NYC Health

Epi Research Report

October 2018

Evaluation of the Quality of Imputed Race and Ethnicity in New York City Administrative Hospitalization Data

Kathleen H. Reilly, Fangtao He, Tim Liao, Sungwoo Lim

Poor concordance of race and ethnicity has been reported for New York State hospitalization data compared with Medicaid and vital statistics data. Accuracy of race and ethnicity data is important for understanding health disparities. The objective of the current study is to determine whether imputation could be a feasible analytic option for improving validity of race and ethnicity variables in the Statewide Planning and Research Cooperative System (SPARCS).

Introduction

Numerous studies have reported poor quality of race and ethnicity information in administrative data.[1-3] In a report from the New York State Department of Health, race and ethnicity variables in SPARCS hospitalization data poorly corresponded to those from Vital Statistics and Medicaid records, with most New York City (NYC) facilities reporting race concordance below 70% and more than a third below 50%.[4, 5] Individual hospitals submit discharge data to SPARCS through a standardized submission system, however the means of collecting the data may differ by facility; therefore, the validity of race and ethnicity values may also vary by facility.[4] As administrative data become more widely available and used for public health program and policy evaluations, especially for understanding racial disparities in health care and health outcomes,[6–8] it is important to find ways to improve the quality of race and ethnicity variables.

Imputation, a technique to replace a missing data point with a probable value, has been empirically and theoretically proven as an effective method to handle missing data, and has been used by several federal agencies such as the U.S. Census Bureau and the Centers for Disease Control and Prevention to impute item nonresponse.[9–11]

Key Points:

- Imputation of race and ethnicity using neighborhood-level data such as United States (U.S.) Census Bureau data may not be able to accurately predict individual race and ethnicity.
- A thorough investigation of the underlying population characteristics should be performed before imputation in order to determine if individual-level inferences should be made.
- If good concordance between individual and neighborhood-level characteristics is observed and the existing data are known to be of poor quality, imputation using Census Bureau data could be considered as an analytic means to improve race and ethnicity information in administrative data for research and surveillance purposes.

1 Pan CX, Glynn RJ, Mogun H, Choodnovskiy I, Avorn J. Definition of race and ethnicity in older people in Medicare and Medicaid. *J Am Geriatr Soc.* 1999;47(6):730–733.
2 Hahn RA, Mulinare J, Teutsch SM. Inconsistencies in coding of race and ethnicity between birth and death in US infants: a new look at infant mortality, 1983 through 1985. *JAMA.* 1992;267(2):259–263.
3 Arday SL, Arday DR, Monroe S, Zhang J. HCFA's racial and ethnic data: current accuracy and recent improvements. *Health Care Financ Rev.* 2000;21(4):107–116.
4 New York State Department of Health. SPARCS - Facility Race/Ethnicity Concordance Reports. 2014; health.ny.gov/statistics/sparcs/reports/race_eth/. Accessed July 19, 2016.
5 Blustein J. The reliability of racial classifications in hospital discharge abstract data. *Am J Public Health.* 1994;84(6):1018–1021.
6 Lu MC, Halfon N. Racial and ethnic disparities in birth outcomes: a life-course perspective. *Matern Child Health J.* 2003;7(1):13–30.
7 Williams DR, Mohammed SA. Discrimination and racial disparities in health: evidence and needed research. *J Behav Med.* 2009;32(1):20–47.
8 Williams DR, Jackson PB. Social sources of racial disparities in health. *Health Aff (Millwood).* 2005;24(2):325–334.

Imputation reduces bias in estimation by preserving the original sample size and treating missing data consistently across all variables. Researchers have previously had success imputing better quality race values for large administrative data sets.[12,13]

To explore whether imputation is a feasible analytic option for improving the validity of race and ethnicity variables in SPARCS data, we conducted two analyses, each setting a random selection of race and ethnicity variables to missing, and examined the validity of imputed values.

