Office of Health Equity
VDH VIRGINIA DEPARTMENT OF HEALTH

August 26, 2020

Imputing Missing Race and Ethnicity Data in COVID-19 Cases

Health equity is a cornerstone of public health. This is especially true during the COVID-19 pandemic when many populations are suffering from both the health and economic consequences of the disease. Good information on disparities in disease incidence, outcomes, and social and economic consequences, is necessary to guide and develop an appropriate response. However, efforts to study these disparities have been hampered by missing data. Almost a quarter of confirmed cases are missing race and ethnicity data. Accounting for this missing data is essential to understanding COVID-19 and to facilitate research into health disparities. Social Epidemiologists from the Office of Health Equity used imputation techniques to estimate race and ethnicity for cases missing that data.

The methodology, along with the results of the first run, are summarized below. These will be updated with more recent data shortly, and monthly thereafter. This data is intended for research purposes and to assist understanding of COVID-19-related health disparities. The COVID-19 Daily Dashboard will continue to report the unimputed data, including the number of cases not reported. Virginia Department of Health Surveillance and Investigations staff continue to pursue multiple strategies to fill in missing data.

Methodology

The **Bayesian Improved Surname Geocoding (BISG)** methodology, developed by the RAND Corporation, has been used by federal agencies, including the Centers for Medicare and Medicaid Services, to account for missing race and ethnicity data. Additionally, the Consumer Financial Protection Bureau uses BISG to address discrimination in the credit industry. BISG uses known case data, including name and geography, to impute missing race and ethnicity data. Reviews from literature indicate that the BISG proxy probability is more accurate than other

Rexford Anson-Dwamena
Priya Pattah
Justin Crow

methodologies like geography-only or surname-only proxy in its ability to predict individual reported race and ethnicity.

The US Census Bureau compiles a list of the surnames reported at least 100 times in each decennial Census, stratified by race and ethnicity. It also provides detailed information on the racial and ethnic composition of geographic areas, at various levels down to the block level. BISG uses these datasets, matched to individual patient names and locations, to estimate the probability individual cases belong to each racial or ethnic category. For purposes of this study, individuals with missing data are assigned the race or ethnicity with the highest probability.

Process

The following are the step by step process for processing the data processing.

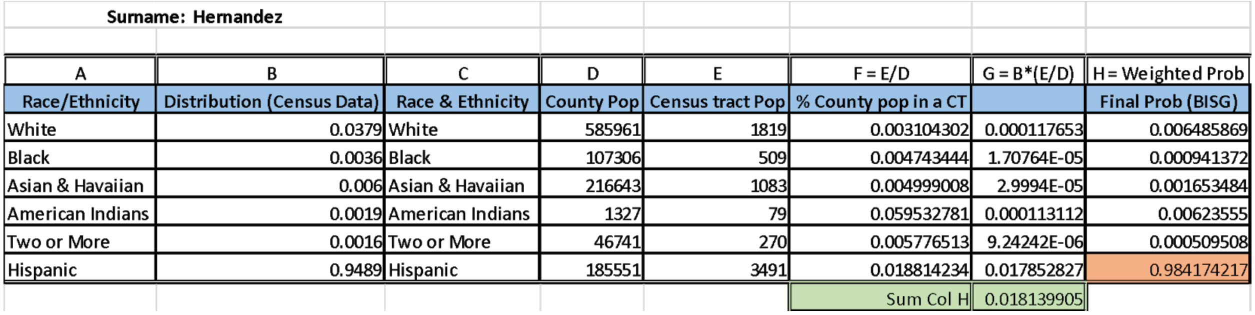
- Applicants' surnames are standardized by removing any special characters such as JR and SR
- Standardized surnames are matched to the census surname list
- For each name that matches the census surname list, the probability of belonging to a given racial or ethnic group is constructed.
- Applicant address information is standardized in preparation for geocoding
- Addresses are mapped into census geographic areas such as census tract
- For geocoded addresses, the proportion of the county adult population for each race and ethnicity residing in the geographic area containing the address or associated with the census tract is calculated.
- Bayes Theorem is used to update the surname-based probabilities constructed

