HHS Public Access

Author manuscript

Traffic Inj Prev. Author manuscript; available in PMC 2022 August 17.

Published in final edited form as:

Traffic Inj Prev. 2021 ; 22(sup1): S32–S37. doi:10.1080/15389588.2021.1955109.

Facilitating research on racial and ethnic disparities and inequities in transportation: Application and evaluation of the Bayesian Improved Surname Geocoding (BISG) algorithm

Emma B. Sartin, PhD, MPH[a], **Kristina B. Metzger, PhD, MPH**[a], **Melissa R. Pfeiffer, MPH**[a], **Rachel K. Myers, PhD, MS**[a,b], **Allison E. Curry, PhD, MPH**[a,b]

[a)]Center for Injury Research and Prevention, Children's Hospital of Philadelphia, Philadelphia, PA

[b)]Division of Emergency Medicine, Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

Abstract

Objective: Racial and ethnic disparities and/or inequities have been documented in traffic safety research. However, race/ethnicity data are often not captured in population-level traffic safety databases, limiting the field's ability to comprehensively study racial/ethnic differences in transportation outcomes, as well as our ability to mitigate them. To overcome this limitation, we explored the utility of estimating race and ethnicity for drivers in the New Jersey Safety and Health Outcomes (NJ-SHO) data warehouse using the Bayesian Improved Surname Geocoding (BISG) algorithm. In addition, we summarize important recommendations established to guide researchers developing and implementing racial and ethnic disparity research.

Methods: We applied BISG to estimate population-level race/ethnicity for New Jersey drivers in 2017 and evaluated the concordance between reported values available in integrated administrative sources (e.g., hospital records) and BISG probability distributions using an area under the receiver operator curve (AUC) within each race/ethnicity category. Overall AUC was calculated by weighting each AUC value by the population count in each reported category. In an exemplar analysis using 2017 crash data, we conducted an analysis of average monthly police-reported crash rates in 2017 by race/ethnicity using the NJ-SHO and BISG sets of race/ethnicity values to compare their outputs.

Results: We found excellent or outstanding concordance (AUC ≥0.86) between reported race/ethnicity and BISG probabilities for White, Hispanic, Black, and Asian/Pacific Islander drivers.

Corresponding author: Emma B. Sartin, PhD, MPH, Center for Injury Research and Prevention, Children's Hospital of Philadelphia, 2716 South Street, 13th Floor, Philadelphia, PA 19146 USA, sartine@chop.edu.
Kristina B. Metzger, PhD, MPH, Center for Injury Research and Prevention, Children's Hospital of Philadelphia, 2716 South Street, 13th Floor, Philadelphia, PA 19146 USA
Melissa R. Pfeiffer, MPH, Center for Injury Research and Prevention, Children's Hospital of Philadelphia, 2716 South Street, 13th Floor, Philadelphia, PA 19146 USA
Rachel K. Myers, PhD, MS, Center for Injury Research and Prevention, Children's Hospital of Philadelphia, 2716 South Street, 13th Floor, Philadelphia, PA 19146 USA
Allison E. Curry, PhD, MPH, Center for Injury Research and Prevention, Children's Hospital of Philadelphia, 2716 South Street, 13th Floor, Philadelphia, PA 19146 USA

CONFLICTS OF INTEREST

The authors have no conflicts of interest to disclose.

We found poor concordance for American Indian/Alaskan Native drivers (AUC= 0.65), and concordance was no better than random assignment for Multiracial drivers (AUC = 0.52). Among White, Hispanic, Asian/Pacific Islander, and American Indian/Alaskan native drivers, monthly crash rates calculated using both NJ-SHO reported race/ethnicity values and BISG probabilities were similar. Monthly crash rates differed by 11% for Black drivers, and by more than 200% for Multiracial drivers.

Conclusion: Findings of excellent or outstanding concordance between and mostly similar crash rates derived from reported race/ethnicity and BISG probabilities for White, Hispanic, Black, and Asian/Pacific Islander drivers (98.9% of all drivers in this sample) demonstrate the potential utility of BISG in enabling research on transportation disparities and inequities. Concordance between race/ethnicity values were not acceptable for American Indian/Alaskan Native and Multiracial drivers, which is similar to previous applications and evaluations of BISG. Future work is needed to determine the extent to which BISG may be applied to traffic safety contexts.

