Race and Ethnicity Modeling Applied to Linked Heath Care Data

Authors: Lisa B. Mirel, Dean Resnick, Jessie L. Parker, Cindy Zhang, and Christine Cox

Background

The National Center for Health Statistics (NCHS) serves as the nation's principal health statistics agency, whose mission is to provide statistical information that can be used to guide actions and policies to improve the health of the American people. NCHS conducts several population-based and health care surveys designed to collect important information about the health of the U.S. population. Through the NCHS Data Linkage Program, data from these surveys are linked to mortality data from the National Death Index (NDI) and other health related administrative data sources.[1] These data linkages are based on both deterministic and probabilistic linkage algorithms, which rely on the exchange and comparison of personally identifiable information (PII) between data sources. The linked data expand the scientific utility of surveys and enable richer analysis than would be possible with each data source alone. The NCHS linked data resources have supported over 1,000 PubMed-indexed scientific publications.[2]

One source that NCHS has used in its linkage program is the National Hospital Care Survey (NHCS)[3]. The NHCS is a provider survey that collects inpatient, emergency department (ED), and outpatient department episode-level data from sampled hospitals. The patient hospital records collected as part of the NHCS were frequently missing race and ethnicity. A recent linkage of the 2016 NHCS to the National Death Index (NDI)[4] highlights an example where a linked administrative source augmented race and ethnicity information that was frequently missing from NHCS patient encounter records. However, since race and ethnicity data obtained from the linked NDI data were only available for patients who died, critically important data gaps remained for researchers wanting to assess mortality rates by race and ethnicity since the denominator would need to include both those who died and who remained alive. To expand the research capabilities of the NHCS, along with its linked data resources, we modeled missing NHCS patient race and ethnicity data using name and geographic race and ethnicity frequencies from 2010 Census data, based on previous methodology. Since neither the NHCS nor the linked NHCS–NDI mortality data include respondent-reported race/ethnicity, we utilized the respondent-reported race and ethnicity information collected in the 2018 National Health Interview Survey (NHIS)[5] to validate the imputation methodology. This paper will describe the imputation methodology and the evaluation methods used to assess imputed race and ethnicity and will demonstrate the potential of this imputation to further health equity research goals.

Materials and Methods

Description of Data Sources

National Hospital Care Survey (NHCS)

The goal of NHCS is to provide reliable and timely healthcare utilization data for hospital-based settings, including prevalence of conditions, health status of patients, health services utilization, and substance-involved ED visits. From participating hospitals, NHCS collects data on all inpatient and ambulatory care visits occurring during the calendar year. The target universe for NHCS is all inpatient discharges and in-person ambulatory care visits in noninstitutional, non-federal hospitals in the 50 states and the District of Columbia that have 6 or more staffed inpatient beds. The patient records collected in the NHCS include patient PII (e.g., name, date of birth (DOB), and Social Security Number (SSN)), which allow for the linkage of episodes of care across hospital units as well as to other data sources, such as the NDI. The 2016 NHCS is not nationally representative due to low response rates (27%), with 158 responding hospitals from the 581 sample[6]. Still, linking NHCS with the NDI does allow for new analyses, such as studying mortality post hospital discharge, along with specific causes of death. The linkage described here includes only patients with at least one inpatient or ED visit reported by hospitals participating in the 2016 NHCS. Less than 1 percent of NHCS records that were eligible for linkage are missing values for name, state of residence, sex, or date of birth.[7]

National Death Index (NDI)

The NDI is a centralized database of United States death record information on file in state vital statistics offices. Working with these state offices, NCHS established the NDI as a resource to aid epidemiologists and other health and medical investigators with their mortality ascertainment activities. The NDI became operational in 1981 and includes death record information for persons who have died in the U.S. or a U.S. territory from 1979 onward. The records, which are compiled annually, include detailed information on the underlying and multiple causes of death. For this analysis, the 2016-2017 NDI records were used.

