doi:10.20944/preprints202201.0227.v1

Title

Enhancing Race and Ethnicity using Bayesian Imputation in an All Payer Claims Database

Authors and Institutions

Sanae El Ibrahimi [1],[2]

Michelle A. Hendricks [1]

Sara E Hallvik [1]

Nazanin Dameshghi [1]

Christi Hildebran [1]

Michael A. Fischer [3],[4]

Scott G. Weiner [3, 4]

Running head

Bayesian Imputation of Race and Ethnicity in APCD

[1] Comagine Health, Division of Research and Evaluation, Portland OR, USA

[2] School of Public Health, Department of Epidemiology and Biostatistics, University of Nevada, Las Vegas NV, USA

[3] Brigham and Women's Hospital, Department of Emergency Medicine, Boston MA, USA

[4] Harvard Medical School, Boston MA, USA

Key Words: Bayesian inference, race and ethnicity imputation, All Payer Claims Database, vital statistics death records, validation

Corresponding author

Sanae El Ibrahimi
Comagine Health
650 NE Holladay St # 1700,
Portland, OR 97232
selibrahimi@comagine.org
(702) 336-4408

Word count: 2969

Number of tables: 2

Figures: 0

Number of References: 36

Acknowledgements

The authors wish to thank the Oregon Data Collaborative and participating data suppliers for making the data available for research.

Abstract

Background: All Payer Claims Databases (APCD) are a rich source of health information, however, race and ethnicity (R&E) data are largely missing. Bayesian Improved Surname Geocoding (BISG) is a common R&E imputation method, yet, validation of BISG in APCDs is lacking. We used the BISG to impute missing R&E in the Oregon APCD.

Methods: BISG imputed R&E for Asian Pacific Islanders (API), Blacks, Hispanics and Whites were contrasted to the gold standard (vital statistics) and sensitivity and specificity improvements were assessed. Logistic regression examined whether missing R&E was random across patient characteristics.

Results: Among 85,857 individuals in the study, 32.1% (n=27,594) had missing R&E. Missing R&E was not randomly distributed. There were higher odds of missingness among males, Whites, those age 65 and older, and commercially insured individuals. Differences in the percent missing were also found by co-morbid conditions and mortality causes. Imputing the missing R&E with BISG method improved the sensitivity to identify White, Black, API, and Hispanics.

Conclusions: APCDs can benefit from enhancing missing R&E with BISG imputation to perform more robust population-health level analyses and identify inequities according to R&E without losing power or dropping non-random records with missing R&E data.

Key Words

Bayesian inference, Race and ethnicity imputation, All Payer Claims Database, vital statistics death records

Introduction

In the everchanging healthcare landscape, data are required to develop new approaches to improve healthcare quality, efficiently use resources, and analyze system performance. To meet this growing need, states are increasingly mandating through legislation that commercial and public payers providing insurance plans in their state submit data to All-Payer Claims Data (APCD) databases that include medical claims, pharmacy claims, insurance enrollment data, provider information, and dental claims [1]. To date, many states have existing (mandated or voluntary), under implementation or strong interest in APCDs [2]. States can use APCDs for many purposes, including to improve health system performance, assess the impact of policy changes, understand key cost and utilization drivers, monitor population health trends, develop interventions, and conduct research [3, 4].

APCDs are a rich source of clinical information that have great potential for providing the data needed to comprehensively address long-standing race and ethnic inequities in the quality of healthcare delivery at the population level [5]. This potential is constrained, however, by the lack of reliable race and ethnicity information in APCDs. Despite long-standing recommendations from the Institute of Medicine, the National Quality Forum, and others that health plans systematically collect race and ethnicity data, implementation has been slow due to privacy concerns and resource limitations [6-9]. Even Medicare and Medicaid plans, which are federally mandated to collect race and ethnicity data by the Affordable Care Act [10], continue to struggle to collect this information completely [11].