Keywords

traffic crashes; minority health; data warehousing; epidemiology; public health informatics; data integration

BACKGROUND

Racial and ethnic disparities and/or inequities have been documented in traffic safety research. For example, while overall rates of crash fatalities have declined substantially over recent decades, across the lifespan individuals who identify as a racial/ethnic minority remain much more likely to be killed in a motor vehicle crash than White individuals (Harper et al. 2015; Centers for Disease Control and Prevention 2019). In addition, compared with White peers, Black and Hispanic/Latino populations have substantially lower rates of behind-the-wheel driver training and restraint use and are more likely to delay licensure because of costs or other concerns (Curry et al. 2012; Li and Pickrell 2018; Tefft et al. 2014). Further, racial and ethnic minority populations report facing more barriers to transportation, which interfere with employment, housing, health, and well-being (Blumenberg and Pierce 2014; Flores and Tomany-Korman 2008; Sanchez et al. 2004; Syed et al. 2013).

The bulk of existing US studies focused on racial and ethnic differences in transportation have relied on interviews, surveys, and/or direct observations. In comparison, few disparity-focused transportation studies have leveraged administrative or population-level data sources (McAndrews et al. 2013; Zhang and Lin 2013). This is likely because these data sources rarely include race and ethnicity information. A notable exception is the Fatality Analysis Reporting System (FARS), which includes race/ethnicity (as listed on death certificates) for individuals killed in crashes (Briggs et al. 2005). As a result, almost all crash studies examining disparities are limited to fatal crashes. Thus, there remains a critical need for novel, foundational studies that characterize transportation-related racial and ethnic disparities and/or inequities, particularly for non-fatal outcomes (e.g., crashes, injuries, restraint use, citations). Further, these studies must be conducted in an ethical way that both (1) ameliorates potential biases commonly cited in data collection, analysis, and

policymaking and (2) considers race and ethnicity as social constructs used to identify populations at-risk for adverse outcomes due to other causal, intervenable variables (e.g., health behaviors, socioeconomic status, environmental factors).

In prior work, we described our efforts to develop a comprehensive, integrated traffic data source in New Jersey—the New Jersey Safety and Health Outcomes (NJ-SHO) data warehouse—to conduct a range of epidemiologic traffic safety studies (Curry et al. 2019, 2021). As in most other US states, race/ethnicity is not collected on the NJ crash report; NJ also does not capture race and ethnicity in their licensing database. Thus, we explored various methods of incorporating race and ethnicity information into the NJ-SHO to improve the ability of the data warehouse to address critical transportation disparity and inequity research questions. In this initial paper, we examine the use of one such method: Bayesian Improved Surname Geocoding (BISG). First, we introduce BISG as a method to estimate race/ethnicity at the population level. We then use traffic safety data from the NJ-SHO to demonstrate how to implement and evaluate this method. Finally, we provide guidance on how to analyze and interpret BISG race/ethnicity estimates in real-world traffic safety research.

Introduction to Bayesian Improved Surname Geocoding (BISG) Methods

The BISG algorithm was developed by the RAND Corporation for use in health disparity research as an approach to produce accurate and reliable group- or population-level race/ethnicity estimates (Adjaye-Gbewonyo et al. 2014; Elliott et al. 2008, 2009). BISG combines information from the 2000 US Census surname list with information on the racial/ethnic composition of each 2010 US Census block group to produce a set of probabilities that an individual belongs to each of six mutually exclusive racial/ethnic groups: White, Hispanic, Black, Asian/Pacific Islander, Multiracial, and American Indian/Alaska Native. BISG has previously been applied to multiple health topics and data sources. For example, this approach was recently applied to national surveillance data to characterize the magnitude of disparities in COVID-19 outcomes (Labgold et al. 2021). However, to our knowledge, BISG has not been applied to transportation data sources. Given that both surname and residential address are routinely collected on crash reports and in driver licensing databases, BISG may be a viable method to incorporate race/ethnicity into traffic data sources that would otherwise not include this information, thus enabling novel research on racial and ethnic disparities and inequities in transportation outcomes.