National Health Interview Survey (NHIS)

NHIS is a nationally representative, cross-sectional household interview survey that serves as an important source of information on the health of the civilian, noninstitutionalized population of the U.S. It is a multistage sample survey with primary sampling units of counties or adjacent counties, secondary sampling units of clusters of houses, tertiary sampling units of households, persons within households, and finally one selected sample adult and sample child. It has been conducted continuously since 1957 and the content of the survey is periodically updated.

Imputation Methodology

The model we developed to conduct these imputations is based on the work described in Elliot, et al.[8] Others have used this methodology as well[9,10]. The model is known as the Bayesian Surname and Geocoding (BSG) method and motivated the development of our enhanced model. With the BSG model, the prior race and ethnicity probabilities are based on geocoded (Census block group) race and ethnicity distributions. The BSG model refines the probabilities (applying Bayes' Theorem) based on whether the last names are on Asian or Hispanic surname lists. In Elliot's application, the posterior distributions are computed based on estimated sensitivity and specificity of Asian and Hispanic- identified last names included on these lists.

Similar to BSG, our model leverages race and ethnicity associations of last names to refine race and ethnicity proportions (i.e., priors) derived from geography. The geocoded Census block[i], rather than block group as in the BSG model, is used to obtain the 2010 decennial census race ethnicity proportions[11]. Our work is different from the BSG approach in three notable ways. First, we use information for first names to make Bayesian adjustments to block-level race and ethnicity frequencies (whereas in BSG they are solely based on last name). Second, we do not adjust these frequencies simply based on whether the names are included on lists but instead use the proportion of persons of each race and ethnicity having these given names. Third, our model does not use estimates of sensitivity and specificity to estimate posteriors but instead is based directly on name-race and ethnicity frequencies.

For the imputation, we classified race and ethnicity as follows (note, these categories are treated as being exclusive, our categorization did not account for multiple race and ethnicity identities):

- Hispanic (*Note: This takes precedence over race. Persons described as Hispanic are not assigned a race group*)
- White (Non-Hispanic)
- Black (Non-Hispanic)
- Asian or Pacific Islander (API, Non-Hispanic)
- American Indian or Alaskan Native (AIAN, Non-Hispanic)

We then generated 5 prior estimates based on the residence block's race/ethnicity distribution:

- P_{B_1} ~ Proportion of block residents who are Hispanic
- P_{B_2} ~ Proportion of block residents who are White, Non-Hispanic

[i] Census blocks, are the smallest geographic area for which the Bureau of the Census collects and tabulates decennial census data. (What are census blocks? (census.gov))

- P_{B_3} ~ Proportion of block residents who are Black, Non-Hispanic
- P_{B_4} ~ Proportion of block residents who are API, Non-Hispanic on the block
- P_{B_5} ~ Proportion of block residents who are AIAN, Non-Hispanic on the block

In cases where block cannot be determined we substitute $P(Race = R_i) \rightarrow P_{B_i}$, where $P(Race = R_i)$ is the estimated U.S. resident population proportion for Race i.

These priors are then adjusted by application of Bayes law using the national name-level proportions for both first and last name for each of these race and ethnicity groupings. Data on name frequency are obtained from two sources:

- For last names, the U.S. Census Bureau tabulation of last names by reported race and ethnicity from 2010 Decennial Census enumeration data was used[12]. For last names reported for 100 or more persons in the Census, tabulations were made by respondent-reported race. By inverting the data on this file, we computed the proportion of each race and ethnicity group members having each surname.
- For first names, we used a race and ethnicity tabulation[13] developed by the Office of the Comptroller of the Currency[ii], using proprietary data from mortgage loan applications from 2007 and 2010 submitted under the requirements of the Home Mortgage Disclosure Act. [iii] We computed the proportion of each race/ethnicity group members having each first name.