Methods

The primary data source was SPARCS, a data system that includes administrative data for all hospital discharges occurring in New York State excluding federal and psychiatric facilities (i.e., an estimated 95% of all hospitalizations in the state).[14] Data elements include patient demographics, diagnoses, treatments, services and associated charges. These data are reported to SPARCS directly by each facility. U.S. Census 2007–2011 American Community Survey (ACS) data were used as an additional source to obtain NYC neighborhood-level predictors. Census data were aggregated by ZIP code tabulation area and matched with patients' residential ZIP codes from the SPARCS data.

Analysis 1

Records from three NYC hospitals (from upper Manhattan, the Bronx and Brooklyn) with good race and ethnicity concordance between SPARCS and Medicaid data and with patients' race and ethnicity distribution similar to that of their service areas were selected from 2012 SPARCS data. Hospitals with SPARCS and Medicaid concordance greater than 70% for race and greater than 90% for ethnicity were considered to have good concordance. To determine the hospital service area, we identified the ZIP codes accounting for 75% of ZIP codes of all patients discharged from these hospitals. The race and ethnicity population counts for the service areas of each hospital were obtained from 2010 U.S. Census data and summed according to each hospital's list of patient ZIP codes. The population counts of each ZIP code were weighted by the proportion of each ZIP code's hospital discharge counts among the total hospital discharges from its ZIP code list. The distribution of race and ethnicity for these hospitals is shown in Appendix 1. For these three hospitals, there was a total of 22,518 unique patients in 2012.

Analysis 2

SPARCS data from 2004 to 2010 were linked to data from the World Trade Center Health Registry (WTCHR), a cohort study of 71,430 people who were exposed to the September 11, 2001, World Trade Center terrorist attacks in NYC. The WTCHR data include three survey periods (2003 to 2004, 2006 to 2007 and 2011 to 2012).

The Registry's enrollment questionnaire included questions about race and ethnicity. Additional information on the WTCHR's methods are available elsewhere.[15] Personal identifying data from the WTCHR and SPARCS were linked electronically using an algorithm that included components of personal identifying information (name, date of birth, social security number and address). The matched data set included 17,165 hospitalizations among 8,858 WTCHR participants.

Variables

The dependent variables of this study were race and ethnicity.[16] Variables from SPARCS for the imputation models included age, sex, county of residence, Clinical Classifications Software (CCS) primary and secondary diagnosis codes, CCS procedure codes, length of stay, source of admission, discharge disposition, primary reimbursement and an indicator for whether the patient had been transferred from the emergency department.

To determine the service area for Analysis 1, variables from 2010 U.S. Census data included aggregated population counts and percentage of each race and ethnicity category for each ZIP code tabulation area.

Variables from the ACS included those related to neighborhood composition such as age, sex,

9 Lillard L, Smith JP, Welch F. What do we really know about wages? The importance of non-reporting and census imputation. *J Political Econ.* 1986:489– 506.

10 Centers for Disease Control and Prevention. Active bacterial core surveillance (ABCs) report emerging infections program network, Streptococcus pneumoniae, 2010. 2012; cdc.gov/abcs/reports-findings/survreports/spneu10.pdf. Accessed September 23, 2016.

11 Bureau of the Census. Current Population Survey Technical Paper 63RV Design and Methodology. 2002; census.gov/prod/2002pubs/tp63rv.pdf. Accessed September 23, 2016.

12 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):330–345.

13 Schenker N, Parker JD. From single-race reporting to multiple-race reporting: using imputation methods to bridge the transition. *Stat Med.* 2003;22(9):1571–1587.

14 New York State Department of Health. 2002 Annual report: the SPARCS data system. Albany, NY: New York State Department of Health;2002.

15 Farfel M, DiGrande L, Brackbill R, et al. An overview of 9/11 experiences and respiratory and mental health conditions among World Trade Center Health Registry enrollees. *J Urban Health.* 2008;85(6):880–909.

