indians/Alaskan Natives (Table 4). The net difference is important for study designs that draw their sampling frame from administrative data sources.

Using the RTI race variable (compared to OASIS) resulted in an overestimation of the number of Hispanics with dementia by a net difference of 4,283 (4.8%) and diabetes by a net difference of 10,477 (5.4%). In contrast, the EDB race variable underestimated the number of Hispanics with dementia by a net difference of 48,407 (−54.8%) and diabetes by a net difference of 114,003 (−59.0%). However, the EDB race variable also produced falsely high estimates of the prevalence of dementia (34.1%) and diabetes (67.9%) in Hispanics. The RTI and OASIS race variables produced similar estimates of the prevalence of dementia (29.0%–29.6%) and diabetes (63.9%–64.9%) among Hispanics.

Among AAPIs, the number of people with dementia was underestimated by a net difference of 1,853 (−6.4%) using the RTI race variable and by 6,032 (−21.1%) using the EDB race variable. The pattern was similar for diabetes in AAPIs, which was underestimated by a net difference of 4,391 (−8.2%) using the RTI race variable, and 12,113 (−22.6%) using the EDB race variable. When the prevalence of dementia and diabetes were calculated for AAPIs using each of the race/ethnicity variables the pattern was similar to that seen for Hispanics, with EDB race overestimating chronic disease burden. Using the RTI and OASIS variables the prevalence of dementia among Asians/Pacific Islanders was 32.1%–32.6%, and 34.1% using EDB race. For diabetes, the prevalence among AAPIs was 59.9%–60.2% using the RTI and OASIS race variables, and 62.7% using EDB race. Full results shown in Table 4.

DISCUSSION

If we believe self-reported race is truly a "gold standard," we must consider more than overall accuracy (kappa statistic > 0.81) and high specificity. Paraphrasing Statalist (statalist.org) expert Clyde Schechter, let's use a simple example: Lou Gehrig's disease or amyotrophic lateral sclerosis (ALS) is a very rare motor neuron disease. If a "test" to diagnose ALS simply results in everyone "not having it," that test will have high specificity, giving the correct answer for well over 99.9% of the population. However, it is useless to find people who actually have ALS. To be useful, you really need to consider two different measures of validity. 1) Sensitivity: the proportion of people who are positive under the gold standard who are also test positive, and 2) specificity: the proportion of people who are negative under the gold standard who also test negative. Referring to Clyde's phony "test" for ALS, the test would have a specificity of nearly 100% but a sensitivity of 0%. Evaluation of tests or measures for which a gold standard exists usually requires looking at both the sensitivity and specificity.

Similarly, the EDB race variable is nearly useless (despite having high specificity) in identifying Medicare beneficiaries who are Hispanic (sensitivity 36.2, kappa 0.50), AIAN (sensitivity 42.9, kappa 0.44), and AAPI (sensitivity 62.5, 0.71). While the RTI race variable is more useful for identifying Hispanics (sensitivity 90.8, kappa 0.87), it still lacks validity for AIAN (sensitivity 43.0, kappa 0.44) and AAPI (sensitivity 74.7, kappa 0.77).

Consistent with prior studies, we found the EDB and RTI race variables contained in Medicare administrative data undercount Hispanics, AAPIs, and AIANs (summarized in Table 5).[18,24,25] While advances have been made in the Medicare Bayesian Improved Surname and Geocoding (MBISG 2.0) algorithm used to calculate racial and ethnic differences in Healthcare Effectiveness Data and Information Set (HEDIS) measures,[26–28] the accuracy statistics are reported as cross-validated Pearson correlations with self-report, in the form of probabilities, precluding direct comparison with current and prior studies listed in Table 5.

From a methodological standpoint, the choice of race/ethnicity data source is essential at the study design stage for health disparities research. The impact of race/ethnicity variable selection on estimates of disease prevalence is of special concern, as we found in the case of dementia prevalence among Hispanics shown in Table 4. When using the EDB race variable, the prevalence of dementia among Hispanics is 18% higher compared to when the RTI race variable is used, with an absolute difference of just over 5 percentage points. A smaller difference (1.5 percentage points) is seen for AAPIs, with virtually no difference for non-Hispanic whites, blacks, and American Indians/Alaskan Natives (AIANs). Compared to the EDB race variable, if the RTI variable was a "race-specific" anti-dementia drug for Hispanics it would be a blockbuster.