So, if someone is named ‘Michael’, the proportions of interest would be…

- Among Hispanic individuals, the proportion which have the first name ‘Michael’
  - $P_{FN_1} = P(FN = \text{‘}Michael\text{’} \,|Hispanic\ and\ any\ race)$

[ii] The OCC is an independent bureau of the U.S. Department of the Treasury. (Organization | OCC (treas.gov))

[iii] Names with proportions based on fewer than 30 observations are not reported on separately, except when all records for a given first name show the same race/ethnicity category -- in which case the threshold for reporting is lowered from 30 to 15 records. The total number of records in the OCC data is close to 2.6 million. These frequencies may be biased by their source as mortgage loan applicants may have a different race /ethnicity distribution than the U.S. population.

- Among White, Non-Hispanic individuals, the proportion which have the first name ‘Michael’
  - $P_{FN_2} = P(FN\ =\ \text{‘}Michael\text{’}\ |White,\ non - Hispanic)$
- Among Black, Non-Hispanic individuals, the proportion which have the first name ‘Michael’
  - $P_{FN_3} = P(FN\ =\ \text{‘}Michael\text{’}\ |Black, non - Hispanic)$
- Among API, Non-Hispanic individuals, the proportion which have the first name ‘Michael’
  - $P_{FN_4} = P(FN\ =\ \text{‘}Michael\text{’}\ |API, non - Hispanic)$
- Among AIAN, Non-Hispanic individuals, the proportion which have the first name ‘Michael’
  - $P_{FN_5} = P(FN\ =\ \text{‘}Michael\text{’}\ |AIAN, non - Hispanic)$

Similarly, if someone has the last name ‘Johnson’ we would compute the proportion for each of the race and/ethnicity groupings: $P_{LN_1} \dots P_{LN_5}$. Last names with population frequencies too small (i.e., < 100) to allow race and ethnicity tabulation are assigned to the category *Other-Name*, and the computed proportion for each of the race and ethnicity grouping among persons falling in the Other-Name group is used.[iv]

Thus, the model would estimate the relative proportion (RP) (compared to values for other race and ethnic groups) of people on the block with the name ‘Michael Johnson’ as (assuming distributional independence) as

$$RP_i \sim P_{B_i} \cdot P_{FN_i} \cdot P_{LN_i} \qquad \textit{(Eq. 1)}$$

Thus, for a given person record, Eq. 1 will generate an expected count of same-named persons on that person’s reported block of residence. Then we compute the posterior distribution across the groups as

$$\mathrm{P(Race\ =\ }R_i\ |\ \mathrm{First\ Name\ =\ }FN\ \mathrm{and\ Last\ Name\ =\ }LN) = \widehat{P_{r_\imath}} = \frac{RP_i}{\sum_{j=1}^{5} RP_j}$$

[iv] Note that generally that among Hispanic and White, Non-Hispanic populations, the proportion falling into catch-all Other Name group is substantially smaller than among members of the other categories.

We then make the $Imputed\ Race = \max\,(P_{r_i})$ and estimate $P(Imputed\ Race = R) =\ P_{r_{\widehat{Imputed\ Race}}}$. Note that the latter is an estimate of the precision of the imputation.

In cases where either first or last name is missing (in contrast to names with low frequencies), we substitute the U.S. proportions for these: $P(Race = R_i) \rightarrow P_{LN_i}$.

One potential strategy to improve the accuracy of race and ethnicity group statistics is to use only records for which race and ethnicity imputation has a high probability of being correct. Because we estimate precision for each imputation, it is feasible to set a threshold for every imputation to increase the overall level of precision for imputed race and ethnicity assignment: $P(Imputed\ Race = R) > t$, where t is an analyst set threshold.

Evaluation

For this evaluation we use the respondent-report of race and ethnicity identity as the gold standard. To assess the validity of the imputation model, we tested the methodology by comparing imputed race and ethnicity to 2018 NHIS respondent-reported values. To evaluate the concordance of respondent reported and imputed values, we assess results using two scenarios, one for the whole set of NHIS records and one for those imputed records with estimated precision greater than 80%. The purpose of this evaluation was to assess how well the imputation performed against a gold standard source of respondent-reported data and not to validate the hospitalization data against the NHIS.