16 Race and ethnicity includes Black, White, Latino, Asian and other race. For the purpose of this report, Latino includes people of Hispanic, Latino or Spanish origin. SPARCS data capture this as "Spanish/Hispanic Origin"; Census and WTCHR data capture this as "Hispanic or Latino."

race, ethnicity, education, income, employment, birth place, languages spoken, migration, marital status, birth rate, military status, transportation, household characteristics, housing financing and property types. Analysis 1 used data from the 2008–2012 ACS and Analysis 2 used data from the 2007–2011 ACS. As a sensitivity analysis, imputation was performed with and without ACS data to determine if inclusion of ACS data improved the validity of the imputed values. For Analysis 2, race and ethnicity variables from the WTCHR were used to validate the imputation.

Statistical analysis

Race and ethnicity from SPARCS facilities with high concordance to Medicaid race and ethnicity data were considered the gold standard for Analysis 1, whereas self-reported race and ethnicity from the WTCHR survey were considered the gold standard for Analysis 2. For both analyses, we randomly selected 10% of hospitalizations with non-missing race and 10% of hospitalizations with non-missing ethnicity and set their race or ethnicity as missing. To impute missing race and ethnicity data, first we used a multivariate sequential regression method where joint conditional density of multiple variables with missing data given observed data was factored into an individual conditional density function for each variable. Next, this individual density was then modeled through a missing value that was drawn from a posterior predictive distribution through various regression models (e.g., logistic model for binary outcomes [ethnicity] and polytomous or generalized logit model for outcomes with multiple categories [race]).[17] Missing race and ethnicity were imputed in a randomly selected 10% sample of observations. Individual-level demographic and clinical characteristics and ZIP code-level neighborhood characteristics were used as auxiliary variables for the imputation. Kappa statistics (which measure interrater agreement) were calculated for the concordance of the gold standard race and ethnicity with imputed values for the 10% randomly selected hospitalizations. Kappa coefficients were calculated separately for race and ethnicity, as well as for a composite race and ethnicity variable, which considered all those with Latino ethnicity as Latino, regardless of race. Sensitivity and specificity were also calculated. A sensitivity analysis was conducted for Analysis 2 that excluded ACS variables to impute race and ethnicity. We performed imputation using the SAS-callable IVEware (Ann Arbor, Michigan, U.S.) and all other analyses were conducted using SAS 9.4 (Cary, North Carolina, U.S.).

This analysis, the WTCHR protocol, and the linkage between the WTCHR and SPARCS were approved by the New York City Department of Health and Mental Hygiene's institutional review board, and verbal consent was obtained from all WTCHR participants with permission to link to other data sources. The linkage between the WTCHR and SPARCS was approved by the New York State Department of Health SPARCS Data Protection Review Board. The New York State Department of Health approved the current study.

TABLE 1. Distribution of race and ethnicity among Analysis 1* and Analysis 2† patients, and the general population of New York City

Race and ethnicity‡	Analysis 1: N (%)	Analysis 2: N (%)	NYC residents (2010)§ N (%)
White	5,915 (26%)	8,094 (47%)	2,722,904 (33%)
Black	10,422 (46%)	3,592 (21%)	1,861,295 (23%)
Latino	4,378 (20%)	3,487 (20%)	2,336,076 (29%)
Asian	412 (2%)	652 (4%)	1,030,914 (13%)
Other	1,391 (6%)	1,340 (8%)	223,944 (3%)
Total	**22,518 (100%)**	**17,165 (100%)**	**8,175,133 (100%)**

* Analysis 1: 2012 hospitalizations for people at facilities with race and ethnicity distribution similar to that of their service areas (upper Manhattan, the Bronx and Brooklyn) and high concordance with Medicaid data. Sources: *Statewide Planning and Research Cooperative System (SPARCS); American Community Survey 2008–2012.*

† Analysis 2: participants in the World Trade Center Health Registry who were hospitalized between 2004 and 2010. Sources: *Statewide Planning and Research Cooperative System (SPARCS), 2012; World Trade Center Health Registry (WTCHR); American Community Survey 2007–2011.*

‡ Race and ethnicity includes Black, White, Latino, Asian and other race. For the purpose of this report, Latino includes people of Hispanic, Latino or Spanish origin. SPARCS data capture this as "Spanish/Hispanic Origin"; Census and WTCHR data capture this as "Hispanic or Latino."