For AAPI populations, our study findings have additional significance. Asian Americans/ Pacific Islanders are the fastest growing population in the U.S., while being the most heterogeneous. Certain AAPI subgroups, such as Filipinos, may be more prone to misclassification using surname-based imputation methods due to the long history of

Spanish colonization in the Philippines. Similarly, the Republic of China (Taiwan) was colonized by the Dutch and Spanish; India was colonized by the Portuguese, Dutch, and British; and Vietnam, Laos, and Cambodia were colonized by the French. In addition, interracial/intercultural marriages frequently result in women changing their last name to that of their husband's family, such that a woman who marries a Filipino-American man might be classified as Hispanic using name-based race algorithms.

While the self-reported race/ethnicity data should always be the first choice, we found the RTI race variable to be very accurate for identifying Hispanics ($k = 0.89$ for males; $k = 0.85$ for females), and non-Hispanic whites ($k = 0.90$) or blacks ($k = 0.96$) of either sex (Table 1). For more granular analyses, and especially research that aims to disentangle race/ethnicity and socio-economic status, a higher level of accuracy may be desired. Researchers who are working with linked administrative and assessment datasets should report racial/ethnic differences based on the self-reported race variable. Reviewers and journal editors should question the source of race/ethnicity data and critically examine the rationale for research which uses the EDB race variable, as it is inappropriate for use beyond studies of black/white disparities. Similarly, studies of nursing home or home health patients should not use the EDB or RTI race variable, as self-reported race collected in the MDS and OASIS assessments is the gold-standard. Finally, future advances in race/ethnicity imputation algorithms at CMS should include and augment self-reported race/ethnicity data from both survey (MCBS, HOS, CAHPS) and assessment (OASIS, MDS) data sources.

This study has several limitations. First, the study population consisted only of Medicare beneficiaries who utilized home health care in calendar year 2015. Second, blacks are slightly overrepresented in the home health care population compared to the full Medicare population (estimated with the RTI race variable). Third, some older adults, especially AAPIs and Hispanics, may retire or seek supportive care outside of the U.S., limiting their access and use of the Medicare home health care benefit, and the generalizability of findings for Medicare beneficiaries living outside the U.S. Finally, AIANs who live on tribal reservations may be underrepresented, in contrast to people who self-identify as American Indian but are not registered tribal members.

In conclusion, administrative datasets are commonly used in reports and studies of minority health and health disparities. Our study highlights the potential for bias and error introduced during the selection of race/ethnicity data source. Our work confirms the advantages of using the RTI race variable compared to EDB race variable. We also show that further reductions in error and bias can be gained by using self-reported race/ethnicity contained in assessment datasets. These findings have important implications for the design of future studies and the interpretation of prior published research on minority health and health disparities. Future work to improve imputation algorithms for Medicare beneficiaries' race/ethnicity should incorporate self-reported race/ethnicity data that is contained in assessment (e.g. MDS, OASIS, IRF-PAI, HIS, HOS) and survey data (CAHPS) to augment existing data sources (EDB, RTI).

Acknowledgements:

APPENDIX

Description and discussion of Medicare beneficiaries who self-identified with two or more race/ethnicity groups during home health assessment (OASIS dataset).

In this supplemental analysis, we focused on the 11,720 beneficiaries who self-identified with two or more racial/ethnic groups during their home health care assessment and were excluded from the main analysis. While this represents a very small fraction (0.28%) of the 4,243,090 Medicare beneficiaries who received home health care in 2015, the number of individuals with multi-racial/ethnic identities is rapidly growing.[29] Of these, 289 people (0.007%)identified with more than two races.

Researchers should be aware of this issue and methods for classifying and modeling individuals who self-report multiple races/ethnicities.[29] For example, of the 4,568 Hispanic individuals who self-reported two races/ethnicities in OASIS, the corresponding RTI race/ ethnicity variable correctly classified 3,194 (70%) as Hispanic but missed/undercounted 1,374 (30%). Additionally, among people who self-identified in OASIS as American Indian or Native Alaskan (AIAN) nearly one-sixth also identified with another race/ethnicity (2,919/18,891).

Appendix Table 1.

