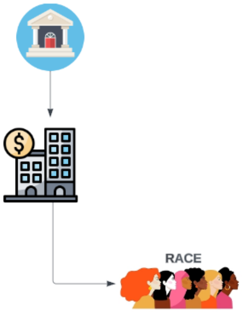
A novel approach on race prediction: Nested dichotomies applied to BISG

Ana Patrón - Piñerez
Supervised by: Arthur Charpentier Agathe Fernandes Machado

July 24, 2024

Regulation and fairness

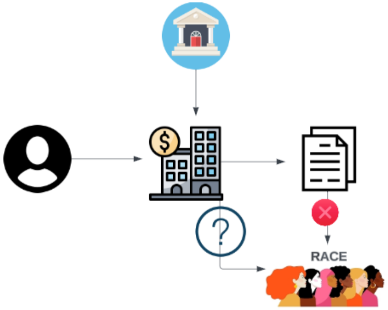
- Colorado SB21-169: The legislation holds insurers accountable for testing their big data systems - including external consumer data and information sources, algorithms, and predictive models - to ensure they are not unfairly discriminating against consumers on the basis of a protected class



Regulation and fairness

Race as a protected variable:

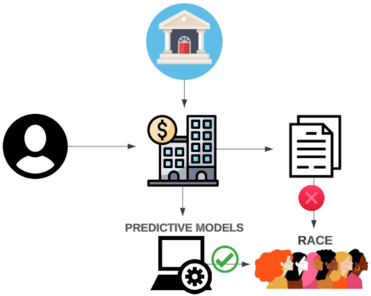
- Civil Rights Act of 1866, 1964 prohibited discrimination based on "race, color or previous condition of servitude"
- In property and casualty (P&C) insurance, race and ethnicity data has not been systematically collected (American Academy of Actuaries, 2022)
- In health insurance, race and ethnicity data are often incomplete and inconsistent (Haley et al. (2022)).



Regulation and fairness

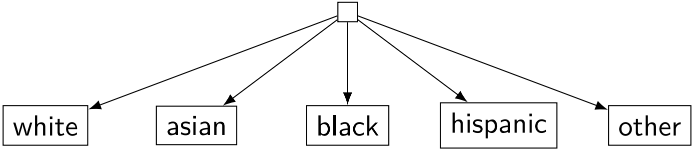
Race as a protected variable:

- statistical methods for imputing or modeling race and ethnicity were in life and health insurance Larry Baeder and Woldeyes (2024).



Setup

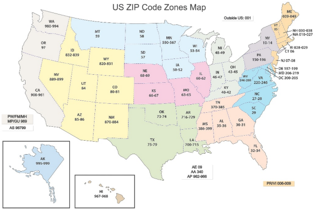
Let i denotes the i-th observation. The goal is to find the probability of individual i belonging to each of the races.



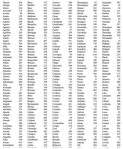
We calculate proxies:
$p(R_i = r_i | G_i = g_i), p(R_i = r_i | S_i = s_i), p(R_i = r_i | G_i = g_i, S_i = s_i), \ldots.$

- R_i: race $\in \{\text{white}, \text{black}, \text{hispanic}, \text{asian}, \text{other}\}$
- S_i: surname from a list of surnames
- F_i: first name from a list of first names
- G_i: geolocation that can be at tract, block, block group, county, place or zcta level.

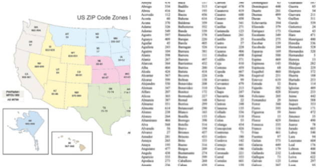
Pre-Bayesian Methods



Geocoding Only (GO): P&C insurance in the 1990s and 2000s (NAIC, 2008)



Surname analysis (SA): Spanish surname lists (Word & Perkins Jr., 1996). Asian surname lists (Lauderdale & Kestenbaum, 2000)



Categorical Surname and Geocoding (CSG) I. SA for Asian and Hispanic, II. GO Black or white/other

Bayesian methods

Bayesian Surname Geocoding (BSG):

- Integrated cohort distributions by surname and geolocation from different datasets using Bayes's theorem (Elliott et al. (2008))

$$p(R_i|S_i) = \frac{p(S_i|R_i)\ p(R_i)}{p(S_i|R_i)p(R_i) + p(S_i|not\ R_i)p(not\ R_i)}$$