Important strengths and limitations of the BISG approach were described in detail in a publication by Fremont *et al.* (2016). First and foremost, BISG was designed as a method to identify disparities *at the population level.* It has been found to accurately estimate membership for the four largest US racial/ethnic group categories: White, Hispanic, Black, and Asian/Pacific Islander. Notably, for racial/ethnic groups that constitute smaller proportions of the US population—American Indian/Alaskan Native and Multiracial—BISG estimates do not have as high concordance with reported values. Conversely, BISG estimates should not be used to classify a specific individual's race/ethnicity, as doing so decreases the overall accuracy of estimates in terms of both efficiency and bias (Fremont et al. 2016). The most accurate estimates of racial/ethnic group compositions or disparities are obtained

by using the BISG algorithm to supplement data sources with reported race/ethnicity; that is, ideally BISG should be used to estimate race/ethnicity for individuals with missing values but should not replace known or reported values (Fremont et al. 2016). Using BISG as a supplement to reported data reduces biases that may arise from a complete case analysis, as race/ethnicity information may not be missing at random, and previous studies have used several methodological approaches to supplement known values with BISG-derived race/ethnicity information in applied analysis (Elliott et al. 2009; Fremont et al. 2016; Labgold et al. 2021).

DATA SOURCE: THE NJ-SHO DATA WAREHOUSE

As described in detail in previous papers, the NJ-SHO contains ~88 million records for 22.3 million NJ residents and is comprised of integrated data from numerous NJ administrative sources for the period of 2004 through 2018 (Curry et al. 2019, 2021). These linked data include (1) driver licensing histories, (2) traffic-related citations and suspensions, (3) police-reported crashes, (4) birth certificates, (5) death certificates, (6) hospital discharges (emergency department, inpatient, and outpatient), and (7) electronic health records (EHR) of patients of the Children's Hospital of Philadelphia network who live in NJ and were born in 1987 through 2000. Race/ethnicity information is available in the original birth, death, hospital discharge, and EHR data sources. For death certificates, race/ethnicity information is provided by next of kin. For EHR and hospital discharge data, race/ethnicity is self- or other- (e.g., caregiver/parent) reported. For birth certificate data, the mother reports separate race/ethnicity values for herself and the father. All activities for this project were approved by the Children's Hospital of Philadelphia Institutional Review Board (IRB 11-008136). All analyses were conducted using SAS version 9.4 (SAS Institute Inc. Cary, NC).

STUDY POPULATION

We identified 6,369,101 individuals from the NJ-SHO population who were aged 17–99 years and had a driver's license at any point in 2017 (hereafter called "drivers"). We collapsed race/ethnicity information across all integrated data sources to create one record per driver, with each distinct value for race and ethnicity reported, regardless of source. We consulted literature describing the accuracy of reported race/ethnicity values in administrative and health data sources and considered the recommendations provided in each of these studies (Agency for Healthcare Research and Quality 2014; Arias et al. 2016; Klinger et al. 2015; West et al. 2005). With these in mind, we developed and applied a hierarchical process to derive a single reported race/ethnicity value for each driver that matched the six mutually exclusive BISG categories (See Supplemental Table 1). Overall, 77.3% (n=4,924,137) of drivers were assigned a race/ethnicity category using reported data integrated from NJ-SHO. A higher proportion of female drivers had a reported race/ethnicity value than male drivers (81.1% and 73.3%, respectively).

CALCULATION AND EVALUATION OF BISG PROBABILITIES

Our approach to derive BISG probabilities and implement them in applied analyses followed strategies recommended in Elliott *et al.* (2009). First, we used each driver's surname and

the geocoded census block group of their most recent residential address (i.e., the address associated with the most recent record for that driver) available in any integrated source to obtain their set of probabilities for the six racial/ethnic groups (Elliott et al. 2009). For each driver, the sum of their six race/ethnicity probabilities equals one. Surname was available for 98.9% (n=6,298,520) of drivers and residential census block group was geocoded for 98.9% (n=6,298,506) of drivers, allowing us to obtain BISG probabilities for 98.9% (n=6,298,506) of drivers.