Medicare beneficiaries who self-identified with two races/ethnicities

	AAPI	White	Black	AIAN
Hispanic	281	3,586	504	197
AAPI		1,596	259	78
White			2,286	2,318
Black				326

Abbreviations: AAPI = Asian American/Pacific Islanders/Native Hawaiians; AIAN = American Indians/Alaskan Natives.

References

1. Perez-Stable EJ, Collins FS. Science visioning in minority health and health disparities. Am J Pub Health 2019;109:S5. [PubMed: 30699033]

2. Department of Health and Human Services. HHS Action Plan to Reduce Racial and Ethnic Health Disparities: A Nation Free of Disparities in Health and Health Care. Washington, DC; 2015.

3. Jeffries N, Zaslavsky AM, Diez Roux AV, et al. Methodological approaches to understanding causes of health disparities. Am J Pub Health 2019;109:S28–S33. [PubMed: 30699015]

4. Duran DG, Perez-Stable EJ. Science visioning to advance the next generation of health disparities research. Am J Pub Health 2019;109:S11–S3. [PubMed: 30699020]

5. Bilheimer LT, Sisk JE. Collecting adequate data on racial and ethnic disparities in health: The challenges continue. Health Affairs 2008;27:383–91. [PubMed: 18332493]

6. Ng JH, Ye F, Ward LM, Haffer SC, Scholle SH. Data on race, ethnicity, and language largely incomplete for managed care plan members. Health Aff (Millwood) 2017;36:548–52. [PubMed: 28264958]

7. Executive Office of the President Office of Management and Budget, Office of Information and Regulatory Affairs. Revisions to the standards for the classification of federal data on race and ethnicity. Washington, DC: Federal Register; 1997:58782–90.

8. Figueroa JF, Zhou X, Jha AK. Characteristics and spending patterns of persistently high-cost Medicare patients. Health Aff (Millwood) 2019;38:107–14. [PubMed: 30615516]

9. Joynt Maddox KE, Chen LM, Zuckerman R, Epstein AM. Association between race, neighborhood, and Medicaid enrollment and outcomes in Medicare home health care. J Am Geriatr Soc 2018;66:239–46. [PubMed: 28975604]

10. Belanger E, Silver B, Meyers DJ, et al. A retrospective study of administrative data to identify high-need Medicare beneficiaries at risk of dying and being hospitalized. J Gen Intern Med 2019.

11. 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;21:91–102. [PubMed: 12026007]

12. Thomson GE, Mitchell, Willimams MB, eds. Examining the Health Disparities Research Plan of the National Institutes of Health: Unfinished Business. Washington, DC: National Academies Press; 2006.

13. Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. Washington, DC: The National Academies Press; 2009.

14. National Research Council. Eliminating Health Disparities: Measurement and Data Needs. Washington, DC: The National Academies Press; 2004 Pages 68–72. 10.17226/10979.

15. Filice CE, Joynt KE. Examining race and ethnicity information in Medicare administrative data. Med Care 2017;55:e170–e6. [PubMed: 29135782]

16. Li Y, Cai X, Glance LG. Disparities in 30-day rehospitalization rates among Medicare skilled nursing facility residents by race and site of care. Med Care 2015;53:1058–65. [PubMed: 26492217]

17. Eicheldinger C, Bonito A. More accurate racial and ethnic codes for Medicare administrative data. Health Care Financ Rev 2008;29:27–42. [PubMed: 18567241]

18. 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. Rockville, MD: Agency for Healthcare Research and Quality; 2008 January 2008. Report No.: AHRQ Publication No. 08–0029-EF.

19. Lim E, Gandhi K, Davis J, Chen JJ. Prevalence of chronic conditions and multimorbidities in a geographically defined geriatric population with diverse races and ethnicities. J Aging Health 2018;30:421–44. [PubMed: 27913765]

20. Centers for Medicare & Medicaid Services Office of Minority Health. The Mapping Medicare Disparities Tool Technical Documentation, Version 6.0. August 31, 2018 https://www.cms.gov/About-CMS/Agency-Information/OMH/Downloads/Mapping-Technical-Documentation.pdf

21. O'Connor M, Davitt JK. The Outcome and Assessment Information Set (OASIS): a review of validity and reliability. Home Health Care Serv Q 2012;31:267–301. [PubMed: 23216513]