To address the ongoing challenges regarding limited race and ethnicity data in administrative datasets, Elliott and colleagues developed the Bayesian Improved Surname Geocoding (BISG) imputation method [12]. BISG estimates the probability that an individual is a member of a given racial or ethnicity category based on their surname and their address. Surname analysis is conducted using the U.S. Census Surname list, which provides common surnames for racial and ethnic groups based on information collected from the decennial census [13]. Using Bayesian estimation, this information is combined with the racial and ethnic composition in the census block group where the individual lives to estimate likely racial and ethnic membership. BISG can provide estimates of racial and ethnic health disparities at the group or population level [14]. It has recently been used to demonstrate significant racial/ethnic disparities in behavioral health care quality measures among Medicare Advantage enrollees [15].

As BISG is increasingly used in healthcare delivery and quality improvement, validation of the methodology in multiple settings is fundamental. Previous validation studies comparing BISG estimates to self-reported race and ethnicity have found that the BISG algorithm reliably predicts categories of racial and ethnic membership among commercial health plan members [16, 17]. BISG performs well for White, Black, Hispanic, and Asian individuals but identification of American Indian/Alaskan Native (AIAN) and Multi-racial individuals is typically poor [14]. A version of BISG has also been developed for use with Medicare data, with similar limitations [18]. However, to date, validation of BISG with a multi-payer data set such as the APCD is lacking [14]. Given the growing prominence of APCDs and their potential to inform state policies to address health disparities in quality of care and patient outcomes, the validation of reliable methods to estimate race and ethnicity in APCDs is an important public health issue.

The aim of this study is to validate and use BISG to estimate distribution of patient characteristics according to race and ethnicity in Oregon's APCD. First, differential missingness of race and ethnicity in the APCD is assessed by patient characteristics. Then, the ability of BISG to accurately impute missing race and ethnicity for Asian/Pacific Islander (API), Black, Hispanic, and White populations was examined using Oregon death certificates as the "gold standard" [19].

Methods

2.1 data sources

The sample was extracted from the Oregon Data Collaborative's voluntary APCD, which includes medical and pharmacy claims and enrollment data for Medicaid, Medicare Advantage, and most Commercial plans in the state, covering about 80% of Oregonians. The study sample was limited to patients with a valid Oregon address and age between 2 and 100 years as of 1/1/2014. Detailed patient demographics (e.g. surnames and addresses) for Medicare Fee-for-Service (FFS) beneficiaries were not available at the time of the study and, therefore, were not included in this analysis. The APCD was linked to the vital statistics death records between 2013-2018 using a probabilistic methodology matching on name and date of birth using the Fastlink R package [20].

Three sources of race/ethnicity were examined: 1) death records race/ethnicity (the gold standard); 2) APCD race/ethnicity from the member eligibility files; and 3) BISG imputed race/ethnicity when missing [16].

2.2 Race/ethnicity variable composition

The initial sample contained 105,240 individuals who had a vital statistics record and were linked with the APCD. Individuals with missing race/ethnicity in vital statistics data were excluded (n=197 (0.19%)).

Vital statistics race/ethnicity

A combined vitals race/ethnicity variable was created and included: American Indian/Alaska Native (AIAN), Asian Pacific Islander (API), Black, Hispanic (any race), White, and "other" race group. AIAN individuals and individuals with more than one race were placed in the "other race" category due to the limitations of low imputation accuracy of BISG for AIAN and multi-race [14].

APCD race/ethnicity

APCD race/ethnicity variables were transformed to align with the same categories used for vital statistics resulting in these APCD self-reported race/ethnicity variable categories: API, Black, Hispanic, other, unknown, White.

BISG race/ethnicity imputation

To impute BISG race/ethnicity for individuals in the APCD, geographic identifiers were allocated to each members' address of residence in the eligibility files using the Census Geocoding Application Performing Interface Services [21]. Addresses were then normalized (i.e. street, city, state, zip) to match the required format to perform batch geocoding. Next, the 12-digit block group (an area containing about 1,000 persons) Federal Information Processing Standard (FIPS) codes were extracted [12, 16]. Lastly, the five race/ethnicity probabilities (API, Black, Hispanic, Other, White) were calculated using the R package wru: *Who Are You? Bayesian Prediction of Racial Category Using Surname and Geolocation* [22] that can be accessed from https://github.com/kosukeimai/wru. The package produced probabilities for each category based on members' surname (last name) compared to the 2010 Census Bureau Surname enumeration list, the Spanish Surname list [23] and the racial/ethnic distribution of the individuals' Census 2000 block group. The AIAN race category in BISG imputation was not included because imputing AIAN from surname and residence for this classification remains unreliable [24].