We then compared the distribution of race/ethnicity using the BISG algorithm and NJ-SHO reported race/ethnicity data. To do this, we restricted the sample to the 76.8% of drivers with available race/ethnicity from both reported NJ-SHO values and BISG estimates (n=4,890,549). Each driver had 2 different sets of race/ethnicity values: 1) a single value derived from race/ethnicity information available from the integrated NJ-SHO data sources and 2) six probabilities derived from the BISG algorithm. The distribution of race/ethnicity in this sub-sample of drivers in 2017 was calculated by using the proportions of NJ-SHO reported values and by calculating the means of BISG probabilities among all drivers, as well as by sex (female and male); these distributions are shown in Table 1. The majority of drivers were classified as White using both the NJ-SHO reported values (64.6%) and the BISG probabilities (67.2%). Compared with the NJ-SHO values, the BISG algorithm indicated a slightly larger proportion of White drivers and a slightly lower proportion of Hispanic drivers. Both categorization methods calculated similar proportions of Black drivers (NJ-SHO: 10.2%; BISG: 10.4%). The proportion of Asian/Pacific Islanders was 5.8% using NJ-SHO values and 6.9% using BISG probabilities. The remaining 2.1% (using NJ-SHO) and 1.3% (using BISG) of drivers were Multiracial or American Indian/Alaska native. Similar patterns were noted for the distribution of race/ethnicity among female and male drivers.

Finally, to conduct an initial evaluation of this approach, we determined the concordance between reported race/ethnicity value and BISG probabilities using area under the receiver operator curve (AUC) within each race/ethnicity category (Elliott et al. 2009; Hosmer et al. 2013). For example, the AUC value for the White race/ethnicity category reflects how well the BISG value for White (a continuous variable with range of 0 to 1) predicts the reported NJ-SHO value for White (a dichotomous variable with values 0 or 1) in the study population. A larger AUC value indicates that the BISG probability is better at distinguishing between NJ-SHO reported values for that category of race/ethnicity; excellent discrimination is considered as $0.8 \leq AUC < 0.9$ and outstanding discrimination as $AUC \geq 0.9$. Overall AUC was calculated by weighting each AUC value by the population count in each reported category. Concordance analyses produced AUC values > 0.85 for 4 of 6 race/ethnicity groups with overall AUC of 0.89, indicating that the BISG probabilities were much more likely than not to correctly distinguish the driver's NJ-SHO reported race/ethnicity category (Table 2). Among all drivers, we found excellent concordance for Hispanic (0.86) group classification and outstanding concordance for White (0.90), Black (0.94), and Asian/Pacific Islander (0.90) group classification. Concordance for the Multiracial group was poor (0.65), and concordance for the American Indian/Alaska Native group was no better than random assignment (0.52). Concordance statistics among female drivers were slightly lower

than among male drivers, but the same general pattern was observed across race/ethnicity categories.

PRACTICAL APPLICATION OF METHODS

To demonstrate the utility of the BISG algorithm for estimating race/ethnicity among drivers, we conducted an analysis of average monthly crash rates in 2017 by race/ethnicity using both the NJ-SHO reported values and BISG probabilities. First, we used linked police-reported crash data and driver licensing history data to determine the number of crashes each driver was involved in and the number of months in which they had a license during 2017. Next, we calculated average monthly crash rates per 10,000 license-months by race/ethnicity using NJ-SHO reported values. For each of the 6 NJ-SHO race/ethnicity groups, we summed the total number of crashes (for the numerator) and the total number of license-months (for the denominator) for all drivers within that group. Since each driver has a set of BISG-generated probabilities for the 6 race/ethnicity groups, an alternative approach was taken to calculate the average monthly crash rates by race/ethnicity using BISG estimated values. For each driver, we applied that driver's 6 BISG probabilities to their total number of crashes and license-months to obtain 6 weighted estimates of the proportion of their crashes and license-months attributed to each race/ethnicity group. Average monthly crash rates for BISG race/ethnicity groups were then calculated as the sum of the weighted number of crashes (for the numerator) and the sum of the weighted number of license-months (for the denominator) within each group. Rates were obtained using Poisson regression models.

Monthly crash rates calculated using both NJ-SHO reported race/ethnicity groups and BISG probabilities are presented in Table 3. The magnitude of the rates was similar (within 2%) for White drivers (the largest group) overall and by sex. For Hispanic drivers and Asian/Pacific Islander drivers (both smaller groups with high concordance), the magnitude of the rates were within 4% and 5%, respectively. Although the concordance of classification of Black drivers was highest of all race/ethnicity groups, we observed a greater difference in the magnitude of the rates using NJ-SHO and BISG values (11%). The crash rates calculated among Multiracial drivers differed greatly (by > 200%). Despite poor concordance for the American Indian/Alaska Native group, the crash rates calculated using NJ-SHO reported values and using BISG probabilities are similar (within 7%). Differences in monthly crash rates using the two race/ethnicity classification methods were similar among female and male drivers (Supplemental Table 2).