COVID-19 data pulled on June 28th had 61,434 records of which 18, 354 had missing race and ethnicity variables (29%). Using the Bayesian Improved Surname Geocoding (BISG) methodology, over 16,300 matched the Census Surname list (88.9 %). The remaining 2,054 Surnames did not match the U. S Census Surname list. This is because, census publishes or provides each surname held by at least 100 enumerated individuals, along with a breakdown of the percentage of individuals with that name belonging to one of six race and ethnicity categories: Hispanic; non-Hispanic White; non-Hispanic Black or African American; non-Hispanic Asian/Pacific Islander; non-Hispanic American Indian and Alaska Native; and non-Hispanic Multiracial. In total, the list provides 151,671 surnames, covering approximately 90% of the U.S. population.

- Calculate the probability of belonging to race or ethnicity r (for each of the six race and ethnicity categories) for a given surname s.

- Calculate the proportion of the population of individuals in race or ethnicity r (for each of the six race and ethnicity categories) that lives in geographic area g.

- Apply Bayes' Theorem to calculate the likelihood that an individual with surname s living in geographic area g belongs to race or ethnicity r. See Example below for Surname "Hernandez"



Surname: Hernandez

A	B	C	D	E	F = E/D	G = B*(E/D)	H = Weighted Prob
Race/Ethnicity	Distribution (Census Data)	Race & Ethnicity	County Pop	Census tract Pop	% County pop in a CT		Final Prob (BISG)
White	0.0379	White	585961	1819	0.003104302	0.000117653	0.006485869
Black	0.0036	Black	107306	509	0.004743444	1.70764E-05	0.000941372
Asian & Havaiian	0.006	Asian & Havaiian	216643	1083	0.004999008	2.9994E-05	0.001653484
American Indians	0.0019	American Indians	1327	79	0.059532781	0.000113112	0.00623555
Two or More	0.0016	Two or More	46741	270	0.005776513	9.24242E-06	0.000509508
Hispanic	0.9489	Hispanic	185551	3491	0.018814234	0.017852827	0.984174217
					Sum Col H	0.018139905	

Using SPSS Modeler, The data was partitioned into 3 sets: 60% into training, 20% into testing, and 20% into validation. The result showed that using the Census Bureau's prior probability together with neighborhood racial and ethnicity composition at the census tract level, we could predict with 99.5% certainty one's race or ethnicity. Imputed race and ethnicity data is linked to COVID-19 case data for further analysis.

Surnames that were not matched with census surname list were linked with auxiliary neighborhood (Census tract) variables like poverty, proportion of race and ethnicity, unemployment, education, etc., to help impute the race and ethnicity using the known race and ethnicity data with the same neighborhood characteristics.

Results

Race

The BISG was trained on case data pulled on June 28, 2020. It imputed race for the 27% of cases missing race data. The greater majority of missing cases were imputed as White, increasing the share significantly from 47% to 59%. In the final dataset, 42% of White cases are imputed. Shares for all other races declined. Notably, 20% of Asian Americans and Pacific Islander cases are imputed in the final dataset, while only 10 of 11,378 Two or More races cases are imputed.

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Race	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
White	20,378	47%	35,085	59%	14,707	42%
Black	9,023	21%	10,062	17%	1,039	10%
Asian or Pacific Islander	2,185	5%	2,723	5%	538	20%
Native American	201	0%	207	0%	6	3%
Other Race or 2+ Races	11,293	26%	11,303	19%	10	0%
Total	43,080	100%	59,380	100%	16,300	27%

Ethnicity

The BISG imputed race for the 27% of cases missing race data. More cases were imputed as Latino, increasing the share from 40 to 45% of cases. In the final dataset, 35% of Latino cases are imputed compared to 21% of Not Latino cases.

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Ethnicity	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
Latino	17,383	40%	26,723	45%	9,340	35%
Not Latino	25,697	60%	32,657	55%	6,960	21%
Total	43,080	100%	59,380	100%	16,300	27%

Regional Results

Central Virginia had the highest share of cases imputed, at just over a third, followed by Northern (29%), Southwestern (27%), Northwestern (21%) and Eastern (18%). Correspondingly, Central and Northern Virginia saw some of the largest shifts in case share. In Central Virginia, the share of cases attributed to White residents increased from 38% to 55%, while the share for Black residents decreased from 40% to 30%. Cases among Latinos increased from 21% to 29%. In Northern Virginia, the share of cases attributed to White residents also increased, from 43% to 57%, but the complementary decrease was more evenly divided among other races.