22. Cohen J A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37–46.

23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74. [PubMed: 843571]

24. Waldo DR. Accuracy and bias of race/ethnicity codes in the Medicare Enrollment Database. Health Care Financing Review 2004;26:61–72. [PubMed: 25371985]

25. Zaslavsky AM, Ayanian JZ, Zaborski LB. The validity of race and ethnicity in enrollment data for Medicare beneficiaries. Health Serv Res 2012;47:1300–21. [PubMed: 22515953]

26. Haas A, Elliott MN, Dembosky JW, et al. Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. Health Serv Res 2019;54:13– [PubMed: 30506674]

27. Dembosky JW, Haviland AM, Haas A, et al. Indirect estimation of race/ethnicity for survey respondents who do not report race/ethnicity. Medical Care 2018.

28. Bykov K, Franklin JM, Toscano M, et al. Evaluating cardiovascular health disparities using estimated race/ethnicity: A validation study. Med Care 2015;53:1050–7. [PubMed: 26565524]

29. Klein DJ, Elliott MN, Haviland AM, et al. A comparison of methods for classifying and modeling respondents who endorse multiple racial/ethnic categories. Med Care 2019;57(6):e-34–e41. [PubMed: 30439794]

Table 1.

Accuracy and agreement of Enrollment Data Base (EDB) and Research Triangle Institute (RTI) race/ethnicity classifications compared to the gold-standard self-reported race/ethnicity variable from the home health care Outcome and Assessment Information Set (OASIS): 2015

	True positive	False negative	False positive	True negative	Sensitivity	Specificity	PPV	Kappa
				EDB race overall kappa = 0.79				
White	3,229,553	69,983	190,434	741,400	97.9	79.6	94.4	0.81
Male	1,259,106	32,469	76,300	281,848	97.5	78.7	94.3	0.80
Female	1,970,447	37,514	114,134	459,552	98.1	80.1	94.5	0.82
Black	513,158	15,630	28,042	3,674,540	97.0	99.2	94.8	0.95
Male	188,826	6,204	10,786	1,443,907	96.8	99.3	94.6	0.95
Female	324,332	9,426	17,256	2,230,633	97.2	99.2	95.0	0.96
AAPI	55,837	33,432	10,462	4,131,639	62.6	99.8	84.2	0.71
Male	21,865	13,808	4,462	1,609,588	61.3	99.7	83.1	0.70
Female	33,972	19,624	6,000	2,522,051	63.4	99.8	85.0	0.72
Hispanic	107,753	190,052	9,079	3,924,486	36.2	99.8	92.2	0.50
Male	43,047	77,852	3,406	1,525,418	35.6	99.8	92.7	0.49
Female	64,706	112,200	5,673	2,399,068	36.6	99.8	91.9	0.50
AIAN	6,892	9,080	8,061	4,207,337	43.2	99.8	46.1	0.44
Male	2,846	3,700	3,018	1,640,159	43.5	99.8	48.5	0.46
Female	4,046	5,380	5,043	2,567,178	42.9	99.8	44.5	0.44
				RTI race overall kappa = 0.89				
White	3,196,880	102,656	41,878	889,956	96.9	95.5	98.7	0.90
Male	1,249,568	42,007	14,562	343,586	96.8	95.9	98.9	0.90
Female	1,947,312	60,649	27,316	546,370	97.0	95.2	98.6	0.90
Black	511,064	17,724	21,699	3,680,883	96.7	99.4	95.9	0.96
Male	188,307	6,723	7,953	1,446,740	96.6	99.5	96.0	0.96
Female	322,757	11,001	13,746	2,234,143	96.7	99.4	95.9	0.96
AAPI	66,696	22,573	15,594	4,126,507	74.7	99.6	81.1	0.77
Male	27,207	8,466	6,086	1,607,964	76.3	99.6	81.7	0.79
Female	39,489	14,107	9,508	2,518,543	73.7	99.6	80.6	0.77
Hispanic	270,370	27,435	48,592	3,884,973	90.8	98.8	84.8	0.87

	True positive	False negative	False positive	True negative	Sensitivity	Specificity	PPV	Kappa
Male	111,779	9,120	16,089	1,512,735	92.5	99.0	87.4	0.89
Female	158,591	18,315	32,503	2,372,238	89.7	98.6	83.0	0.85
AIAN	6,872	9,100	7,852	4,207,546	43.0	99.8	46.7	0.45
Male	2,839	3,707	2,928	1,640,249	43.4	99.8	49.2	0.46
Female	4,033	5,393	4,924	2,567,297	42.8	99.8	45.0	0.44

Abbreviations: AAPI = Asian American/Pacific Islanders/Native Hawaiians; AIAN = American Indians/Alaskan Natives.