- where $p(R_i), p(not\ R_i)$ are the prior probabilities of belonging and not belonging to a specific race/ethnicity cohort based solely on geolocation, respectively.
- $p(S_i|R_i)$, $p(S_i|not\ R_i)$ are computed depending on lists of Asian or Hispanic surnames (for more information see the appendix)

Bayesian Methods

Bayesian Improved Surname Geocoding (BISG):

- Different surname data (U.S. Census Bureau of 2010, lists for all the races) and conditions the prior probability of race/ethnicity on surname instead of geolocation (Elliott et al. (2009))

$$p(R_i|G_i, S_i) = \frac{p(R_i|S_i)\ p(G_i|R_i)}{\sum_{r\in R} p(R_i|S_i)\ p(G_i|R_i)} \tag{1}$$

Intuition behind (1):

I. Independence assumption

- Given the race, the geolocation is not informative about the surname and viceversa.

$$G_i \perp\!\!\!\perp S_i | R_i \qquad (Assumption \quad 1)$$

II. General properties

- From Bayes formula: $p(R_i, G_i) = p(R_i|G_i)\, p(G_i)$
- Properties of joint distribution: $p(R_i, G_i, S_i) = p(G_i|R_i, S_i)\, p(R_i|S_i)\, p(G_i)$
- Law of total probability : $p(G_i, S_i) = \sum_{r \in R} p(R_i, G_i, S_i)$

Intuition behind (1):

$$p(R_i|G_i,S_i) = \frac{p(R_i,G_i,S_i)}{p(G_i,S_i)}$$
$$= \frac{p(G_i|R_i,S_i)\;p(R_i|S_i)\;p(G_i)}{p(G_i)\sum_{r\in R} p(G_i|R_i,S_i)\;p(R_i|S_i)}$$

By assumption 1, we arrive to:

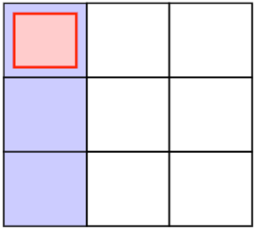
$$p(R_i|G_i,S_i) = \frac{p(R_i|S_i)\;p(G_i|R_i)}{\sum_{r\in R} p(R_i|S_i)\;p(G_i|R_i)} \tag{2}$$

Intuition behind (1):

Figure 1: $p(R_i|S_i)$ obtained from US Census Surname List 2010

surname	p_whi	p_bla	p_his	p_asi	p_oth
SMITH	0.7090	0.2311	0.0240	0.0050	0.0308
JOHNSON	0.5897	0.3463	0.0236	0.0054	0.0350
WILLIAMS	0.4575	0.4768	0.0249	0.0046	0.0363
BROWN	0.5795	0.3560	0.0252	0.0051	0.0342
JONES	0.5519	0.3848	0.0229	0.0044	0.0361
GARCIA	0.0538	0.0045	0.9203	0.0141	0.0073
MILLER	0.8411	0.1076	0.0217	0.0054	0.0243
DAVIS	0.6220	0.3160	0.0244	0.0049	0.0327
RODRIGUEZ	0.0475	0.0054	0.9377	0.0057	0.0036
MARTINEZ	0.0528	0.0049	0.9291	0.0060	0.0073
HERNANDEZ	0.0379	0.0036	0.9489	0.0060	0.0035
LOPEZ	0.0486	0.0057	0.9292	0.0102	0.0063
GONZALEZ	0.0403	0.0035	0.9497	0.0038	0.0027

Figure 2: $p(G_i|R_i)$ is the racial composition of each geolocation. We apply bayes $\frac{p(R_i|G_i)\,p(G_i)}{p(R_i)}^*$, obtained from US Census 2010



$$*\frac{p(R_i|G_i)\,p(G_i)}{p(R_i)} = \frac{\frac{\#\text{ counts for race r in geolocation g}}{\#\text{ counts for geolocation g}}\frac{\#\text{ counts for geolocation g}}{\#\text{ total observations}}}{\frac{\#\text{ total counts for race r}}{\#\text{ total observations}}} = \frac{\#\text{ counts for race r in geolocation g}}{\#\text{ total counts for race r}}$$

Summarizing

Algorithm 1 BISG

Input surname list and census counts

for do $R_i \in \mathcal{R}$

Select from the voters file $p(R_i|S_i)$

Compute from census $p(G_i|R_i)$.