§ Source: U.S. Census Bureau, 2010 and 2000 Census Public Law 94-171 Files and 1990 STF1 Population Division - NYC Department of City Planning (May 2011). www1.nyc.gov/assets/planning/download/pdf/data-maps/nyc-population/census2010/t_pl_p2a_nyc.pdf.

17 Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol.* 2001;27(1):85–96.

Results

Analysis 1

In 2012, there were 22,518 hospitalizations at the three selected hospitals. The race and ethnicity of the patients at these hospitals were not representative of all NYC residents (Table 1). Nearly half of the hospitalized patients were Black (n=10,422; 46%), less than one-third were White (n=5,915; 26%), one-fifth were Latino (n=4,378; 20%), and the remainder were Asian (n=412; 2%) or identified as other race (n=1,391; 6%) (Table 1).

Of the 22,518 hospitalizations, race and ethnicity data were imputed for the 10% randomly set to missing (2,283 hospitalized people and 2,339 hospitalized people, respectively). Table 2 depicts the relationship between imputed and actual race and ethnicity. Overall, the kappa coefficient of actual versus imputed race data was 0.37. Among 844 hospitalizations of White patients, 558 were predicted to be White using imputation (66%), while 679 out of 1,092 hospitalizations of Black patients (62%) were correctly imputed as Black race. For Asian patients, imputation had a very low sensitivity (i.e., only three cases were imputed as Asian patients). For these race groups except for Asians, the specificity was similar to the sensitivity (Whites: 78%; Blacks: 79%; Asians: 93%).

Imputation of Latino ethnicity had a kappa coefficient of 0.67. Of 465 hospitalizations among Latino patients, 240 (52%) were imputed as Latino, while 1,644 out of 1,874 hospitalizations for non-Latino patients (88%) were imputed correctly. Lastly, the kappa coefficient of the composite race and ethnicity variable was 0.59. The sensitivities for White and Black patients were 68% and 80%, respectively, while the specificities for these two groups were 88% and 90%.

The imputed race and ethnicity values derived from the data set used for the sensitivity analysis that did not use ACS data demonstrated slightly better validity than those from the data set that used ACS data for imputation. The kappa coefficients were 0.45 for imputed race, 0.92 for imputed ethnicity and 0.63 for imputed composite race and ethnicity (data not shown). Sensitivity was high for Black (75% and 87%) and Latino patients (73% and 95%), moderate for White patients (66% and 67%) and very low (4% and 45%) for Asian patients. Specificity was high for all groups, as follows: Asian (98% and 99%), Latino (93% and 98%), White (81% and 89%) and Black (76% and 87%).

TABLE 2. Kappa, sensitivity and specificity for race and ethnicity for 10% imputed sample from three New York City hospitals' SPARCS records (2012) matched to reported value**

	Race††	Ethnicity††	White		Black		Asian		Latino	
Analysis 1	**Kappa**	**Kappa**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
Race and ethnicity	0.37	0.67	66%	78%	62%	79%	7%	93%	52%	88%
Composite race and ethnicity variable	0.59		68%	88%	80%	90%	55%	97%	73%	94%

** Analysis 1: 2012 hospitalizations for people at facilities with race and ethnicity distribution similar to that of their service areas (upper Manhattan, the Bronx and Brooklyn) and high concordance with Medicaid data. Sources: *Statewide Planning and Research Cooperative System (SPARCS); American Community Survey 2008–2012.*

†† Race and ethnicity: includes White, Black, Latino, Asian and other race; For the purpose of this report, Latino includes people of Hispanic, Latino or Spanish origin. SPARCS data capture this as "Spanish/Hispanic Origin"; American Community Survey data capture this as "Hispanic or Latino."