To estimate the optimal probability thresholds to create binary BISG race/ethnicity variables, the Youden Index optimal cut off [25] was applied which maximizes the difference between the true positives and the false positives on the receiver operating characteristic (ROC) curve. The Youden index produced overlapping cutoffs for Whites (0.44) and Blacks (0.48) where some individuals would be predicted to belong to both White and Black race. Therefore, discriminative thresholds of at least 0.5 were used to assign patients to a single race/ethnic group. Sensitivity analyses between a 0.5 and 0.75 cutoff were conducted based on previous studies [17, 26]. A total of 98.4% of the study population met the starting cutoff of 0.5 for the assignment of race ethnicity, while only 83.1% met criteria at the 0.75 threshold. In the sensitivity analyses, there was a considerable decrease in sensitivity and specificity when the 0.75 is applied compared to the 0.5. Based on this, the 0.5 cutoff was used to assign BISG race/ethnicity. There were 141 records for which BISG imputation was undefined, i.e. none of the race categories reached the 0.5 cutoff, these were excluded. BISG probabilities were not imputed for individuals without a geocoded address (missing address completely or partially, PO boxes). There were 19,045 records with both missing APCD self-reported and BISG imputed race and ethnicity. These records were excluded from the initial sample resulting in a total sample of 85,857 individuals.

Enhanced APCD race/ethnicity

BISG imputed probabilities were used to assign race/ethnicity for APCD records where it was missing.

2.3 Analysis variables

To assess the differential missingness of race and ethnicity in the APCD records, an indicator was created of whether APCD race ethnicity was missing. Patient characteristics included age at death, gender, payer, year of death, comorbidities and the top ten causes of death. Because age at death from the vital statistics was used, the age distribution is skewed towards older ages; therefore, age was grouped as: <65, 65-74, 75-84, 85+. The payer attributed to each patient was the plan that the individual had for most of the year of death. Comorbidities were flagged using the R "comorbidity" package [27, 28]. This program uses the International Classification of Diseases (ICD) ninth and tenth revision codes based on the Quan et al. definitions [29]. Comorbidities present in 10% of the sample or more are reported.

The top ten causes of death are based on the national "leading causes of death" report [30]. Causes of death were identified using ICD-10 from the underlying causes of death including diseases of heart (I00–I09, I11, I13, I20–I51), malignant neoplasms (C00–C97), accidents (unintentional injuries) (V01–X59, Y85–Y86), chronic lower respiratory diseases (J40–J47), cerebrovascular diseases (I60–I69), Alzheimer's disease (G30), diabetes mellitus (E10–E14), influenza and pneumonia (J09–J18), nephritis, nephrotic syndrome and nephrosis (N00–N07, N17–N19, N25–N27), intentional self-harm (suicide) (U03,X60–X84,Y87.0). Of note, during the study period, there were no deaths directly attributed to Alzheimer's disease in this sample.

2.3 Analysis

Patient characteristics were tabulated by status of missing race and ethnicity in the APCD and this distribution was tested statistically with the Chi Square tests. Multivariable logistic regression was used to assess factors associated with increased odds of missing race and ethnicity in the APCD. Validation of the various race/ethnicity sources and improvements in sensitivity were compared against the gold standard vital statistics.

SAS Studio 9.3 (SAS Institute, Cary, NC) and RStudio version 1.3 were used for all analyses. This study was not deemed Human Subject Research by the Partners Healthcare Human Research Committee (protocol 2018P001185/PHS) and patient consent was not collected.

Results

The analytical file included 85,857 APCD records of which 32.1% (n=27,594) had missing APCD race/ethnicity. Individuals in the sample were mostly aged 65 years and older (68.5%), female (52.2%), White race based on vital statistics (91.8%) and were Medicare insured at the time of death (70.2%). Almost 30% of individuals had uncomplicated hypertension. Uncomplicated diabetes, chronic pulmonary disease and cardiac arrhythmia were each present in about 20% of the sample. The leading cause of death was malignant neoplasms (21.8%) followed by heart disease (15.9%) and chronic lower respiratory diseases (6.2) (Table 1).