DISCUSSION

Historically, traffic safety researchers have had only a few large sources of data to leverage for investigations of racial and ethnic disparities and inequities, as the majority of administrative traffic databases do not routinely collect race/ethnicity information. Thus, we have embarked on an effort to identify methods and approaches to derive race and ethnicity. In this initial paper, we introduced the notion of employing the BISG algorithm —which has been applied to other areas of health research to supplement or derive race and ethnicity information in large data sources—for use in traffic safety. We then evaluated the algorithm's performance at estimating race/ethnicity within a large, integrated traffic

safety data source. Finally, we provided an initial exemplar analysis of how BISG-estimated race/ethnicity values may be used to examine racial and ethnic disparities in crash rates.

First, we demonstrated that we were able to calculate BISG race/ethnicity probabilities for almost all (98.9%) of drivers in our sample using surname and residential address, two fields commonly available in licensing and crash data. With respect to the performance of BISG-derived probabilities compared with reported race/ethnicity values available from integrated sources within the NJ-SHO, we found that BISG probabilities exhibited excellent concordance (AUC ≥ 0.86) for the four racial/ethnic groups (White, Hispanic, Black, and Asian/Pacific Islander) that constitute the vast majority (≈98%) of drivers in our sample. On the other hand, our application of BISG produced unacceptable concordance estimates for Multiracial and American Indian/Alaskan Native drivers, which constitute a much smaller proportion (≈2%) of our study sample. These concordance estimates are consistent with those reported in previous applications of BISG (Elliott et al. 2009; Fremont et al. 2016). Finally, we demonstrated in a relatively simple exemplar analysis how BISG probabilities can be applied in traffic safety contexts. When we derived crash rates using reported race/ethnicity and BISG probabilities, we did not observe substantial variations in crash rates among 3 of the 4 racial/ethnic groups with high concordance. The magnitude of the crash rate estimates among Black drivers—the group with the highest concordance—differed by about 11%, which warrants further investigation. Further, we found that BISG probabilities were similar to reported race/ethnicity data among both female drivers and male drivers; concordance statistics and crash rate estimates were similar. Taken together, our strong concordance statistics and similar crash rates among the vast majority of drivers lead us to conclude that BISG may be a promising method to incorporate race/ethnicity information into large traffic data sources that historically collect surnames and residential addresses, even when no race/ethnicity information is originally available (see Hesketh et al. 2020 for an applied example).

As previous studies have indicated, the most accurate estimates of racial/ethnic group composition and disparities are derived when BISG is applied to supplement missing race/ethnicity data (Fremont et al. 2016). Notably, we limited the sample in this paper only to the three-fourths of drivers who had known race/ethnicity values in order to enable direct comparisons between known and BISG-derived race/ethnicity information. However, previous studies have indicated that limiting applied analyses to complete cases—that is, those with known values—may not accurately quantify disparities or inequities, as race/ethnicity information is often not missing at random. Thus, in future work we plan to evaluate BISG's utility in supplementing available race/ethnicity data within the entire NJ-SHO population.

While our goal is to promote racial and ethnic disparity research in traffic safety contexts, there are important ethical implications for researchers to consider before embarking in this line of work (for more recommendations and their explanations, see Flanagin et al. 2021 and Kaplan and Bennett 2003). First, race and ethnicity are social constructs and should therefore be used to identify populations at-risk for adverse outcomes due to other causal, intervenable variables. As such, using race/ethnicity as a routine descriptor in analyses may manifest or support a belief that health disparities are caused by race/

ethnicity instead of underlying constructs (e.g., health behaviors, socioeconomic status, environment); therefore, reasons for including race and ethnicity variables—as well as how these variables were assigned or categorized—must be explicitly stated. Further, all conceptually relevant factors (examples from Kaplan and Bennett include socioeconomic status, racism and discrimination, wealth, age, language, religion, health beliefs and practices, and environmental exposures) should be considered when interpreting racial and ethnic differences. Building on this, Kaplan and Bennett suggest every effort should be made to adjust for socioeconomic status and social class in analyses, which are the most common source of bias in racial and ethnic differences.