Calculate $p(R_i|S_i, G_i) \leftarrow p(R_i|S_i) * p(G_i|R_i)$

Normalize $p(R_i|S_i, G_i) \leftarrow \frac{p(R_i|S_i,G_i)}{\sum_{R\in\mathcal{R}} p(R_i|S_i,G_i)}$

end for

Output vector of probabilities $(p(R_i|S_i, G_i))_{R_i\in\mathcal{R}}$

To evaluate the performance of the model, you should have a dataset with self-reported race, S, F, G from voters file or healthcare

Bayesian Improved First Surname Geocoding (BIFSG)

We also assume that once we know the race, the geolocation is not informative about the first name and viceversa.

$$G_i \perp\!\!\!\perp F_i | R_i \qquad (Assumption \quad 1)$$

Thus

$$p(R_i|G_i, F_i, S_i) = \frac{p(S_i|R_i)\; p(F_i|R_i)\; p(R_i|G_i)}{\sum_{r \in R} p(S_i|R_i)\; p(F_i|R_i)\; p(R_i|G_i)}$$

Variable importance

"*As can be seen, half of the total predictive power of BISG is unique to surnames, about a quarter is unique to location. As expected, these proportions vary strongly by race/ethnicity, with surnames alone responsible for only 33% of BISG's predictive power for Blacks.*" Elliott et al. (2009)

Algorithm/Cohort	white	black	asian	hispanic	other	overall
SA	0.95	0.09	0.56	0.84	0.01	0.75
BISG	-0.02	+0.41	+0.03	+0.02	+0.05	+0.04
BIFSG	+0.03	+0.04	-0.02	-0.03	+0.08	+0.02

Table 1: Differences in accuracy compared to the methodology above (increasing number of explanatory variables). Examples for Asian individuals: Yu Kyle, Pham Sam.

↪ Optimizing the BISG methodology with the variables considered (F, S, G): using first names (F) exclusively for identifying White and Black individuals.

Bayesian methods

Fully Bayesian Improved Surname Geocoding (fBISG)

BISG suffers from two data problems regarding minorities:

- the census often contains zero counts
  ↪ fBISG uses a measurement error model so that zero values mean low probability instead of nonexistence
- many surnames are missing from the census data
  ↪ fBISG also supplemens the surname list with additional data from voter files from six Southern states

fBISG: Methodology

BISG (Elliott et al., 2009)

$$P(R_i|S_i, G_i) \propto P(S_i|R_i)P(R_i|G_i)$$

$P(R_i = r|G_i = g) \propto N_{rg}$, obtained from US **census data**. [a]

[a] https://www.census.gov/data.html

fBISG (Imai and Khanna, 2016)

$$P(R_i|S_i, G_i) \propto P(R_i|S_i)P(G_i|R_i)$$

$P(R_i = r|G_i = g, R_{-i}) \propto n_{rg}^{-i} + N_{rg} + 1 > 0$, with:

- the term $+1$ arises from the assumption of a Dirichlet prior distribution over the race distribution for geolocation g,
- n_{rg}^{-i} is obtained using Gibbs sampling (Robert and Casella, 1999) on the dataset of individuals whose race is being predicted, by conditioning on the race of other individuals R_{-i} in geolocation g.

fBISG: Results

AUC BY METHODOLOGY

	Area under ROC				
	Hispanic	**Asian**	**Black**	**White**	**Other**
BISG	0.92	0.82	0.92	0.90	0.59
fBISG with zero-count correction	0.96	0.91	0.94	0.91	0.57
fBISG with additional surname data	0.96	0.91	0.96	0.91	0.58
fBISG with first name	0.97	0.93	0.97	0.94	0.61
fBISG with first and middle name	0.98	0.94	0.98	0.95	0.62

Source: (Imai, Olivella, & Rosenman, 2022).

But...