Analysis 2

Included WTCHR participants were not representative of all NYC residents (Table 1), as participants were predominantly White and lived in lower Manhattan and Staten Island (data not shown). Sensitivity was highest for White race imputed using ACS data. Many of the Black, Latino and Asian WTCHR participants resided in neighborhoods that were racially diverse or predominantly White; neighborhood-level characteristics did not improve the prediction for these participants. Whereas 60% (5560/9196) of White participants lived in ZIP codes that were greater than 60% White, only 35% (1375/3893) of Black participants lived in ZIP codes that were greater than 60% Black, 6% of Latino participants lived in ZIP codes that were greater than 60% Latino, and 2% of Asian participants lived in ZIP codes that were greater than 60% Asian.

The kappa coefficient of self-reported versus imputed race and ethnicity was less than 0.4 for both 10% imputed sample data sets (Table 3). Sensitivity was reasonably high among White participants (72%), moderate for Black participants (53% and 56%), low for Latino participants (30% and 46%) and very low for Asian participants (8% and 12%). Specificity was highest for Asian (95% and 96%), Latino (87%) and Black (83% and 85%) participants, but moderate for White (65% and 71%) participants.

For the imputed race and ethnicity values from the data set used for the sensitivity analysis that did not use ACS data, all metrics demonstrated lower validity than those from the data set that used ACS data for imputation. The kappa coefficient of self-reported versus imputed race and ethnicity was less than or equal to 0.2 for all 10% imputed sample data sets (data not shown). Sensitivity was moderate among White participants (59% and 66%), low for Black participants (35% and 38%), low for Latino participants (22% and 40%) and very low for Asian participants (6% and 11%). Specificity was highest for Asian (96% and 97%), Latino (86% and 87%) and Black (77% and 80%) participants, but lower for White (55% and 59%) participants.

TABLE 3. Kappa, sensitivity and specificity for race and ethnicity for 10% imputed sample of New York City SPARCS hospitalization records (2004–2010) matched to World Trade Center Health Registry enrollees[‡]

	Race[§§]	Ethnicity[§§]	White		Black		Asian		Latino	
Analysis 2	**Kappa**	**Kappa**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
Race and ethnicity	0.30	0.35	72%	65%	53%	83%	12%	96%	46%	87%
Composite race and ethnicity variable	0.29		72%	71%	56%	85%	8%	95%	30%	87%

[‡] Analysis 2: participants in the World Trade Center Health Registry who were hospitalized between 2004 and 2010. Sources: *Statewide Planning and Research Cooperative System (SPARCS), 2012; World Trade Center Health Registry (WTCHR); American Community Survey 2007–2011.*

[§§] Race and ethnicity: includes White, Black, Latino, Asian and other race; For the purpose of this report, Latino includes people of Hispanic, Latino or Spanish origin. SPARCS data capture this as "Spanish/Hispanic Origin"; American Community Survey and WTCHR data capture this as "Hispanic or Latino."

Discussion

We found that validity of the imputed data varied by race. When patients' race and ethnicity, as reported in SPARCS, was similar to the predominant race and ethnicity of the hospital service area, imputed values were quite consistent with the SPARCS values.

For Analysis 1, imputed race was less than the "good" threshold of 70%, but much better for patients with White and Black race than for Asian race. We believe that this result reflects the extent to which each race group is represented in the neighborhoods served by the three selected hospitals. When imputing Asian race using neighborhood-level characteristics, the small sample size for Asian New Yorkers (2%) in these three hospitals limited the sensitivity of the imputation.