To our knowledge, this is the first study to evaluate BISG for use in traffic safety administrative data. However, there are several important limitations to note. First, the extent to which reported race/ethnicity is available within the NJ-SHO may vary for different populations. This is demonstrated by the higher proportion of female drivers who have a reported race/ethnicity value, which may be due to having an increased likelihood of linking with a hospital discharge record than male drivers because of childbirth. Additionally, the accuracy of BISG probabilities may decrease for some subpopulations in which changes in surname (e.g., females) and residential addresses are more common. This may influence the concordance between BISG produced estimates of race/ethnicity and reported values in our sample population. Concordance may also vary in populations with different distributions of race/ethnicity groups than NJ. Further work is needed to develop methods that better estimate race/ethnicity values for American Indian/Alaska Native and Multiracial populations (Elliott, 2009; Fremont et al. 2016), as BISG probabilities for these populations have consistently demonstrated poor concordance with reported values. Surname has been found to be a poor indicator of individuals who self-identify as Multiracial (Elliott, 2009) and in many communities across the US, individuals who self-identify as American Indian/Alaska Native are dispersed across geographic areas, reducing the usefulness of the census block groups in generating probabilities. There also remain lingering questions in the literature regarding the accuracy of reported race/ethnicity available from our original integrated data sources (e.g., death records), in particular for American Indian/Alaskan Native and Multiracial groups (Agency for Healthcare Research and Quality 2014). Due to the nature of how data are obtained and recorded in our integrated data sources, we are unable to directly examine the accuracy of race/ethnicity values recorded. Lastly, our analysis of crash rates among our racial/ethnic groups is intended to be an exemplar analysis of applying BISG with a limited dataset; therefore, the current crash rates presented should not be interpreted as real-world results. Regarding future directions, we plan to apply and further evaluate and then apply the BISG algorithm within the NJ-SHO to address novel research questions. More broadly, research is needed to directly compare reported race/ethnicity information across and within administrative data sources. Future work should also apply BISG and evaluate how it performs in other traffic safety databases and in identifying disparities in traffic safety outcomes.

To date, the majority of population-level traffic safety studies focused on identifying disparities have been limited to fatal crashes; incorporating race/ethnicity data into non-fatal traffic administrative data sources would undoubtedly catalyze the field's ability to identify, understand, and ameliorate disparities and inequities in transportation outcomes.

Our initial effort suggests that the BISG algorithm is a promising method to incorporate race and ethnicity for the majority of individuals in population-level crash, licensing, and other transportation databases via use of ubiquitously-collected name and address data, regardless of whether these data are linked to external sources with known race and ethnicity information. Further, taken with previous validation studies, our preliminary findings suggest applying BISG to traffic safety analyses may also reduce potential biases commonly cited in data collection and analysis, ultimately promoting more effective traffic safety interventions and equitable policies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors would first like to thank the RAND Corporation for sharing their BISG algorithm code with us. We also gratefully acknowledge Heather Griffis (PhD), Vicky Tam (MA), Oluwatimilehin (Timmy) Okunowo (MPH), and Christina Labows (BA) for their contributions to this project. This project was supported by funding from the Eunice K. Shriver National Institute of Child Health and Human Development (R21HD098276, PI: AEC and R21HD092850, PI: AEC).

DATA AVAILABILITY

The data that support the findings of this study are available from the NJ Motor Vehicle Commission (MVC), NJ Department of Transportation, and NJ Department of Health. Restrictions apply to the availability of these data, which were used under a Memorandum of Agreement and a Data Use Agreement for this study. Data may be available from the authors with the permission of these NJ governmental agencies, a Collaborative Research Agreement with the authors, and with certain restrictions.

REFERENCES

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):268–283. [PubMed: 23855558]

Agency for Healthcare Research and Quality. Racial Misclassification and Disparities in Mortality among AI/AN and Other Races, Washington. Healthc. Cost Util. Proj 2014. Available at: www.hcup-us.ahrq.gov/datainnovations/raceethnicitytoolkit/or26.jsp.

Arias E, Heron MP, Hakes JK. The validity of race and Hispanic origin reporting on death certificates in the United States: an update.; 2016.