[Place Table 1 here]

Missing APCD race and ethnicity data were not randomly distributed across patient characteristics. The logistic regression results showed multiple factors that were significantly associated with the odds of missing APCD self-reported race and ethnicity. For instance, a gradual increase in the frequency of missing race and ethnicity with increasing age was observed. Those in the 85 years of age and older age category had almost 4-fold increased odds of missing race and ethnicity (adjusted odds ratio (aOR) 3.74, 95% confidence level (CI) 3.51-3.98) compared to patients less than 65 years of age. Males were also more likely to be missing race and ethnicity (aOR 1.46, 95% CI 1.41-1.51) compared to females. Compared to vital statistics Whites, all other racial and ethnic groups had lower odds of missing APCD race and ethnicity. Missing race and ethnicity was more common in more recent years. Compared to Medicaid enrollees, individuals with commercial insurance had a considerably higher likelihood of missing APCD race and ethnicity (aOR 43.8, 95% CI 40.17-47.77). The likelihood of missing race and ethnicity in the APCD differed significantly for patients with different comorbidities. Some conditions had increased likelihood such as uncomplicated hypertension, cardiac arrhythmia, solid tumors and renal failure, while others had lower likelihood such as diabetes, chronic pulmonary disease, and depression. Individuals who died from suicide had a two-fold increased odds of having missing APCD race and ethnicity (aOR 2.10, 95% CI 1.84-2.40), while those who died of diabetes had similar odds (OR 0.91, 95% 0.79-1.04) (Table 1).

Table 2 contrasts the distribution and changes in sensitivity and specificity of the APCD race and ethnicity variables before and after enhancement with BISG imputation compared to the "gold standard" vital statistics race and ethnicity. Enhancing missing race and ethnicity in the APCD improves sensitivity for Whites (from 61.1% to 97.2%), Blacks (from 81.7% to 90%), Hispanics (from 51.1% to 66.2%) and API (from 62.6% to 78.2%). In contrast, using BISG imputation alone was not sufficient to capture race and ethnicity for this sample as demonstrated by lower sensitivity and especially specificity for Blacks (26.8%) (Table 2).

[Place Table 2 here]

Discussion

This is the first study to our knowledge to examine the ability of the BISG algorithm to accurately impute race and ethnicity when missing in an APCD. Findings suggest that missingness of race and ethnicity variables in the APCD is common. Moreover, missingness does not occur equally; it is more likely among males, Whites, older and commercially insured individuals. There were also significant differences between missing and non-missing records in representation of comorbid conditions and mortality cause. This study found that imputing the missing race/ethnicity with the BISG method greatly improved the sensitivity to detect White, Black, API, and Hispanic groups.

Missing race/ethnicity in administrative data sets is a known issue [18, 26, 31]. The APCD data set, by including most of a state's population, can be valuable to answer many important health questions. However, with a high proportion of the data set missing race/ethnicity, the ability of APCD-based analyses to assess health outcomes is limited. Enhancing APCD data with indirect imputation based on surname and address can allow for more robust population health studies. This is particularly important as excluding records with missing race/ethnicity may introduce bias, especially if that missingness is not random as observed in this study. For instance, Grundmeier and colleagues simulated not-randomly missing race/ethnicity data to compare the association of race/ethnicity with pediatric health outcomes when race/ethnicity is imputed vs. excluded. They found that imputing missing data with BISG reduces bias compared to when only data with non-missing values are used [32]. This study found similar results.

Compared to other validation studies, findings from this study show that BISG alone was less optimal to estimate race and ethnic categories correctly, particularly for Blacks (low sensitivity and specificity). For example, Adjaye-Gbewonyo et al. found that BISG captured 71.8% of Blacks while in this study the sensitivity was less optimal at 58% only [33]. This is likely driven by the racial composition of the sample studied. This study is drawn from a mostly White state (Oregon) where Blacks represent a very slim share of the population (2.2%), whereas Adjaye-Gbewonyo and colleagues performed BISG validation on a population form Georgia, a state with a higher than average Black population (32.6%) [34].