- Minorities continue to be underestimated. They are absorbed by the majority
- How can we give more power to the minorities?

white	**black**	**asian**	**hispanic**	**other**
0.57	0.11	0.05	0.17	0.09

Table 2: Proportion of races in the Census decennial of 2020 at tract level, US

Motivation:

surname	first	middle	race
Pacheco	Emalee	Julie	Hispanic

bisg white	bisg hispanic	bisg asian	bisg black	bisg other
0.50	0.47	0.001	0.007	0.012

Table 3: Example of probabilities marginally distant. Evaluated on data from an insurance application - SOA

white	non white
0.43	0.57

Table 4: Example of binomial prediction

Motivation

- Mistakes are highly punished, a single mistake along the path to a leaf node results in an incorrect prediction. Frank and Kramer (2004)
- Our hypothesis: relaxed metrics show significant improvements. In some cases probabilities are marginally distant, and the model is penalized.
- We evaluate Recall (identification) and Precision (the model is sharp)

Table 5: Performance metrics BISG. Taking two highest probabilities (taking highest probability)

Cohort/Metric	Recall	Precision
white	0.96 (0.93)	0.87 (0.79)
black	0.68 (0.47)	0.76 (0.66)
asian	0.65 (0.57)	0.82 (0.81)
hispanic	0.88 (0.84)	0.88 (0.88)
other	0.14 (0.03)	0.45 (0.36)

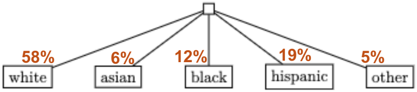


Figure 3: Multiclass configuration

Balanced problems →

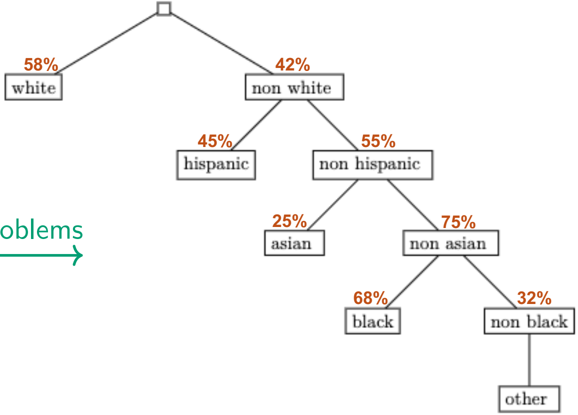
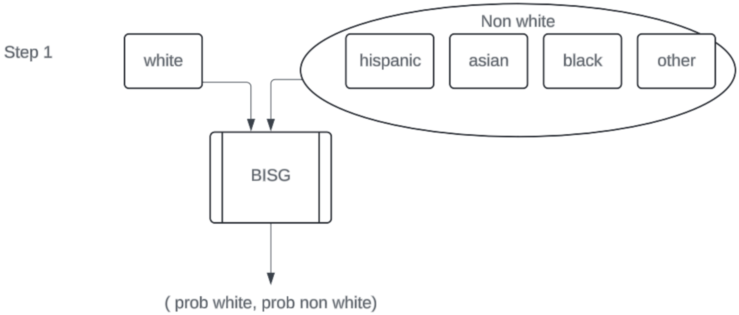
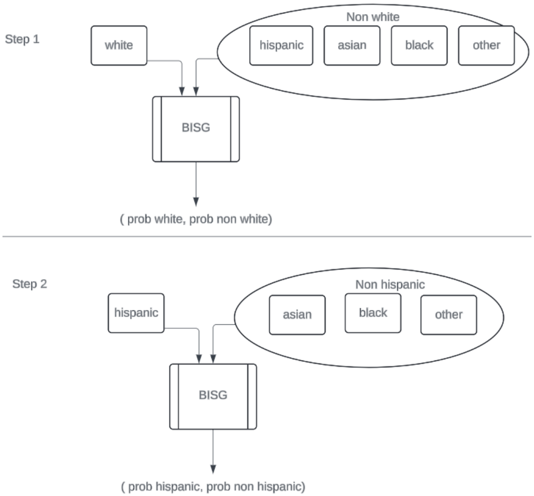