Cross Tabulation

We provide summaries of cross-classification of respondent-reported race/ethnicity to modeled race and ethnicity by means of 2 x 2 cross-tabulations for each race and ethnicity group. Positive predictive value (true positive/(true positive + false positive)) and negative predicative value (true negative/(true negative + false positive) are based on unweighted counts and are presented for each 2 x 2 cross tabulation.

Race and Ethnicity Category Alignment

The categories of respondent-reported race and ethnicity provided on the 2018 NHIS public-use data files do not perfectly align with those used for modeling, which conform to the categorizations used in the input sources. While the categories Hispanic, non-Hispanic White, and non-Hispanic Black are essentially identical in the two classification schemes, the NHIS public use data categorizes the remaining persons as either Asian or Other, and the model categorizes the remaining persons as either "Asian or Pacific Islander (API)" or "American Indian or Alaskan Native (AIAN)". Here the difference is that Pacific Islanders are joined with AIAN in the NHIS public use file reporting, but they are joined with Asians in the model classification (this was necessary because of how name/race tallies and Census tabulations were provided). Because the U.S. Pacific Islander population is relatively small compared to both the Asian and AIAN populations (1.4 million Pacific Islander[14] compared to 18.9 million Asian[15] and 5.7 million AIAN[16]), this incongruence only moderately affects these comparisons.

Because NHIS restricted-use data does allow race and ethnicity categorization that is consistent with that used in the model, a sensitivity analysis was performed using the restricted-use data which confirmed that the misalignment does not substantially affect the evaluation of concordance of race and ethnicity assignment between reported and modeled (data not shown).

Next, for each of the five mutually exclusive race and ethnicity categories, binary indicator variables were created (one for reported race and ethnicity and one for modeled race and ethnicity):

- 1 = Yes, person has been assigned to this group
- 0 = No, person has **not** been assigned to this group

Cohen's Kappa Statistics and Pearson Correlations

The similarity of these assignments was evaluated by computing Cohen's kappa coefficient for each of the five race/ethnicity categories for males and females separately, and for three age groups: <18, 18-64, and 65+.These coefficients were evaluated according to the recommended scale interpretations developed by J. Richard Landis and Gary G. Koch[17]:

- Kappa ≤ 0 → no agreement
- 0 < Kappa ≤ 0.20 → slight agreement
- 0.21 ≤ Kappa ≤ 0.40 → fair agreement
- 0.41 ≤ Kappa ≤ 0.60 → moderate agreement
- 0.61 ≤ Kappa ≤ 0.80 → substantial agreement
- 0.81≤ Kappa ≤ 1 → almost perfect agreement.

Demonstration of Imputation with NHCS Data

To demonstrate the feasibility of using the imputed race and ethnicity values with linked data, we compute mortality rates by race and ethnicity, age group, and sex within 0-30, 31- 60 and 61- 90 days of hospital discharge from the 2016 NHCS linked to the NDI. This demonstration was limited to linkage-eligible NHCS patients. Information on the linkage of NHCS patients to the NDI has been described elsewhere.[18] As nearly 70% of the patient records collected in the 2016 NHCS are missing race and ethnicity data, these statistics could not be reliably produced without the use of the imputed values. In addition, although counts of deaths can be examined by race and ethnicity from the linked death records, it is not possible to accurately assess mortality rates for analyses which require the race and ethnicity of non-deceased persons in the denominator without imputation. The mortality rates below are shown by age, sex, and imputed race and ethnicity. It should be noted that we do not show AIAN-identified patients in this tabulation due to disclosure and precision issues related to small sample size.

Results

Cross Tabulations

The following tables include the cross-tabulations of race and ethnicity groups by imputed vs. reported values for NHIS 2018 (first for all survey respondents and second for those with estimated imputation precision > 80%).