Blumenberg E, Pierce G. A driving factor in mobility? Transportation's role in connecting subsidized housing and employment outcomes in the moving to opportunity (MTO) program. J. Am. Plan. Assoc 2014;80(1):52–66.

Briggs NC, Levine RS, Haliburton WP, Schlundt DG, Goldzweig I, Warren RC. The Fatality Analysis Reporting system as a tool for investigating racial and ethnic determinants of motor vehicle crash fatalities. Accid. Anal. Prev 2005;37:641–649. [PubMed: 15949455]

Centers for Disease Control and Prevention. Web-based Injury Statistics Query and Reporting System (WISQARS).; 2019. Available at: www.cdc.gov/ncipc/wisqars.

Curry AE, García-España JF, Winston FK, Ginsburg KR, Durbin DR. Variation in teen driver education by state requirements and sociodemographics. Pediatrics. 2012;129(3):453–457. [PubMed: 22331344]

Curry AE, Pfeiffer MR, Carey ME, Cook LJ. Catalyzing traffic safety advancements via data linkage: Development of the New Jersey Safety and Health Outcomes (NJ-SHO) data warehouse. Traffic Inj. Prev 2019;20(sup2):S151–S155. [PubMed: 31714800]

Curry AE, Pfeiffer MR, Metzger KB, Carey ME, Cook LJ. Development of the integrated New Jersey Safety and Health Outcomes (NJ-SHO) data warehouse: catalysing advancements in injury prevention research. Inj. Prev 2021;In press.

Elliott MN, Fremont AM, 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 P1):1722–1736. [PubMed: 18479410]

Elliott MN, Morrison PA, Fremont AM, McCaffrey DF, Pantoja P, Lurie N. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. Heal. Serv. Outcomes Res. Methodol 2009;9(2):69–83.

Flanagin A, Frey T, Christiansen SL, Bauchner H. The Reporting of Race and Ethnicity in Medical and Science Journals: Comments Invited. JAMA - J. Am. Med. Assoc 2021;325(11):1049–1052.

Flores G, Tomany-Korman SC. Racial and ethnic disparities in medical and dental health, access to care, and use of services in US children. Pediatrics. 2008;121(2):e286–e298. [PubMed: 18195000]

Fremont A, Weissman J, Hoch E, Elliott M. When Race/Ethnicity Data Are Lacking: Using Advanced Indirect Estimation Methods to Measure Disparities.; 2016.

Harper S, Charters TJ, Strumpf EC. Trends in socioeconomic inequalities in motor vehicle accident deaths in the United States, 1995–2010. Am. J. Epidemiol 2015;182(7):606–614. [PubMed: 26354899]

Hesketh M, Wuellner S, Robinson A, Adams D, Smith C, Bonauto D. Heat related illness among workers in Washington State: A descriptive study using workers' compensation claims, 2006–2017. Am. J. Ind. Med 2020;63(4):300–311. [PubMed: 31994776]

Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. John Wiley & Sons; 2013.

Kaplan JB, Bennett T. Use of Race and Ethnicity in Biomedical Publication. J. Am. Med. Assoc 2003;289(20):2709–2716.

Klinger EV, Carlini SV, Gonzalez I, Hubert SS, Linder JA, Rigotti NA, Kontos EZ, Park ER, Marinacci LX, Haas JS. Accuracy of race, ethnicity, and language preference in an electronic health record. J. Gen. Intern. Med 2015;30(6):719–723. [PubMed: 25527336]

Labgold K, Hamid S, Shah S, Gandhi NR, Chamberlain A, Khan F, Khan S, Smith S, Williams S, Lash TL, Collin LJ. Estimating the unknown: greater racial and ethnic disparities in COVID-19 burden after accounting for missing race and ethnicity data. Epidemiology. 2021;32(2):157–161. [PubMed: 33323745]

Li HR, Pickrell TM. Occupant Restraint Use in 2016: Results from the NOPUS Controlled Intersection Study. Washington, D.C., U.S.A.; 2018.

McAndrews C, Beyer K, Guse CE, Layde P. Revisiting exposure: fatal and non-fatal traffic injury risk across different populations of travelers in Wisconsin, 2001–2009. Accid. Anal. Prev 2013;60:103–112. [PubMed: 24036316]

Sanchez TW, Stolz R, Ma JS. Inequitable effects of transportation policies on minorities. Transp. Res. Rec 2004;1885(1):104–110.