The lower prediction ability of BISG in a less diverse population is likely due to relying on the Surname Census list. Hispanic and Asian individuals may have more distinctive surnames, which is not the case for surnames of Whites and Blacks. For instance, “Williams”, the third most common surname in the U.S. [13], has close to 50/50 probability of either White or Black race. Therefore, in a predominantly White population, the ability of the Surname list to correctly predict Whites vs Blacks may be diminished. BISG could be further improved by incorporating information from the “first name” into the prediction algorithm, particularly for Blacks. Ioan Voicu proposed the Bayesian Improved First Name Surname Geocoding (BIFSG) using a list of first names from mortgage applications [35]. While the author noted amelioration in the predictions compared to BISG, further work is needed using a more representative first name list other than those of mortgage applicants.

Another potential reason for the lower BISG prediction of non-White race/ethnic groups is the use of geographic distribution of race/ethnicity which may be influenced by the level of residential segregation in communities. Higher level of segregation is likely to result in more precise prediction of the race/ethnic groups. The state of Oregon is somewhat more integrated than other states [36]. Having diverse communities might make it harder for the BISG algorithm to attribute deterministic probabilities for the racial make-up of a community resulting in less accurate predictions. This study showed that using the BISG method to “enhance” the existing race/ethnicity information is more suitable than relying on BISG predictions alone especially in less diverse and more integrated communities.

Limitations

This study has a number of limitations. The use of death data means that the sample for this study skews older than the Oregon population, in which the race and ethnic distributions may be different. Using a more representative data source with respect to age and that includes a more reliable capture of AIAN race is warranted. Because of data use agreement restrictions, the voluntary APCD used in this study did not include Medicare fee-for-service patients, however, this may have attenuated the skewness towards older patients. Not all individuals had an address that could be properly geocoded. Finally, given the less diverse Oregon population reflected in the high percent White and the low discriminative power of surname probabilities for White and Black surnames, we decided to drop any missing BISG imputed probabilities. Future studies may add the first name to improve the prediction for Blacks.

Conclusion

Health equity requires accurately assessing the burden of disease for the different race and ethnic groups to reduce disparities and appropriately allocate resources and interventions. Given the growing use of APCDs and their potential for developing population-based interventions, it is critical to develop methods that allow for the non-biased estimation of race/ethnicity. Having complete race/ethnicity information in administrative data sets would allow to perform more robust population-health level analyses to identify inequities according to race/ethnicity, without losing power or dropping non-random records with missing race data.

Abbreviations:

AIAN: American Indian Alaska Native; APCD: All Payer Claims Data; API: Asian Pacific Islander; BISG: Bayesian Improved Surname and Geolocation; FIPS: Federal Information Processing Standard; ICD: International Classification of Diseases; OR: Odds ratio