Table 1a. Comparison of respondent-reported Hispanic ethnicity to imputed Hispanic ethnicity, 2018 National Health Interview Survey

	Respondent reported: Hispanic ethnicity		
Imputed: Hispanic ethnicity	Yes	No	Total
All NHIS respondents			
Yes………………………………………………….	9,651	1,179	10,830
No……………………………………………………	2,830	59,171	62,001
Total………………………………………………	12,481	60,350	72,831
NHIS respondents with imputation precision > 80%			
Yes………………………………………………….	8,644	535	9,179
No……………………………………………………	1,708	48,871	50,579
Total………………………………………………	10,352	49,406	59,758

All NHIS respondents

Positive predictive value: 89.1% (9,651/10,830)

Negative predictive value: 95.4% (59,171/62,001)

NHIS respondents with imputation precision > 80%

Positive predictive value: 94.2% (8,644/9,179)

Negative predictive value: 96.6% (48,871/50,579)

Table 1b. Comparison of respondent-reported non-Hispanic Black race/ethnicity to imputed non-Hispanic Black race/ethnicity, 2018 National Health Interview Survey

	Respondent reported: non-Hispanic Black		
Imputed: non-Hispanic Black	Yes	No	Total
All NHIS respondents			
Yes………………………………………………	5,753	2,224	7,977
No……………………………………………….	2,734	62,120	64,854
Total……………………………………………	8,487	64,344	72,831
NHIS respondents with imputation precision > 80%			
Yes………………………………………………	4,342	606	4,948
No……………………………………………….	1,387	53,423	54,810
Total……………………………………………	5,729	54,029	59,758

All NHIS respondents

Positive predictive value: 72.1% (5,753/7,977)

Negative predictive value: 95.8% (62,120/64,854)

NHIS respondents with imputation precision > 80%

Positive predictive value: 87.8% (4,342/4,948)

Negative predictive value: 97.5% (53,423/54,810)

Table 1c. Comparison of respondent-reported non-Hispanic White race/ethnicity to imputed non-Hispanic White race/ethnicity, 2018 National Health Interview Survey

	Respondent reported: non-Hispanic White		
Imputed: non-Hispanic White	Yes	No	Total
All NHIS respondents			
Yes………………………………………………	43,275	6,268	49,543
No…………………………………………………	2,947	20,341	23,288
Total……………………………………………	46,222	26,609	72,831
NHIS respondents with imputation precision > 80%			
Yes………………………………………………	38,599	3,906	42,505
No…………………………………………………	906	16,347	17,253
Total……………………………………………	39,505	20,253	59,758

All NHIS respondents

Positive predictive value: 87.3% (43,275/49,543)

Negative predictive value: 87.3% (20,341/23,288)

NHIS respondents with imputation precision > 80%

Positive predictive value: 90.8% (38,599/42,505)

Negative predictive value: 94.7% (16,347/17,253)

Table 1d. Comparison of respondent-reported non-Hispanic Asian race/ethnicity to imputed non-Hispanic Asian and Pacific Islander (API) race/ethnicity, 2018 National Health Interview Survey

	Respondent reported: non-Hispanic Asian		
Imputed: non-Hispanic API	Yes	No	Total
All NHIS respondents			
Yes………………………………………………..	2,910	1,216	4,126
No…………………………………………………	1,545	67,160	68,705
Total………………………………………………	4,455	68,376	72,831
NHIS respondents with imputation precision > 80%			
Yes………………………………………………..	2,495	460	2,955
No…………………………………………………	923	55,880	56,803
Total………………………………………………	3,418	56,340	59,758

All NHIS respondents

Positive predictive value: 70.5% (2,910/4,126)

Negative predictive value: 97.8% (67,160/68,705)

NHIS respondents with imputation precision > 80%

Positive predictive value: 84.4% (2,495/2,955)

Negative predictive value: 98.4% (55,880/56,803)

NOTE: The NHIS respondent reported race and ethnicity category of non-Hispanic Asian does not include respondents who reported race and ethnicity as Pacific Islander.