Syed ST, Gerber BS, Sharp LK. Traveling towards disease: transportation barriers to health care access. J. Community Health 2013;38(5):976–993. [PubMed: 23543372]

Tefft BC, Williams AF, Grabowski JG. Driver licensing and reasons for delaying licensure among young adults ages 18–20, United States, 2012. Inj. Epidemiol 2014;1(1):1–8. [PubMed: 27747675]

West CN, Geiger AM, Greene SM, Harris EL, Liu IA, Barton MB, Elmore JG, Rolnick S, Nekhlyudov L, Altschuler A, Herrinton LJ, Fletcher SW, Emmons KM. Race and ethnicity: comparing medical records to self-reports. JNCI Monogr. 2005;2005(35):72–74.

Zhang Y, Lin G. Disparity surveillance of nonfatal motor vehicle crash injuries. Traffic Inj. Prev 2013;14(7):697–702. [PubMed: 23944196]

Table 1.

Distribution of race/ethnicity categories among NJ drivers using NJ-SHO reported data and BISG probabilities, overall and by sex, 2017

Race/ethnicity	Total (N=4,890,549)		Female (N=2,656,188)		Male (N=2,234,361)	
	NJ-SHO (%)	BISG (%)	NJ-SHO (%)	BISG (%)	NJ-SHO (%)	BISG (%)
White	64.6	67.2	63.5	66.7	66.1	67.8
Hispanic	17.3	14.1	17.5	14.0	17.0	14.3
Black	10.2	10.4	10.7	11.0	9.6	9.8
Asian/Pacific Islander	5.8	6.9	6.1	7.1	5.4	6.8
Multiracial	1.5	1.2	1.6	1.2	1.3	1.2
American Indian/Alaska Native	0.6	0.1	0.6	0.1	0.6	0.1

Abbreviations: NJ: New Jersey; NJ-SHO: New Jersey Safety and Health Outcomes; BISG: Bayesian Improved Surname Geocoding. *Note*: The 6 race/ethnicity categories are mutually exclusive. Categories presented in descending frequency.

Table 2.

Concordance statistics (area under the receiver operator curve [AUC]) for race/ethnicity of NJ drivers, comparing NJ-SHO reported data and BISG probabilities, overall and by sex, 2017.

Race/ethnicity	Total (N=4,890,549) AUC	Female (N=2,656,188) AUC	Male (N=2,234,361) AUC
Overall (weighted)	0.89	0.89	0.90
White	0.90	0.90	0.91
Hispanic	0.86	0.85	0.88
Black	0.94	0.94	0.95
Asian/Pacific Islander	0.90	0.89	0.92
Multiracial	0.65	0.65	0.65
American Indian/Alaska Native	0.52	0.52	0.53

Abbreviations: AUC: Area under the receiver operating curve; NJ: New Jersey; NJ-SHO: New Jersey Safety and Health Outcomes; BISG: Bayesian Improved Surname Geocoding.

Note: Darker shaded boxes indicate outstanding concordance (AUC ≥ 0.9); lighter shaded boxes indicate excellent concordance (0.8 ≤ AUC < 0.9).

Traffic Inj Prev. Author manuscript; available in PMC 2022 August 17.

Table 3.

Average monthly crash rates per 10,000 license-months among NJ drivers using NJ-SHO reported data and BISG probabilities for race/ethnicity, 2017

Race/ethnicity	NJ-SHO Rate (per 10k)	BISG Rate (per 10k)	Absolute Difference[a]	Change in Magnitude[b] (%)
White	44.3	43.6	0.7	2
Hispanic	70.1	73.1	3.0	4
Black	70.6	78.6	8.0	11
Asian/Pacific Islander	49.4	47.0	2.4	5
Multiracial	70.7	223.6	152.9	216
American Indian/Alaska Native	55.6	59.5	3.9	7

Abbreviations: NJ: New Jersey; NJ-SHO: New Jersey Safety and Health Outcomes; BISG: Bayesian Improved Surname Geocoding.

[a]Absolute difference equals the absolute value of (NJ-SHO rate minus BISG rate).

[b]Change in magnitude equals the absolute difference, divided by NJ-SHO rate, and then multiplied by 100.