References

1. Porter J, Love D, Costello A, Peters A, Rudolph B. All-Payer Claims Database Development Manual: Establishing a Foundation for Health Care Transparency and Informed Decision Making. Institute for Health Policy and Practice (IHPP); 2015.
2. APCD Council. Interactive State Report Map (2020). updated 2021. Available from: https://www.apcdcouncil.org/state/map.
3. Porter J, Love D. The ABCs of APCDs: How States Are Using Claims Data to Understand and Improve Care. California Health Care Foundation; 2018.
4. APCD Council, National Association of Health Data Organizations, New Hampshire Institute for Health Policy and Practice. Informing Health System Change - Use of All-Payer Claims Databases. 2018.
5. US Department of Health Human Services. 2016 National Healthcare Quality and Disparities Report: Agency for Healthcare Research and Quality Rockville, MD; (2017). Available from: https://www.ahrq.gov/research/findings/nhqrdr/nhqdr16/index.html?utm_source=ahrq&utm_medium=en2&utm_term=&utm_content=2&utm_campaign=ahrq_en8_15_2017.
6. Landon BE, Normand S-LT, editors. National voluntary consensus standards for ambulatory care: Measurement challenges in small group settings2006: National Quality Forum.
7. Nerenz DR, Carreon R, Veselovskiy G. Race, ethnicity, and language data collection by health plans: findings from 2010 AHIPF-RWJF survey. J Health Care Poor Underserved. 2013;24(4):1769-83.
8. Smedley BD, Stith AY, Nelson AR. Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care.(2003). Unequal treatment: Confronting racial and ethnic disparities in health care. 2003.
9. Escarce JJ, Carreon R, Veselovskiy G, Lawson EH. Collection of race and ethnicity data by health plans has grown substantially, but opportunities remain to expand efforts. Health Aff (Millwood). 2011;30(10):1984-91.
10. Patient Protection and Affordable Care Act. §3022 and §10307, Pub L No. 111-148, 124 Stat 119. 2010.
11. 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(3):548-52.
12. 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. 2008;43(5 Pt 1):1722-36.
13. Comenetz J. Frequently occurring surnames in the 2010 census. 2016.
14. Fremont A, Weissman JS, Hoch E, Elliott MN. When race/ethnicity data are lacking: using advanced indirect estimation methods to measure disparities. Rand Health Q. 2016;6(1):16.
15. Breslau J, Elliott MN, Haviland AM, Klein DJ, Dembosky JW, Adams JL, et al. Racial And Ethnic Differences In The Attainment Of Behavioral Health Quality Measures In Medicare Advantage Plans. Health Aff (Millwood). 2018;37(10):1685-92.
16. Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N. Using the Census Bureau’s surname list to improve estimates of race/ethnicity and associated disparities. Health Serv Outcomes Res Methodol. 2009;9(2):69-83.
17. Adjaye-Gbewonyo D, Bednarczyk RA, Davis RL, Omer SB. Using the bayesian improved surname geocoding method (BISG) to create a working classification of race and ethnicity in a diverse managed care population: a validation study. Health Serv Res. 2014;49(1):268-83.
18. Haas A, Elliott MN, Dembosky JW, Adams JL, Wilson-Frederick SM, Mallett JS, et al. Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. Health Serv Res. 2019;54(1):13-23.

19. Arias E, Heron MP, Hakes JK. The validity of race and Hispanic origin reporting on death certificates in the United States: an update. 2016(148):1-23.

20. Enamorado T, Fifield B, Imai K. Using a probabilistic model to assist merging of large-scale administrative records. Available at SSRN 3214172. 2018;113(2):353-71.

21. U.S. Census Bureau. Geocoding Services Web Application Programming Interface (API) (2019). cited 2020. Available from: https://geocoding.geo.census.gov/geocoder/Geocoding_Services_API.pdf.

22. Imai K, Khanna K. Improving ecological inference by predicting individual ethnicity from voter registration records. Polit Anal. 2016;24(2):263-72.

23. Word DL, Coleman CD, Nunziata R, Kominski R. Demographic aspects of surnames from census 2000. Unpublished manuscript, Retrieved from http://citeseerx ist psu edu/viewdoc/download. 2008.

24. Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. Health Services and Outcomes Research Methodology. 2009;9(2):69-83.

25. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biometrical Journal: Journal of Mathematical Methods in Biosciences. 2005;47(4):458-72.

26. Derose SF, Contreras R, Coleman KJ, Koebnick C, Jacobsen SJ. Race and ethnicity data quality and imputation using U.S. Census data in an integrated health system: the Kaiser Permanente Southern California experience. Med Care Res Rev. 2013;70(3):330-45.

27. Gasparini A. Comorbidity: an R package for computing comorbidity scores. Journal of Open Source Software. 2018;3(23):648.

28. Elixhauser A, Steiner CA, Harris D, Coffey R. Comorbidity measures for use with administrative data. Med Care. 1998(36):8-27.

29. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi J-C, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005;43(11):1130-9.

30. Heron MP. Deaths: leading causes for 2017. 2019.

31. Brown DP, Knapp C, Baker K, Kaufmann M. Using Bayesian Imputation to Assess Racial and Ethnic Disparities in Pediatric Performance Measures. Health Serv Res. 2016;51(3):1095-108.

32. Grundmeier RW, Song L, Ramos MJ, Fiks AG, Elliott MN, Fremont A, et al. Imputing missing race/ethnicity in pediatric electronic health records: reducing bias with use of US census location and surname data. Health Serv Res. 2015;50(4):946-60.