Table 1e. Comparison of respondent-reported non-Hispanic Other race/ethnicity to imputed non-Hispanic American Indian or Alaskan Native (AIAN) race/ethnicity, 2018 National Health Interview Survey

	Respondent reported: non-Hispanic Other		
Imputed: non-Hispanic AIAN	Yes	No	Total
All NHIS respondents			
Yes…………………………………………………	192	146	338
No……………………………………………………	994	71,499	72,493
Total………………………………………………	1,186	71,645	72,831
NHIS respondents with imputation precision > 80%			
Yes…………………………………………………	141	30	171
No……………………………………………………	613	58,974	59,587
Total………………………………………………	754	59,004	59,758

All NHIS respondents

Positive predictive value: 56.8% (192/338)
Negative predictive value: 98.6% (71,499/72,493)

NHIS respondents with imputation precision > 80%

Positive predictive value: 82.5% (141/171)
Negative predictive value: 99.0% (58,974/59,587)

NOTE: The NHIS respondent reported race and ethnicity category of non-Hispanic Other includes persons who reported race and ethnicity as Pacific Islander.

Across all comparisons, there was variation in the agreement of the race and ethnicity imputation and respondent report. Hispanic persons had the highest positive predictive value (89.1%), and non-Hispanic AIAN persons had the highest negative predictive value (99.0%). When limiting results to modeled race and ethnicity assignments with precision over 80%, all race and ethnicity groups have positive predicted values over 84% and negative predictive values over 95% except for non-Hispanic AIAN persons (positive predictive value of 56.8%). The model tends to over impute non-Hispanic White persons and under impute the remaining groups (particularly non-Hispanic AIAN persons). Likely this is due to the population size of non-Hispanic White persons and imputation assignment being given, for each survey respondent, to the category with the highest probability.

Cohen's Kappa Statistics

Table 2a and 2b provide an assessment of agreement between respondent-reported and imputed race and ethnicity by sex and age using Cohen's Kappa statistics; first among all survey respondents and then among respondents with estimated imputation precision > 80%.

Table 2a. Cohen's Kappa statistics comparing respondent-reported to modeled race and ethnicity group assignments for all NHIS 2018 respondents

	All			Age	Age	Age
Race/Ethnicity	Respondents	Female	Male	< 18	18-64	65+
Hispanic	0.80	0.78	0.82	0.77	0.80	0.82
Non-Hispanic Black	0.66	0.68	0.64	0.60	0.67	0.71
Non-Hispanic White	0.72	0.71	0.73	0.66	0.73	0.76
Non-Hispanic Asian	0.66	0.64	0.68	0.59	0.68	0.69
Non-Hispanic Other	0.25	0.25	0.24	0.30	0.21	0.26

NOTE: NHIS 2018 Public Use Files, n=72,831

Table 2b. Cohen's Kappa statistics comparing respondent-reported to modeled race and ethnicity group assignments for NHIS 2018 respondents with estimated race and ethnicity imputation precision > 80%

	All			Age	Age	Age
Race/Ethnicity	Respondents	Female	Male	< 18	18-64	65+
Hispanic	0.86	0.85	0.88	0.83	0.87	0.90
Non-Hispanic Black	0.80	0.82	0.77	0.75	0.80	0.86
Non-Hispanic White	0.81	0.81	0.82	0.76	0.82	0.87
Non-Hispanic Asian	0.77	0.76	0.79	0.72	0.78	0.79
Non-Hispanic Other	0.30	0.32	0.28	0.38	0.26	0.29

NOTE: NHIS 2018 Public Use Files, n=59,758

Based on Landis-Koch assessment schema for Cohen's Kappa statistic, for respondents with estimated imputation precision > 80%, the agreement of modeled race and ethnicity assignment is either substantial or almost perfect except for non-Hispanic Other (where it is only fair). We assessed if the low

Cohen’s Kappa was due to misalignment based on the restricted data for NHIS. When using the same categories as the model assignment the Cohen’s Kappa Statistics changed from 0.77 for non-Hispanic Asian to 0.79 for non-Hispanic API and from 0.30 for non-Hispanic Other to 0.40 for non-Hispanic AIAN. Agreement is particularly good for assignment to Hispanic ethnicity (Cohen’s Kappa Statistic>=0.83 (almost perfect) for those with a precision of > 80%). Differences between agreement level by sex are small, but among age groups, agreement is generally better for the older respondents (except among non-Hispanic Other where cell sizes are small).