Central Health Region

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Race	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
White	2,729	38%	5,986	55%	3,257	54%
Black	2,918	40%	3,329	30%	411	12%
Asian or Pacific Islander	162	2%	218	2%	56	26%
Native American	19	0%	21	0%	2	10%
Other Race or 2+ Races	1,392	19%	1,396	13%	4	0%
Total	7,220	100%	10,950	100%	3,730	34%

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Ethnicity	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
Latino	1,523	21%	3,131	29%	1,608	51%
Not Latino	5,697	79%	7,819	71%	2,122	27%
Total	7,220	100%	10,950	100%	3,730	34%

Eastern Health Region

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Race	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
White	2,480	42%	3,525	49%	1,045	30%
Black	2,956	50%	3,173	44%	217	7%
Asian or Pacific Islander	149	3%	159	2%	10	6%
Native American	25	0%	27	0%	2	7%
Other Race or 2+ Races	308	5%	309	4%	1	0%
Total	5,918	0%	7,193	100%	1,275	18%

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Ethnicity	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
Latino	781	13%	1,068	15%	287	27%
Not Latino	5,137	87%	6,125	85%	988	16%
Total	5,918	100%	7,193	100%	1,275	18%

Northern Health Region

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Race	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
White	9,146	43%	17,073	57%	7,927	46%
Black	2,166	10%	2,472	8%	306	12%
Asian or Pacific Islander	1,749	8%	2,183	7%	434	20%
Native American	145	1%	147	0%	2	1%
Other Race or 2+ Races	8,046	38%	8,050	27%	4	0%
Total	21,252	0%	29,925	100%	8,673	29%

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Ethnicity	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
Latino	12,031	57%	18,305	61%	6,274	34%
Not Latino	9,221	43%	11,620	39%	2,399	21%
Total	21,252	100%	29,925	100%	8,673	29%

Northwestern Health Region

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Race	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
White	4,481	67%	6,265	74%	1,784	28%
Black	705	11%	770	9%	65	8%
Asian or Pacific Islander	105	2%	133	2%	28	21%
Native American	11	0%	11	0%	0	0%
Other Race or 2+ Races	1,338	20%	1,338	16%	0	0%
Total	6,640	0%	8,517	100%	1,877	22%

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Ethnicity	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
Latino	2,602	39%	3,539	42%	937	26%
Not Latino	4,038	61%	4,978	58%	940	19%
Total	6,640	100%	8,517	100%	1,877	22%

Southwestern Health Region

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Race	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
White	1,542	75%	2,236	80%	694	31%
Black	278	14%	318	11%	40	13%
Asian or Pacific Islander	20	1%	30	1%	10	33%
Native American	1	0%	1	0%	0	0%
Other Race or 2+ Races	209	10%	210	8%	1	0%
Total	2,050	0%	2,795	100%	745	27%

As of June 28, 2020	Original Data		Final Data		Cases Imputed	
Ethnicity	**Count**	**Share of Total**	**Count**	**Share of Total**	**Count**	**Share Imputed**
Latino	446	22%	680	24%	234	34%
Not Latino	1604	78%	2,115	76%	511	24%
Total	2,050	100%	2,795	100%	745	27%

References

Elliott, M. N. 2009. "Use of Indirect Measures of Race/Ethnicity to Target Disparities" [accessed on October 18, 2011].

Elliott, M. N., A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie. 2008. "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity." Health Services

Fiscella, K., and A. M. Fremont. 2006. "Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity." Health Services Research 41 (4 Pt 1): 1482–500.

Gazmararian, J., R. Carreon, N. Olson, and B. Lardy. 2012. "Exploring Health Plan Perspectives in Collecting and Using Data on Race, Ethnicity, and Language." American Journal of Managed Care 18 (7): e254–61.