Table 2.

Data underlying agreement and validity statistics.

EDB race compared to OASIS race — "true positive" cases and "PPV" on diagonal (n (%); diagonal values in bold. Cells off the diagonal are annotated "False positive" (upper/right) and "False negative" (lower/left).)

EDB race	Total	OASIS race White	Black	AAPI	Hispanic	AIAN
Total	4,231,370 (100)	3,299,536 (78.0)	528,788 (12.5)	89,269 (2.1)	297,805 (7.0)	15,972 (0.4)
White	3,419,987 (100)	**3,229,553 (94.4)**	8,376 (0.2)	8,532 (0.3)	167,495 (4.9)	6,031 (0.2)
Black	541,200 (100)	17,524 (3.2)	**513,158 (94.8)**	1,155 (0.2)	8,167 (1.5)	1,196 (0.2)
AAPI	66,299 (100)	7,786 (11.7)	728 (1.1)	**55,837 (84.2)**	1,124 (1.7)	824 (1.2)
Hispanic	116,832 (100)	6,695 (5.7)	1,481 (1.3)	753 (0.6)	**107,753 (92.2)**	150 (0.1)
AIAN	14,953 (100)	6,614 (44.2)	491 (3.3)	439 (2.9)	517 (3.5)	**6,892 (46.1)**
Other/unknown	72,099 (100)	31,364 (43.5)	4,554 (6.3)	22,553 (31.3)	12,749 (17.7)	879 (1.2)

RTI race compared to OASIS race — "true positive" cases and "PPV" on diagonal (n (%); diagonal values in bold)

RTI race	Total	OASIS race White	Black	AAPI	Hispanic	AIAN
Total	4,231,370 (100)	3,299,536 (78.0)	528,788 (12.5)	89,269 (2.1)	297,805 (7.0)	15,972 (0.4)
White	3,238,758 (100)	**3,196,880 (98.7)**	7,682 (0.2)	6,589 (0.2)	21,941 (0.7)	5,666 (0.2)
Black	532,763 (100)	17,265 (3.2)	**511,064 (95.9)**	960 (0.2)	2,298 (0.4)	1,176 (0.2)
AAPI	82,290 (100)	11,948 (14.5)	1,160 (1.4)	**66,696 (81.1)**	1,373 (1.7)	1,113 (1.4)
Hispanic	318,962 (100)	37,670 (11.8)	4,175 (1.3)	6,214 (2.0)	**270,370 (84.8)**	533 (0.2)
AIAN	14,724 (100)	6,570 (44.6)	474 (3.2)	362 (2.5)	446 (3.0)	**6,872 (46.7)**
Other/unknown	43,873 (100)	29,203 (66.6)	4,233 (9.7)	8,448 (19.3)	1,377 (3.1)	612 (1.4)

OASIS race compared to EDB race — "true positive" cases and "sensitivity" on diagonal (n (%); diagonal values in bold. Cells off the diagonal are annotated "False positive" / "False negative".)

OASIS race	Total	EDB race White	Black	AAPI	Hispanic	AIAN	Other/unknown
Total	4,231,370 (100)	3,419,987 (80.8)	541,200 (12.8)	66,299 (1.6)	116,832 (2.8)	14,953 (0.4)	72,099 (1.7)
White	3,299,536 (100)	**3,229,553 (97.9)**	17,524 (0.5)	7,786 (0.2)	6,695 (0.2)	6,614 (0.2)	31,364 (1.0)
Black	528,788 (100)	8,376 (1.6)	**513,158 (97.0)**	728 (0.1)	1,481 (0.3)	491 (0.1)	4,554 (0.9)
AAPI	89,269 (100)	8,532 (9.6)	1,155 (1.3)	**55,837 (62.6)**	753 (0.8)	439 (0.5)	22,553 (25.3)
Hispanic	297,805 (100)	167,495 (56.2)	8,167 (2.7)	1,124 (0.4)	**107,753 (36.2)**	517 (0.2)	12,749 (4.3)
AIAN	15,972 (100)	6,031 (37.8)	1,196 (7.5)	824 (5.2)	150 (0.9)	**6,892 (43.2)**	879 (5.5)