Figure 4: Nested dichotomies configuration

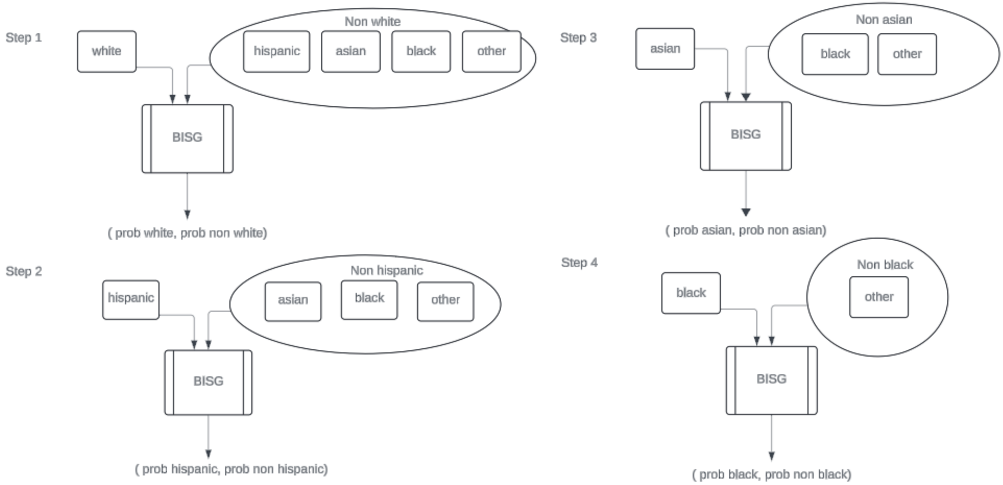
Nested dichotomies



Nested dichotomies



Nested dichotomies



Regular approach (R)

$$p(R_i = r | G_i = g, S_i = s) = \prod_{k \in \mathcal{P}_r} (\mathbb{I}(r \in \mathcal{R}_{k1}) p(r$$
$$+ (\mathbb{I}(r \in \mathcal{R}_{k2}) p(r \in \mathcal{R}_{k2} | G_i, S_i, R_i \in \mathcal{R}_k)$$

- $\mathcal{P}_r$: path to the leaf node corresponding to class r
- $\mathbb{I}$: indicator function
- $\mathcal{R}_k$: set of classes present at node k
- $\mathcal{R}_{k1}, \mathcal{R}_{k2} \subset \mathcal{R}_k$: sets of classes present at the left and right child of node k, respectively.

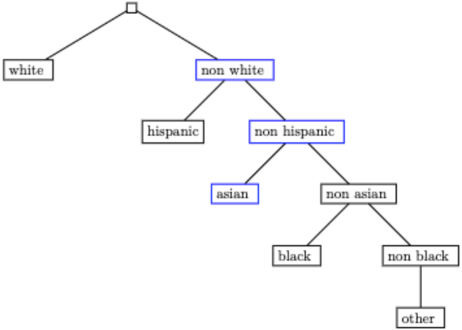


Figure 5: Regular approach to compute probabilities e.g. $p(R_i = asian | G_i, S_i)$.

Nested dichotomies

The order in which the tree was built is irrelevant under R:

Theorem (Theorem of Conditional Independence)

This theorem states that if A, B, and C are events in a sample space, and it holds that:

$$P(A) = P(A \mid B) \cdot P(B \mid C) \cdot P(C)$$

then it also holds that:

$$P(A) = P(A \mid C) \cdot P(C \mid B) \cdot P(B)$$

This implies that the probability of A is independent of the order in which events B and C are conditioned, provided that the conditional probabilities are defined and nonzero.

Thresholds approaches

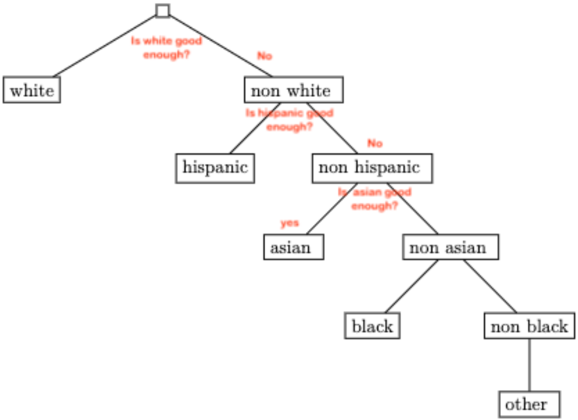


Figure 6: Threshold approach in which the individual is marked as asian

Thresholds approaches

Once we built the optimal tree[1]:

- **I. Discard Sequentially (DS):** ask sequentially if the prediction is good enough (given the optimized threshold).
  - If yes, stop
  - If not, continue to the next layer
  - The last layer is the default option
- **II. Discard Sequentially Strengthened (DSS):** Discard Sequentially + BUT If any of the predictions is not good enough, then take the maximum among the predictions.

[1]White was fixed at first level and other at the last level, and we evaluate the 6 possibles combinations for asian, hispanic and black race considering perfomances metrics

Results I

Table 6: Recall

Cohort/Metric	BISG	R	DS	DSS
white	0.93	0.93	0.83	0.83
black	0.47	0.41	0.58	0.64
asian	0.58	0.55	0.61	0.61
hispanic	0.84	0.84	0.85	0.85
other	0.03	0.04	0.12	0.09

Table 7: Precision

BISG	R	DS	DSS
0.78	0.78	0.83	0.83
0.66	0.66	0.55	0.51
0.81	0.81	0.72	0.72
0.88	0.87	0.86	0.86
0.36	0.21	0.10	0.14

Results II

Table 8: Recall

Cohort/Metric	BIFSG	R	DS	DSS
white	0.44	0.97	0.87	0.87
black	0.60	0.33	0.66	0.70
asian	0.58	0.51	0.66	0.66
hispanic	0.93	0.81	0.80	0.80
other	0.06	0.00	0.15	0.11

Table 9: Precision

BIFSG	R	DS	DSS
0.90	0.78	0.89	0.89
0.43	0.70	0.55	0.54
0.79	0.85	0.59	0.59
0.36	0.83	0.83	0.83
0.17	0.00	0.18	0.20

*Note: first name is included only for white and black cohorts

Future work

- Explore Nested Dichotomies applied to fully Bayesian Improve Surname Geocoding (fBISG)
- Apply the algorithm to UK case. The challenge is to find a dataset of self-reported race, with geolocation and surname.
- Investigate more on calibration properties of the bayesian approaches and the extension proposed

References I

Census, U. (2010). Census Surname Data.

Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., and Lurie, N. (2008). A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health services research*, 43:1722–1736.

Elliott, M. N., Morrison, P. A., Fremont, A. M., McCaffrey, D. F., Pantoja, P. M., and Lurie, N. (2009). Using the census bureau' s surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9:69–83.

Frank, E. and Kramer, S. (2004). Ensembles of nested dichotomies for multi-class problems. In *Proceedings of the 21st International Conference on Machine Learning*, volume 69 of *ACM International Conference Proceeding Series*. ACM.

Haley, J. M., Dubay, L., Garrett, B., Caraveo, C. A., Schuman, I., Johnson, K., Hammersla, J., Klein, J., Bhatt, J., Rabinowitz, D., et al. (2022). Collection of race and ethnicity data for use by health plans to advance health equity. *Working Paper*.

Imai, K. and Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2):263–272.

References II

Imai, K., Olivella, S., and Rosenman, E. T. (2022). Addressing census data problems in race imputation via fully bayesian improved surname geocoding and name supplements. *Science Advances*, 8(49):eadc9824.

Larry Baeder, Erica S. Baird, P. B. J. L. C. S. K. T.-D. G. U. N. W. and Woldeyes, M. (2024). Statistical methods for imputing race and ethnicity. *Society of Actuaries*.

Leathart, T., Frank, E., Pfahringer, B., and Holmes, G. (2018). On the calibration of nested dichotomies for large multiclass tasks. *ArXiv*, abs/1809.02744.

Robert, C. P. and Casella, G. (1999). *The Gibbs Sampler*, pages 285–361. Springer New York, New York, NY.

Word, D. L. and Jr, R. C. P. (1996). Building a spanish surname list for the 1990's—a new approach to an old problem. *U.S. Census Bureau*.

Table A.1 (parenthetical values are calculated from the 1,973,362 patients in the primary data set).