33. Adjaye-Gbewonyo D, Bednarczyk RA, Davis RL, Omer SB. Using the Bayesian Improved Surname GeocodingMethod (BISG) to Create a Working Classification ofRace and Ethnicity in a Diverse Managed Care Population: AValidation Study. Health Services Research. 2014;49(1):268-83.

34. Census Bureau. Population estimates, July 1, 2019 (2019). cited 2020 September. Available from: https://www.census.gov/quickfacts/fact/table/US/PST045219.

35. Voicu I. Using first name information to improve race and ethnicity classification. Statistics and Public Policy. 2018;5(1):1-13.

36. Oregon Health Authority. Social Determinants of Health: Residential racial segregation (2019). Available from: https://www.oregon.gov/oha/PH/ABOUT/Documents/indicators/segregation.pdf.

Statements and Declarations:

Funding:

This work was funded by NIH grant 1-R01-DA044167. The funding source was not involved in the study design, collection, analysis and interpretation of data, nor in the writing of the report or the decision to submit the report for publication. The conclusions in this article are those of the authors and do not necessarily represent the official position of the funder. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Competing Interests

The authors have no relevant financial or non-financial interests to disclose

Authors' contributions

S. E. conceptualize, analyzed the data and wrote the manuscript, M. H. reviewed the analysis and the manuscript, S. H. reviewed the analysis and the manuscript, N. H. prepared the data for analysis, C. H. reviewed the manuscript, M. F. reviewed the manuscript, S. W. conceptualized and reviewed the manuscript. All authors reviewed the manuscript.

Ethics approval and consent to participate:

This study was not deemed Human Subject Research by the Partners Healthcare Human Research Committee (protocol 2018P001185/PHS) and patient consent was not collected.

Consent for publication

All authors consent to publish this manuscript.

doi:10.20944/preprints202201.0227.v1

Table 1. Patient characteristics and odds ratios of missing self-reported APCD race and ethnicity

Patient characteristics	**Total**	**Non-missing APCD race & ethnicity**	**Missing APCD race & ethnicity**	**Odds of missing race & ethnicity**
	n (%)	**n (%)**	**n (%)**	OR (95% CI)
Total sample	85857 (100)	58263 (67.9)	27594 (32.1)	
Age at death				
<65	27035 (31.5)	21263 (36.5)	5772 (20.9)	Ref
65-74	15975 (18.6)	10747 (18.5)	5228 (19.0)	2.22 (2.08-2.37)
75-84	18115 (21.1)	11439 (19.6)	6676 (24.2)	3.04 (2.85-3.25)
85+	24732 (28.8)	14814 (25.4)	9918 (35.9)	3.74 (3.51-3.98)
Gender				
Female	44804 (52.2)	31397 (53.9)	13407 (48.6)	Ref
Male	41053 (47.8)	26866 (46.1)	14187 (51.4)	1.46 (1.41-1.51)
Vital race and ethnicity				
White	78855 (91.8)	52732 (90.5)	26123 (94.7)	Ref
Black	1575 (1.8)	1363 (2.3)	212 (0.8)	0.889 (0.793-0.99)
Hispanic	2319 (2.7)	1766 (3.0)	553 (2.0)	0.58 (0.51-0.65)
API	1727 (2.0)	1266 (2.2)	461 (1.7)	0.41 (0.35-0.48)
Other	1381 (1.6)	1136 (2.0)	245 (0.9)	0.62 (0.53-0.73)
Year of death				
2013	9463 (11.3)	7260 (12.7)	2203 (8.3)	Ref
2014	11762 (14.1)	8285 (14.5)	3477 (13.1)	1.41 (1.31-1.51)
2015	13434 (16.1)	9524 (16.7)	3910 (14.7)	1.41 (1.32-1.51)
2016	14727 (17.6)	9579 (16.8)	5148 (19.4)	1.81 (1.70-1.94)
2017	17102 (20.4)	10999 (19.3)	6103 (23.0)	1.79 (1.68-1.91)
2018	17196 (20.6)	11453 (20.1)	5743 (21.6)	1.56 (1.47-1.67)
Payer [†]				
Commercial	6098 (7.3)	1902 (3.3)	4196 (15.8)	43.76 (40.17-47.77)
Medicaid	18873 (22.6)	17822 (31.2)	1051 (4.0)	Ref
Medicare	58713 (70.2)	37376 (65.5)	21337 (80.3)	5.27 (4.91-5.67)
Elixhauser comorbidities, prevalence 10%+ [‡]				
Hypertension, uncomplicated	25235 (29.4)	15341 (26.3)	9894 (35.9)	1.34 (1.29-1.39)
Diabetes, uncomplicated	18610 (21.7)	13171 (22.6)	5439 (19.7)	0.83 (0.79-0.87)
Chronic pulmonary disease	17154 (20.0)	12433 (21.3)	4721 (17.1)	0.76 (0.73-0.80)
Cardiac arrhythmias	16799 (19.6)	10176 (17.5)	6623 (24.0)	1.37 (1.32-1.43)
Congestive heart failure	14142 (16.5)	9619 (16.5)	4523 (16.4)	0.84 (0.80-0.88)
Solid tumor, without metastasis	11540 (13.4)	6569 (11.3)	4971 (18.0)	1.50 (1.43-1.58)
Diabetes, complicated	10018 (11.7)	7350 (12.6)	2668 (9.7)	0.76 (0.72-0.81)
Depression	9614 (11.2)	7279 (12.5)	2335 (8.46)	0.74 (0.70-0.78)
Fluid and electrolyte disorders	9495 (11.1)	6570 (11.3)	2925 (10.6)	0.89 (0.84-0.94)
Renal failure	8599 (10.0)	5552 (9.5)	3047 (11.0)	1.19 (1.13-1.26)
Top 10 causes of death [‡, §]				
Malignant neoplasms	18744 (21.8)	11232 (19.3)	7512 (27.2)	1.62 (1.55-1.70)
Heart disease	13645 (15.9)	8581 (14.7)	5064 (18.4)	1.37 (1.31-1.44)