OASIS race compared to RTI race — "true positive" cases and "sensitivity" on diagonal (n (%); diagonal values in bold)

OASIS race	Total	RTI race White	Black	AAPI	Hispanic	AIAN	Other/unknown
Total	4,231,370 (100)	3,238,758 (76.5)	532,763 (12.6)	82,290 (1.9)	318,962 (7.5)	14,724 (0.4)	43,873 (1.0)
White	3,299,536 (100)	**3,196,880 (96.9)**	17,265 (0.5)	11,948 (0.4)	37,670 (1.1)	6,570 (0.2)	29,203 (0.9)
Black	528,788 (100)	7,682 (1.5)	**511,064 (96.7)**	1,160 (0.2)	4,175 (0.8)	474 (0.1)	4,233 (0.8)
AAPI	89,269 (100)	6,589 (7.4)	960 (1.1)	**66,696 (74.7)**	6,214 (7.0)	362 (0.4)	8,448 (9.5)
Hispanic	297,805 (100)	21,941 (7.4)	2,298 (0.8)	1,373 (0.5)	**270,370 (90.8)**	446 (0.1)	1,377 (0.5)
AIAN	15,972 (100)	5,666 (35.5)	1,176 (7.4)	1,113 (7.0)	533 (3.3)	**6,872 (43.0)**	612 (3.8)

Table 3.

Cases misclassified by EDB or RTI race/ethnicity compared to OASIS.

Study population (N = 4,321,370)						
	Misclassified	**OASIS race self-reported as:**				
EDB race misclassified as:	**318,177(7.5%)**	**White**	**Black**	**AAPI**	**Hispanic**	**AIAN**
White	190,434	-	8,376 4.4%	8,532 4.5%	167,495 88.0%	6,031 3.2%
Black	28,042	17,524 62.5%	-	1,155 4.1%	8,167 29.1%	1,196 4.3%
AAPI	10,462	7,786 74.4%	728 7.0%	-	1,124 10.7%	824 7.9%
Hispanic	9,079	6,695 73.7%	1,481 16.3%	753 8.3%	-	150 1.7%
AIAN	8,061	6,614 82.1%	491 6.1%	439 5.5%	517 6.4%	-
Other / unknown	72,099	31,364 43.5%	4,554 6.3%	22,553 6.3%	12,749 17.7%	879 1.2%
RTI race misclassified as:	**179,488 (4.2%)**	**White**	**Black**	**AAPI**	**Hispanic**	**AIAN**
White	41,878	-	7,682 18.3%	6,589 15.7%	21,941 52.4%	5,666 13.5%
Black	21,699	17,265 80.0%	-	960 4.4%	2,298 10.6%	1,176 5.4%
AAPI	15,594	11,948 76.6%	1,160 7.4%	-	1,373 8.8%	1,113 7.1%
Hispanic	48,594	37,670 77.5%	4,175 8.6%	6,214 12.8%	-	533 1.1%
AIAN	7,852	6,570 83.7%	474 6.0%	362 4.6%	446 5.7%	-
Other / unknown	43,873	29,203 66.6%	4,233 9.7%	8,448 19.3%	8,448 19.3%	612 1.4%

Abbreviations: AAPI = Asian American / Pacific Islanders / Native Hawaiians; AIAN = American Indians / Alaskan Natives

Table 4.

Medicare beneficiaries by racial/ethnic group with a dementia or diabetes diagnosis flag in the Medicare Chronic Conditions Warehouse: Number, prevalence, ratio, and net difference of between OASIS, RTI, and EDB race variables.