Table A.1: Probabilities of Joint Surname Test Results by True Race/Ethnicity

	On Asian Surname List (AS=1)	*On Spanish Surname List but Not Asian List (HS=1 & AS=0)*	*On Neither Surname List (AS=HS=0)*
Self-Reported Asian	d (0.515)	(1-g)(1-d) (0.011)	g(1-d) (0.474)
Self-Reported Hispanic	1-e (0.004)	ef (0.801)	e(1-f) (0.195)
Self-Reported Black or NW White	1-e (0.004)	e(1-g) (0.022)	Eg (0.973)

Figure 7: Conditional probabilities $p(S_i|R_i)$.

- For Asian List, sensitivity (d) is $p(AS = 1|Asian)$ and specificity (e)is $p(AS = 0|Not\ Asian)$
- f and g is sensitivity and specificity, respectively, of the Hispanic List

Let $\mathcal{R} = \{\text{white, black, hispanic, asian, other}\}$

Algorithm 2 Nested dichotomies

Input voters file and census

Initialize $n \leftarrow |\mathcal{R}|, \quad k \leftarrow 1, \quad \mathcal{R}_k \leftarrow \mathcal{R}$
while $k \leq n$ **do**
  Select $r_k \in \mathcal{R}_k$
  Update $\mathcal{R}_k \leftarrow \mathcal{R}_k - \{r_k\}$
  Find $p(r_k|S_i), p(\mathcal{R}_k|S_i) \leftarrow \sum_{r' \in \mathcal{R}_k} p(r'_k|S_i)$ and compute $p(G_i|r_k),\ p(G_i|\mathcal{R}_k)$.
  Calculate $p(r_k|S_i, G_i) \leftarrow p(r_k|S_i) * p(G_i|r_k)$ and $p(\mathcal{R}_k|S_i, G_i) \leftarrow p(\mathcal{R}_k|S_i) * p(G_i|\mathcal{R}_k)$
  Normalize $p(r_k|S_i, G_i) \leftarrow \frac{p(r_k|S_i,G_i)}{p(r_k|S_i,G_i)+p(\mathcal{R}_k|S_i,G_i)}$ and $p(\mathcal{R}_k|S_i, G_i) \leftarrow \frac{p(\mathcal{R}_k|S_i,G_i)}{p(r_k|S_i,G_i)+p(\mathcal{R}_k|S_i,G_i)}$
  Update $p(r'_k|S_i) \leftarrow \frac{p(r'_k|S_i)}{1-p(r_k|S_i)}$ for $r'_k \in \mathcal{R}_k$
  k $+= 1$
end while
Output $\{\ p(r_k|S_i, G_i), p(\mathcal{R}_k|S_i, G_i)\}_{k=1}^{n}$

Table 10: Recall

Cohort/Metric	BIFSG	R	DS	DSS
white	0.68	0.97	0.87	0.87
black	0.66	0.35	0.60	0.64
asian	0.56	0.54	0.64	0.67
hispanic	0.70	0.68	0.66	0.57
other	0.04	0.00	0.17	0.11

Table 11: Precision

BIFSG	R	DS	DSS
0.82	0.78	0.89	0.89
0.40	0.60	0.53	0.50
0.61	0.51	0.38	0.38
0.50	0.89	0.83	0.83
0.30	0.00	0.16	0.18

Appendix

Distribution of self reported race in the evaluation dataset

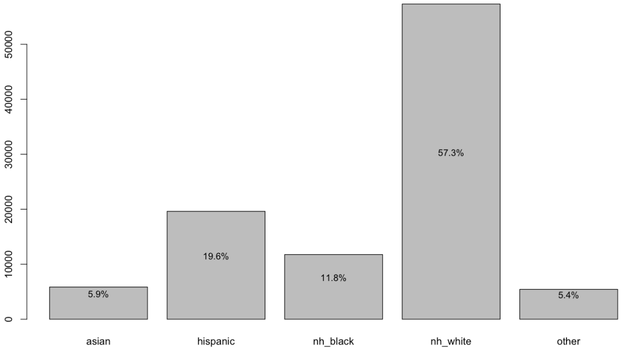


Figure 8: Data from SOA - health insurance.Includes first name for white and black cohorts