Results for Demonstration with NHCS Data

Tables 3a and 3b illustrate the mortality rates for 0-30, 31-60, 61-90 days post final hospital discharge by imputed race and ethnicity category and age group and sex for all linkage-eligible 2016 NHCS patients and for linkage-eligible 2016 NHCS patients with estimated race and ethnicity imputation greater than 80%.

Table 3a. 2016 NHCS mortality rates by time after final hospital discharge for linkage-eligible NHCS 2016 patients by age, sex and imputed race and ethnicity

		Mortality Rates Percentages								
		Age Category								
		18-39 years			40-64 years			65+ years		
		0-30 days	31-60 days	61-90 days	0-30 days	31-60 days	61-90 days	0-30 days	31-60 days	61-90 days
Race/Eth	Sex									
All	*Female*	0.19	0.24	0.27	1.69	2.02	2.29	6.82	8.35	9.42
	Male	0.53	0.63	0.71	2.59	3.12	3.53	9.07	10.96	12.28
Hispanic	*Female*	0.15	0.18	0.20	1.14	1.35	1.53	5.38	6.58	7.41
	Male	0.42	0.51	0.56	2.12	2.54	2.82	7.67	9.12	10.32
Non-Hispanic Black	*Female*	0.22	0.26	0.29	1.69	2.04	2.31	6.91	8.47	9.42
	Male	0.56	0.67	0.74	2.51	3.07	3.46	8.81	10.71	12.01
Non-Hispanic White	*Female*	0.20	0.25	0.28	1.80	2.16	2.44	6.98	8.56	9.68
	Male	0.56	0.66	0.74	2.69	3.25	3.67	9.26	11.21	12.56
Non-Hispanic API	*Female*	0.11	0.12	0.15	1.70	2.04	2.27	7.04	8.06	8.80
	Male	0.61	0.71	0.76	2.68	3.50	3.96	8.97	10.70	11.62

NOTE: n=3,744,405. The small number of 2016 NHCS patients imputed to non-Hispanic, AIAN are not included in this tabulation.

Table 3b. 2016 NHCS mortality rates by time after final discharge for linkage-eligible NHCS patients by age, sex and imputed race and ethnicity with precision > 80%

		Mortality Rates Percentages								
		Age Category								
		18-39 years			40-64 years			65+ years		
		0-30 days	31-60 days	61-90 days	0-30 days	31-60 days	61-90 days	0-30 days	31-60 days	61-90 days
Race/Eth	Sex									
All	*Female*	0.21	0.26	0.29	1.74	2.08	2.35	6.78	8.31	9.40
	Male	0.55	0.66	0.74	2.59	3.13	3.54	9.04	10.92	12.23
Hispanic	*Female*	0.18	0.21	0.22	1.12	1.32	1.51	4.93	6.04	6.84
	Male	0.43	0.52	0.57	2.11	2.53	2.79	7.36	8.65	9.77
Non-Hispanic Black	*Female*	0.22	0.27	0.30	1.77	2.15	2.38	6.96	8.38	9.37
	Male	0.59	0.71	0.80	2.47	3.04	3.46	8.54	10.46	11.81
Non-Hispanic White	*Female*	0.21	0.27	0.30	1.83	2.18	2.47	6.93	8.51	9.64
	Male	0.57	0.68	0.76	2.69	3.24	3.68	9.22	11.16	12.50
Non-Hispanic API	*Female*	0.18	0.18	0.20	1.78	2.34	2.67	7.16	8.23	8.98
	Male	*1.11*	*1.29*	*1.3*	*2.85*	*4.14*	*4.48*	*10.29*	*12.00*	*12.65*

NOTE: n=2,382,229. The small number of 2016 NHCS patients imputed to Non-Hispanic, AIAN are not included in this tabulation.