Alzheimer's Disease and Dementia Sample:	OASIS n (prevalence)	RTI n (prevalence)	Ratio RTI: OASIS	Net Difference RTI - OASIS	EDB n (prevalence)	Ratio EDB: OASIS	Net Difference EDB - OASIS
	Study population with Alzheimer's disease or dementia diagnosis flag in Medicare Chronic Conditions Warehouse (1,195,145/4,231,370 = 28.2%)						
White	932,097 (28.2%)	918,955 (28.4%)	0.99	− 1.4%	966,390 (28.3%)	1.04	+ 3.7%
Black	142,112 (26.9%)	143,526 (26.9%)	1.01	+ 1.0%	145,651 (26.9%)	1.02	+ 2.5%
AAPI	28,653 (31.1%)	26,800 (32.6%)	0.94	− 6.5%	22,621 (34.1%)	0.79	− 21.1%
Hispanic	88,272 (29.6%)	92,555 (29.0%)	1.05	+ 4.9%	39,865 (34.1%)	0.45	− 54.8%
AIAN	4,011 (25.1%)	3,662 (24.9%)	0.91	− 8.7%	3,721 (24.9%)	0.93	− 7.2%
Other / unknown	n/a	9,647 (22.0%)			16,897 (23.4%)		

Diabetes Mellitus Sample:	OASIS n (prevalence)	RTI n (prevalence)	Ratio RTI: OASIS	Net Difference RTI - OASIS	EDB n (prevalence)	Ratio EDB: OASIS	Net Difference EDB - OASIS
	Study population with diabetes mellitus diagnosis flag in Medicare Chronic Conditions Warehouse (2,018,686/4,231,370 = 47.7%)						
White	1,435,215 (43.5%)	1,406,189 (43.4%)	0.98	− 2.0%	1,516,959 (44.4%)	1.06	+ 5.7%
Black	326,941 (61.8%)	329,158 (61.8%)	1.01	+ 0.7%	334,490 (61.8%)	1.02	+ 2.3%
AAPI	53,705 (60.2%)	49,314 (59.9%)	0.92	− 8.2%	41,592 (62.7%)	0.77	− 22.6%
Hispanic	193,274 (64.9%)	203,751 (63.9%)	1.05	+ 5.4%	79,271 (67.9%)	0.41	− 59.0%
AIAN	9,551 (59.8%)	8,902 (60.5%)	0.93	− 6.8%	9,050 (60.5%)	0.95	− 5.2%
Other / unknown	n/a	21,372 (48.7%)			37,324 (51.8%)		

Abbreviations: AAPI = Asian American / Pacific Islanders / Native Hawaiians; AIAN = American Indians / Alaskan Natives

Note: Denominator for prevalence of chronic illness is based on the racial/ethnic subgroup population specific to each race variable (OASIS, EDB, RTI).

Table 5.

Comparison of agreement and validity measures for EDB and RTI race with prior studies.

Author(s)	Year	Reference	Sample Size	Race	EDB				RTI (Administrative Data)			
					Sensitivity	Specificity	PPV	Kappa	Sensitivity	Specificity	PPV	Kappa
Waldo	1998–2001	1998–2001 MCBS	32,038 32,038	White	96.5	88.2	98.2	0.81				
				Black	95.6	99.6	96.5	0.96				
				AAPI	54.0	99.8	70.0	0.61				
				Hispanic	35.7	99.9	97.5	0.50				
				AIAN	20.6	99.9	69.5	0.32				
Eicheldinger & Bonito	2003	2000–2002 CAHPS	830,728	White	99.3	61.7	91.7	0.71	-	-	-	-
				Black	97.4	98.8	86.3	0.91	-	-	-	-
				AAPI	54.7	99.8	84.5	0.66	79.2	99.7	81.5	0.80
				Hispanic	29.5	99.9	92.7	0.43	76.6	99.2	84.5	0.79
				AIAN	35.7	99.9	59.9	0.45	-	-	-	-
[This study]	2015	2015 OASIS	4,231,370	White	97.9	79.6	94.4	0.81	96.9	95.5	98.7	0.90
				Black	97.0	99.2	94.8	0.95	96.7	99.4	95.9	0.96
				AAPI	62.6	99.8	84.2	0.71	74.7	99.6	81.1	0.77
				Hispanic	36.2	99.8	92.2	0.50	90.8	98.7	84.8	0.87
				AIAN	43.2	99.8	46.1	0.44	43.0	99.8	46.7	0.45

Abbreviations: AAPI = Pacific Islanders; AIAN = American Indians / Alaskan Natives