Chronic lower respiratory diseases	5276 (6.2)	4028 (6.9)	1248 (4.5)	0.94 (0.87-1.02)
Accidents	5099 (5.9)	3692 (6.3)	1407 (5.1)	1.34 (1.24-1.45)
Cerebrovascular diseases	3628 (4.2)	2381 (4.1)	1247 (4.5)	1.21 (1.12-1.31)
Intentional self-harm (suicide)	2045 (2.4)	1347 (2.3)	698 (2.5)	2.10 (1.84-2.40)
Diabetes mellitus	1461 (1.7)	1129 (1.9)	332 (1.2)	0.91 (0.79-1.04)
Nephritis, nephrotic syndrome and nephrosis	921 (1.1)	653 (1.1)	268 (1.0)	1.01 (0.86-1.18)
Influenza and pneumonia	693 (0.8)	486 (0.8)	207 (0.8)	1.14 (0.95-1.36)

CI: confidence interval; OR: odds ratio

[†] Payer corresponds to the payer on the year of death

[‡] Reference = no event

[§] There were no deaths associated with Alzheimer's in the study period

Table 2. Improvement in the sensitivity and specificity of race and ethnicity after enhancement with the BISG

Race and Ethnicity	Vital statistics[†]		APCD				BISG imputation				Enhanced APCD			
	n	%	n	%	Sensitivity	Specificity	n	%	Sensitivity	Specificity	n	%	Sensitivity	Specificity
White	78,855	92	52,457	61	65.2	98	63,890	74	78.4	98	78,308	91	97.2	97.9
Black	1,575	1.8	1,403	1.6	81.7	91.7	3,400	4	58	26.8	2,160	2.5	90	65.7
Hispanic	2,319	2.7	2,082	2.4	51.1	56.9	2,182	2.5	56.1	56.9	2,662	3.1	66.2	57.7
API	17,27	2	1,231	1.4	62.6	87.1	1,346	1.6	60.5	77.6	1,599	1.9	78.2	84.4
Other [‡]	1,381	1.6	1,090	1.3			180	0.2			1,128	1.3		
Missing			27,594	32			14,859	17						

APCD: All Payer claims data; API: Asian/Pacific Islander; BISG: Bayesian Improved Surname Geocoding Method

[†] Vital statistics race/ethnicity is considered the Gold Standard

[‡] The "other" category is not homogenous across the different race/ethnicity source variables, therefore, sensitivity is not calculated for this category