As noted in Tables 3a and 3b, the mortality rates appear to vary for some race and ethnicity groups, depending on the level of precision used for the imputation. For example, non-Hispanic API males have a mortality rate of about 0.61% in 0-30 days after hospital discharge, but when we limited the results to just those with a precision of greater than 80%, the mortality rate for that same time frame was 1.11%. In addition, Tables 3a and 3b, illustrate variation in mortality rates among race and ethnicity groups that could not be determined before the use of imputation. This could be important if death rates differed by race and ethnicity within a certain sex. For example, the mortality rate for males aged 18-39 years within 0-30 days of final hospital discharge is 0.55%, however, for non-Hispanic API males in that same age group and timeframe from final hospital discharge (using the precision of greater than 80%) is 1.11%.

Conclusion

NCHS population and provider surveys are designed to collect information about the health of the U.S. population and are used to inform scientific research and health policy. NCHS conducts linkages of these survey data to health-related administrative data sources to expand the scientific utility of surveys and enable richer analysis than would be possible with each data source alone.

This paper highlights the value of leveraging linked survey data to demonstrate the potential enhancements of imputing race and ethnicity information. The NHCS collects patient encounter records from participating hospitals to provide reliable and timely healthcare utilization data for hospital-based settings. However, patient hospital records collected as part of the NHCS were frequently missing race and ethnicity, a key health-related covariate. This research demonstrates that it is possible to reliably impute such information using Bayesian techniques applied to data obtained from other sources. As an added benefit, the imputation strategy employed here is relatively straightforward to apply and uses publicly available sources to develop the race and ethnicity distributions. Finally, this work demonstrates the importance of applying appropriate statistical techniques to impute critically important health information to enable further study of the role of race and ethnicity in health outcomes. More robust statistical analyses enabled through race and ethnicity imputation is an important step in supporting a wide variety of health equity research goals.

[1] https://www.cdc.gov/nchs/data-linkage/index.htm

[2] https://www.cdc.gov/nchs/data/datalinkage/LinkedMortalityFilesCitationList_508.pdf

[3] https://www.cdc.gov/nchs/nhcs/index.htm

[4] https://www.cdc.gov/nchs/ndi/index.htm

[5] https://www.cdc.gov/nchs/nhis/index.htm

[6] https://www.cdc.gov/nchs/data/datalinkage/NHCS16-NDI16-17-Methodology-Analytic-Consider.pdf

[7] https://www.cdc.gov/nchs/data/datalinkage/NHCS16-NDI16-17-Methodology-Analytic-Consider.pdf

[8] Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., & Lurie, N. (2008). A new method for estimating race/ethnicity and associated disparities where administrative records lack . respondent-reported race/ethnicity. *Health services research*, *43*(5p1), 1722-1736.

[9] 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 Jun;70(3):330-45. doi: 10.1177/1077558712466293. Epub 2012 Nov 20. PMID: 23169896.

[10] Grundmeier RW, Song L, Ramos MJ, Fiks AG, Elliott MN, Fremont A, Pace W, Wasserman RC, Localio R. Imputing Missing Race/Ethnicity in Pediatric Electronic Health Records: Reducing Bias with Use of U.S. Census Location and Surname Data. Health Serv Res. 2015 Aug;50(4):946-60. doi: 10.1111/1475-6773.12295. Epub 2015 Mar 11. PMID: 25759144; PMCID: PMC4545341.

[11] ***Documentation:*** https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.2010.html

Data: https://www2.census.gov/census_2010/redistricting_file--pl_94-171/

[12] Comenetz, J. (2016). Frequently occurring surnames in the 2010 census. *United States Census Bureau*. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

[13] Tzioumis, K. (2018). Demographic aspects of first names. *Scientific data*, *5*(1), 1-9. https://www.nature.com/articles/sdata201825

[14] Native Hawaiian/Other Pacific Islander - The Office of Minority Health (hhs.gov)

[15] Asian American - The Office of Minority Health (hhs.gov)

[16] American Indian/Alaska Native - The Office of Minority Health (hhs.gov)

[17] Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data*". Biometrics.* 33 *(1): 159–174.*

[18] https://www.cdc.gov/nchs/data/datalinkage/NHCS16-NDI16-17-Methodology-Analytic-Consider.pdf