For Analysis 2, imputation for White race using the WTCHR population performed better than those for other race groups and Latino ethnicity, and was above the 70% "good" threshold. For those who are multiracial, racial identification may be contextual and depend on the gender, socioeconomic status and religion of the individual, and racial identity may vary over time.[18,19] A study validating Veterans Affairs administrative data with self-report survey data similarly found that agreement was higher for White and Black participants compared with those of other races, and those who self-reported multiple races were more likely to be misclassified.[7] Racial misclassification in administrative data may be more common for those who are not White or Black if the data are not self-reported.[20]

Poor performance in Analysis 2 using WTHCR data could be attributed to lack of generalizability to the NYC population, specifically for those who are not White. Because Black, Latino and Asian participants in the WTCHR population tended to live in racially and ethnically diverse or predominantly White neighborhoods, the performance of imputation was poor. Similarly, a race validation study of women in North Carolina found that surrogating census block-level race information from the U.S. Census was only successful for the race that was the larger proportion of the population (White) and not the race that was the smaller proportion of the population (Black).[21]

This finding demonstrates how the feasibility of imputing race for a data set depends on the variables used for the imputation, and highlights the importance of having an understanding of the underlying characteristics of the population and how covariates chosen for the imputation could have an effect on the imputed values.

It is likely that the imputation performed better for White participants since the WTCHR population has a high proportion of White (54%) participants, and most White participants lived in predominantly White neighborhoods. On the other hand, Black, Latino and Asian participants among the WTCHR population tend to live in more racially diverse or predominantly White neighborhoods; using neighborhood characteristics for these race and ethnic groups resulted in poorer predictions for Black and Asian races and Latino ethnicity.

Imputation of race and ethnicity using neighborhood-level predictors may not be appropriate for individuals living in racially and ethnically diverse neighborhoods.

18 Davenport LD. The Role of Gender, Class, and Religion in Biracial Americans' Racial Labeling Decisions. Am Sociol Rev. 2016;81(1):57–84.
19 Doyle JM, Kao G. Are racial identities of multi-racials stable? Changing self-identification among single and multiple race individuals. Soc Psychol Q. 2007;70(4):405–423.
20 Edgar HJ, Daneshvari S, Harris EF, Kroth PJ. Inter-observer agreement on subjects' race and race-informative characteristics. PLoS ONE. 2011;6(8):e23986.

Imputation of race and ethnicity using ACS data or other aggregate neighborhood data source may not be appropriate for those living in racially and ethnically mixed neighborhoods because neighborhood-level racial and ethnic distribution may not be able to predict individual race and ethnicity.

For Analysis 1, data were not validated with external data, whereas for Analysis 2, external data were self-reported, and thus more accurate. Even though the three selected NYC hospitals for Analysis 1 were based on the New York State concordance report, which was validated using Medicaid data,4 the validity of race and ethnicity in the Medicaid data may not be as good as that of self-reported data.

The current study demonstrates that imputation of race and ethnicity using neighborhood-level predictors may not be appropriate for individuals living in racially and ethnically diverse neighborhoods because neighborhood-level racial and ethnic distribution may not be able to predict individual race and ethnicity. A thorough understanding of the race and ethnicity data, the neighborhood population and important limitations of imputation in this context are needed before imputation in order to determine if individual-level inferences should be made. If good concordance between individual and neighborhood-level characteristics is observed, imputation could be considered as an analytic means to improve race and ethnicity information for research and surveillance purposes in administrative data with known poor-quality race and ethnicity information.

Suggested citation: Reilly KH, He F, Liao T, Lim S. Evaluation of the quality of imputed race and ethnicity in New York City administrative hospitalization data. New York City Department of Health and Mental Hygiene: Epi Research Report, October 2018; 1-8.

Acknowledgments: The authors acknowledge Regina Zimmerman, Sara Archie, Hannah Gould, Hannah Jordan, Cheryl Stein, Charon Gwynn and James Hadler of the NYC DOHMH for reviewing previous drafts of this report.

21 Kwok RK, Yankaskas BC. The use of census data for determining race and education as SES indicators: a validation study. *Ann Epidemiol.* 2001;11(3):171–177.


NYC
Health